



SELF-ORGANIZING SYSTEMS AND DISEASE MODELLING

From Molecules to Populations

PhD Thesis in The Physics of Complex Systems

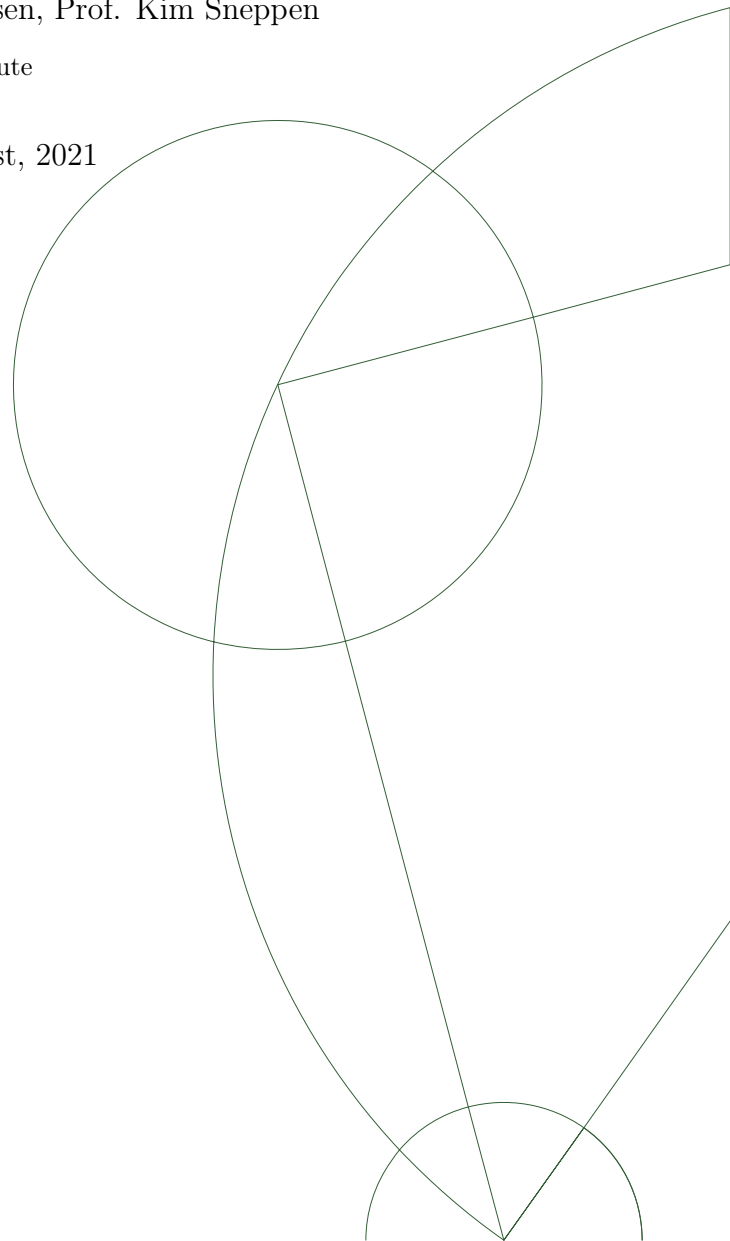
This thesis has been submitted to the PhD School of The Faculty of Science, University of Copenhagen.

Bjarke Frost Nielsen

Supervisors: Prof. Joachim Mathiesen, Prof. Kim Sneppen

Niels Bohr Institute

Tuesday 31st August, 2021



Abstract

Part 1: How are organs and complex biological shapes formed by polarized cells? Biological shape formation in the animal kingdom is characterized by enormous diversity and remarkable robustness. We present a computational point-particle model of cells endowed with apical-basal and planar cell polarity and use it to simulate tube-forming processes in vasculogenesis, gastrulation and neurulation. We find that a simple set of dynamically enforced rules can adequately capture these very different processes.

Part 2: How is the crystallization of stripe-forming block copolymers affected by substrate curvature? It is known that pattern formation of lamellar phase smectic block copolymers is sensitive to intrinsic curvature. In our study, we show that the systematic inclusion of finite thickness fundamentally changes this picture, resulting in a coupling between stripe orientation and the extrinsic curvature of the substrate. We simulate the model obtained and show how extrinsic as well as intrinsic curvature can be used to guide pattern formation.

Part 3: How does heterogeneity and superspreading affect the mitigation of infectious diseases? The spread of infectious diseases can be affected by heterogeneities of many different kinds, from differences in the disease progression to social and behavioural differences and heterogeneities in infectiousness or susceptibility. Using agent-based models, we focus on two different types of heterogeneity, namely social activity in the form of contact rates and network structure, and transmission overdispersion in the form of superspreading. We find that superspreading has profound implications for the effectiveness of lockdown-like mitigation strategies and that heterogeneous social activity is generally beneficial for contact tracing. Finally, we show that superspreading and non-pharmaceutical interventions may conspire to affect the evolution of a highly overdispersed pathogen such as SARS-CoV-2.

Dansk resumé

Del 1: Hvordan dannes organer og komplekse biologiske strukturer af polariserede celler? Dannelse af biologiske former i dyreriget er præget af en slående diversitet og robusthed. Vi præsenterer en computationel punktpartikelmodel af celler understøttet med to polariteter: apical-basal (AB) og planar cell polarity (PCP). Ved hjælp af denne model simulerer vi rørdannelsesprocesser i vaskulogenese, gastrulation og neuralrørsdannelse. Vi finder at et sæt af simple, dynamisk opretholdte regler kan indfange disse meget forskelligartede processer.

Del 2: Hvordan påvirkes sribedannende blok-copolymerers krystallisering af substratets krumning? Det er kendt at mønsterdannelsen i smektiske blok-copolymerer i lamel-fasen er følsom overfor intrinsisk krumning. I vores studie viser vi, at systematisk inklusion af en endelig filmtykkelse ændrer billedet fundamentalt og leder til en kobling mellem sribeorienteringen og substratets ekstrinsiske krumning. Vi simulerer den resulterende model og viser hvordan ekstrinsisk såvel som intrinsisk krumning kan bruges til at kontrollere mønsterdannelsen.

Del 3: Hvordan påvirker heterogenitet og superspredning afbødningen af smitsomme sygdomme? Smitsomme sygdommes spredning kan påvirkes af mange former for heterogenitet, fra forskelle i sygdomsforløb til sociale og adfærdsmæssige forskelle og heterogeniteter i smitsomhed eller modtagelighed. Ved hjælp af agentbaserede modeller fokuserer vi på heterogenitet på to områder, nemlig social aktivitet, i form af kontaktrater og netværksstruktur, samt overdispersion i sygdomsoverførsel (superspredning). Vi finder at superspredning har betragtelige konsekvenser for virkningen af lockdown-/nedlukningslignende afbødningsstrategier og at heterogen social aktivitet generelt er gavnligt for kontaktopsporing. Endelig viser vi at superspredning og ikke-farmakologiske indgreb sammen kan påvirke evolutionen af et udpræget superspredende patogen såsom SARS-CoV-2.

Contents

Introduction	5
1 Biological self-organization and shape formation: Epithelial cells	7
1.1 Budding, wrapping and invagination	8
1.2 Vasculogenesis	13
1.3 Discussion	18
1.4 Publications for Chapter 1	20
Manuscript: Model to link cell shape and polarity with organogenesis	21
Manuscript: Self-assembly, buckling and density-invariant growth of three-dimensional vascular networks	42
2 Chemical self-organization: Block copolymers	51
2.1 Introduction	51
2.2 Methods	52
2.3 Results	57
2.4 Discussion	60
2.5 Publications for Chapter 2	61
Manuscript: Substrate curvature governs texture orientation in thin films of smectic block copolymers	62
3 Spreading and heterogeneity: COVID-19	76
3.1 Superspreading	77
3.2 Contact tracing in heterogeneous social networks	92
3.3 Discussion	98
3.4 Publications for Chapter 3	100
Manuscript: Overdispersion in COVID-19 increases the effectiveness of limiting non-repetitive contacts for transmission control	101
Manuscript: COVID-19 superspreading suggests mitigation by social network modulation	108
Manuscript: Lockdowns exert selection pressure on overdispersion of SARS-CoV-2 variants	115
Manuscript: Differences in social activity increase efficiency of contact tracing	122
Manuscript: The COVID-19 pandemic: Key considerations for the epidemic and its control	142
Additional publications	174
Manuscript: Newton-Cartan submanifolds and fluid membranes	175

The reductionist hypothesis does not by any means imply a “constructionist” one. The ability to reduce everything to simple fundamental laws does not imply the ability to start from those laws and reconstruct the universe. The constructivist hypothesis breaks down when confronted with the twin difficulties of scale and complexity.

More Is Different
P. W. Anderson

ACKNOWLEDGEMENTS

I would like to thank my supervisors, Joachim Mathiesen and Kim Sneppen for always welcoming debate and for acknowledging and emphasizing the creative aspects of the scientific process.

Furthermore, I would like to express my gratitude towards Ala Trusina, Gaute Linga, Julius Kirkegaard, Lone Simonsen, Kristian Stølevik Olsen, Emil Have, Robert Taylor, Amalie Christensen, Andreas Eilersen and Søren Ørskov. Last – but certainly not least – I thank my wife, Dr. Olivia Frost Lorentzen for her unwavering support and constructive honesty.

INTRODUCTION

This thesis is split into three distinct parts, not because each is completely unrelated to the others, but because treating each part as a self-contained story only makes the exposition clearer. The thesis is written in the style of a synopsis. As such, it provides a moderate level of detail about each of the subjects it covers, with the understanding that a more detailed exposition is available in the published manuscripts. The relevant manuscripts are included at the end of each chapter.

The first part concerns research we have done on the subject of the dynamics of polar cells in the context of morphogenesis – the formation of biological shapes. In many organisms, some cells – epithelial cells in particular – express one or more polarities. These polarities facilitate spatial symmetry breaking and affect the mechanical and adhesive properties of cells and are instrumental in the formation of cell sheets and elongated structures. Our research is an attempt to create a simple dynamical model of these polarized cells and their interactions, in order to simulate processes in organ formation and identify the necessary and sufficient mechanisms for embryonic shape formation to occur.

The model is fully local, meaning that shapes arise on the basis of interaction between cells, and not due to externally orchestrated signals. This self-organizing system can lead to some highly nontrivial shapes, such as the one illustrated in Figure 1. Furthermore, our model allows for self-assembly of complex structures from random initial conditions, a property which formed the basis of our studies of vasculogenesis.

The second part concerns another self-organizing system, albeit of a non-living nature. Diblock copolymers have the ability to spontaneously assemble into nano-scale striped patterns which have several promising applications. Among these are the use of thin films of block copolymers as etch masks in lithographic fabrication of microelectronic circuitry elements. However, for this to bear fruit, a method for guiding the pattern formation is necessary. Our contribution to this problem consists of a continuum model of how patterns form on curved surfaces, which shows how one may direct the patterns formed by means of manipulating the geometry of the underlying substrate. We simulate the equations obtained in several cases and demonstrate the kinds of patterns that can then form on cylindrical surfaces, Gaussian bumps, saddle geometries and irregular (but smooth) bumpy surfaces.

The third and final part concerns the spread of disease in a population under non-pharmaceutical interventions, and how the spread is affected by certain heterogeneities.

This story consists of two separate arcs. One has to do with contact tracing – specifically TTI, test-trace-isolate – and how realistic heterogeneities of social contact networks and contact rates impact the effectiveness of this mitigation strategy.

The second story arc concerns transmission heterogeneity. One of the more striking features of the COVID-19 pandemic has been overdispersed transmission - i.e. *superspreading* - the phenomenon that some individuals infect very many while the majority of infected persons hardly infect at all. Our work shows that this feature of the disease renders the epidemic highly vulnerable to lockdown-type interventions. These interventions thus have a much larger effect than would be the case in a disease with a similar basic reproductive number but more homogeneous spreading. This result is connected to our finding that random, non-repeated contacts make an outsized contribution to a superspreading epidemic. In a separate paper, we studied the effects of social network modulation on an overdispersed disease. Here we found that a superspreading disease is highly sensitive to reductions in personal contact network size as well as to the clustering of said network. The last

project of this thesis concerns the evolution of new SARS-CoV-2 variants with different levels of overdispersion. This was inspired by observations that the Alpha variant, while exhibiting a higher mean respiratory viral load, actually exhibited a lower relative variation when compared with the ancestral variant. Our simulations showed that non-pharmaceutical interventions may exert a selection pressure, favouring the development of more homogeneously spreading variants.

CHAPTER 1

BIOLOGICAL SELF-ORGANIZATION AND SHAPE FORMATION: EPITHELIAL CELLS

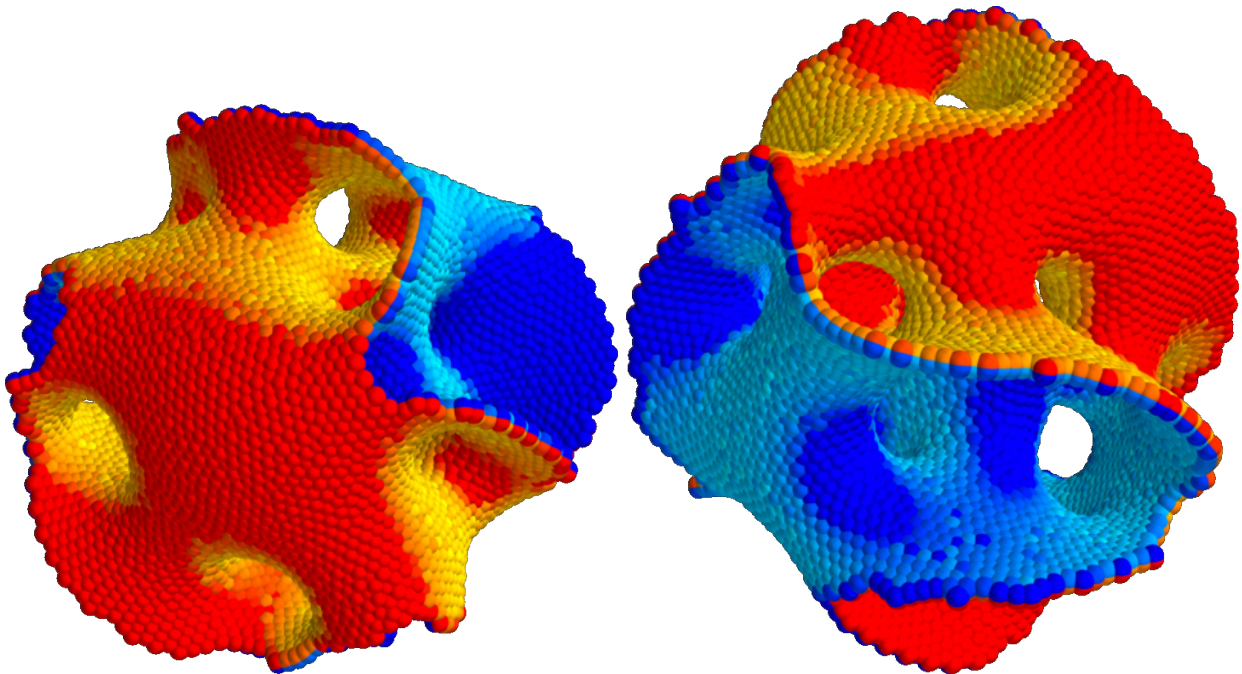


Figure 1: A structure composed of AB-polarized cells, formed by self-organized aggregation of an initially randomized collection of cells. Blue and red color schemes distinguish apical from basal, while the shading indicates apical-basal (AB) polarity orientation.

The question of how organs are formed at the scale of interacting biological cells can be posed and answered at several different levels. There are questions of how cells communicate locally, which biochemical signaling agents are released and how the temporal and spatial regulation of these organizing molecules occurs. None of these questions have been the focus of our work in this area. Rather, we have taken the somewhat coarse-grained approach of attempting to answer the following question:

“Given a few empirical rules, can one formulate a simple 3-dimensional dynamical model of polarized cell-cell interactions which can reproduce central processes and transitions in organogenesis?”

Our goal was to reproduce morphogenic transitions in a model where

- All interactions are local (i.e. cell-to-cell neighbour interactions),
- Two types of cellular polarity are included: apical-basal (AB) and planar cell polarity (PCP),

- Cells are represented as point particles.

Our main focus has been on transitions which involve *sheets* of cells. Either the self-assembly of cell monolayers (sheets) or transitions of preexisting cell sheets which result in topological changes. The work presented here is a continuation of theoretical work published in [1] and [2].

Simply put, the two types of polarity mentioned above each serve a well-defined primary role. Apical-Basal (AB) polarity induces and maintains a sheet-like structure by compelling a cell to align its own AB polarity with those of its neighbours and to preferentially adhere to cells which lie in the plane orthogonal to this polarity. Planar Cell Polarity (PCP) is slightly more involved. Its primary purpose in our model is to facilitate convergent extension, a biological mechanism behind the elongation of tissues by means of T1 transitions.

1.1 BUDDING, WRAPPING AND INVAGINATION

This subsection is based entirely on the work published as Ref. [3].

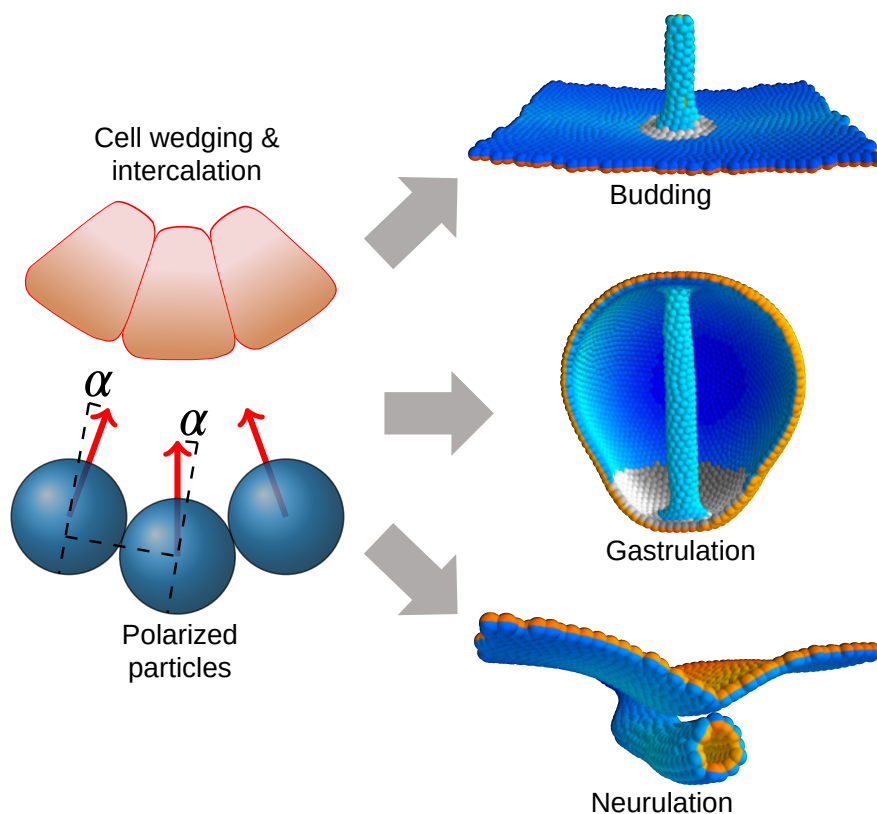


Figure 2: Wedging and intercalation of polarized cells drive central processes in development. Figure from Ref. [3].

In morphogenesis, the topology-changing formation of tubular structures from flat sheets of cells has widespread importance. How, for instance, do gut and neural tubes form from flat sheets of epithelial cells? In *Drosophila* gastrulation and mammalian neurulation, a (section of a) flat sheet of cells *wraps* in an axisymmetric manner until the edges meet and fuse to form a tube (a simulation of this is visualized in Figure 4D-F). The “orthogonal process” also occurs, with a tube being formed perpendicularly to the initial sheet of cells (illustrated in Figure 4A-C). This mode of tube formation is called budding. One example of this is the gut formation – gastrulation – of sea urchins. Budding is ubiquitous in organ development, with examples including salivary glands and trachea in *Drosophila* and lung and kidney development in vertebrates [4].

Several mechanisms may participate in these transitions, but it is not clear to what extent they are

necessary, sufficient or complementary. The mechanisms behind sheet-to-tube transitions include cell shape changes – apical and basal constriction, where either the apical [5] or basal [6, 7] sides of cells constrict to render the cell wedge-shaped – and convergent extension by directed cell intercalation [4, 8]. Additionally, spatially restricted cell division and apoptosis can contribute to the formation of tubular structures [4].

Below, we give a description of our model of interacting cells in 3D, equipped with apical-basal and planar cell polarity as well as a mechanism for generating spontaneous curvature in the cell sheet, mimicking apical or basal constriction. We use this model to investigate the necessary and sufficient mechanisms for epithelial sheet-to-tube transitions.

1.1.1 METHODS

Our work builds on the basic model originally developed in [2]. We have extended the model to include cell wedging effects, which we will introduce below. First, we will describe the base model in the absence of wedging.

The Base Model. Cells are treated as point particles and interact with neighbouring cells through a pair potential V_{ij} (with indices i and j labeling cells). The potential has a rotationally symmetric (i.e. polarity independent) repulsion term and a polarity-modulated attraction term. The potential can be formulated in dimensionless form in terms of the inter-cell distance r_{ij} as follows:

$$V_{ij} = e^{-r_{ij}} - [\lambda_1 S_{ij}(A) + \lambda_2 S_{ij}(AP) + \lambda_3 S_{ij}(P)] e^{-r_{ij}/\beta}. \quad (1.1)$$

Here, β determines the energetically optimal distance (i.e. the equilibrium distance) between cell centers and is fixed at $\beta = 5$, since this ensures an equilibrium distance of 2 (corresponding to measuring all lengths in units of the typical cell radius). The parameters λ_n ($n = 1, 2, 3$) are coupling constants which determine the strength of polar interactions and satisfy a normalization condition $\sum_n \lambda_n = 1$. There are two types of polarities in the model, modeled as unit vector quantities associated with each cell. The vector \mathbf{p} represents apical-basal (AB) polarity while \mathbf{q} represents planar cell polarity (PCP). The three polarity-dependent factors are given by the following expressions:

$$S_{ij}(A) = (\mathbf{p}_i \times \hat{\mathbf{r}}_{ij}) \cdot (\mathbf{p}_j \times \hat{\mathbf{r}}_{ij}) \quad (1.2)$$

$$S_{ij}(AP) = (\mathbf{p}_i \times \mathbf{q}_j) \cdot (\mathbf{p}_j \times \mathbf{q}_i) \quad (1.3)$$

$$S_{ij}(P) = (\mathbf{q}_i \times \hat{\mathbf{r}}_{ij}) \cdot (\mathbf{q}_j \times \hat{\mathbf{r}}_{ij}) \quad (1.4)$$

Since each of these expressions are scalar quadruple products, they can be rewritten as sums (of products) of inner products, which is perhaps more illuminating:

$$S_{ij}(A) = \mathbf{p}_i \cdot \mathbf{p}_j - (\mathbf{p}_i \cdot \hat{\mathbf{r}}_{ij})(\mathbf{p}_j \cdot \hat{\mathbf{r}}_{ij}), \quad \text{using that } \hat{\mathbf{r}}_{ij} \cdot \hat{\mathbf{r}}_{ij} = 1, \quad (1.5)$$

$$S_{ij}(AP) = (\mathbf{p}_i \cdot \mathbf{p}_j)(\mathbf{q}_i \cdot \mathbf{q}_j) - (\mathbf{p}_i \cdot \mathbf{q}_i)(\mathbf{p}_j \cdot \mathbf{q}_j), \quad (1.6)$$

$$S_{ij}(P) = \mathbf{q}_i \cdot \mathbf{q}_j - (\mathbf{q}_i \cdot \hat{\mathbf{r}}_{ij})(\mathbf{q}_j \cdot \hat{\mathbf{r}}_{ij}). \quad (1.7)$$

The equilibrium configuration is somewhat more easily gleaned from this description, since the dynamics is such that the quantities which enter with a positive sign will be maximized while those with a negative sign will be minimized. Close to equilibrium, we furthermore expect (heuristically)

- $\mathbf{p}_i \cdot \mathbf{p}_j = 1 + \mathcal{O}(\varepsilon)$
- $\mathbf{p}_i \cdot \hat{\mathbf{r}}_{ij} = \mathcal{O}(\varepsilon)$
- $\mathbf{q}_i \cdot \mathbf{q}_j = 1 + \mathcal{O}(\varepsilon)$
- $\mathbf{p}_i \cdot \mathbf{q}_j = \mathcal{O}(\varepsilon)$

$$\bullet \mathbf{q}_i \cdot \hat{\mathbf{r}}_{ij} = \mathcal{O}(\varepsilon)$$

with ε a small deviation from the optimal configuration. We mention this mostly in order to build intuition about each term: for instance, $\mathbf{p}_i \cdot \mathbf{p}_j - 1$ is a “first order term” while e.g. $(\mathbf{p}_i \cdot \hat{\mathbf{r}}_{ij})(\mathbf{p}_j \cdot \hat{\mathbf{r}}_{ij})$ is a second order term in this description, and should thus be subdominant. A systematic classification and description of polar point particle models as a near-equilibrium expansion is underway, but not yet finished for inclusion in this thesis.

The overall purpose of $S_{ij}(A)$ is to dynamically organize cells into a sheet with AB polarities

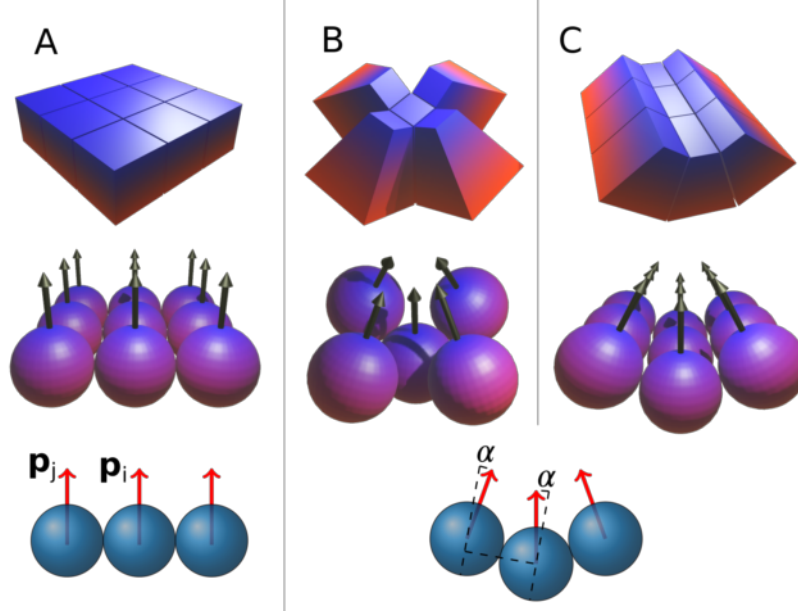


Figure 3: In the absence of wedging (panel A), apical-basal (AB) polarity favours a flat sheet. Wedging is introduced by an interaction which energetically favours tilted AB polarities. Wedging may be either isotropic (panel B) or anisotropic (pane C) with the latter leading to axially symmetric bending of sheets.

orthogonal to the local cell layer. $S_{ij}(AP)$ mainly exists to ensure orthogonality of apical-basal and planar cell polarities. Finally, $S_{ij}(P)$ keeps planar cell polarities uniform across neighbouring cells (dynamically enforcing $\mathbf{q}_i \cdot \mathbf{q}_j \approx 1$). It should be noted that while $S_{ij}(A)$ and $S_{ij}(P)$ are similar, save for depending on AB polarity and PCP, respectively, the AB-related coupling constant λ_1 must always be greater than the PCP-associated λ_3 . Otherwise, the two types of polarity will switch roles. In other words, the symmetry between AB and PCP is broken simply by one coupling much more strongly to the local geometry than the other does.

When parametrized in this way ($\lambda_1 \sim \lambda_2 \gg \lambda_3$), planar cell polarity has the effect of facilitating elongation of structures (lateral organization) through convergent extension.

The time development of the model is governed by overdamped (*relaxational*) dynamics under the above potential:

$$\frac{\partial \mathbf{r}_i}{\partial t} = -\frac{\partial V_i}{\partial \mathbf{r}_i} + \eta, \quad (1.8)$$

$$\frac{\partial \mathbf{p}_i}{\partial t} = -\frac{\partial V_i}{\partial \mathbf{p}_i} + \eta, \quad (1.9)$$

$$\frac{\partial \mathbf{q}_i}{\partial t} = -\frac{\partial V_i}{\partial \mathbf{q}_i} + \eta, \quad (1.10)$$

where $V_i = \sum_j V_{ij}$ is the total potential experienced by the cell i (so the sum index j runs over the neighbours of the i 'th cell). The neighbours are determined geometrically by a Voronoi-like line-of-sight criterion, which is described in more detail in [2]. Lastly, η is a Gaussian noise term with vanishing mean. Some of our simulations also include cell division. This is incorporated as a Poisson (i.e. constant-rate) process with daughter cells being spawned randomly at a distance of

one cell radius from the mother cell.

Wedging. The potential as described above – specifically the $S_{ij}(A)$ term – favours the formation of a flat cell sheet. We introduce wedging by extending this model with a single parameter, α . This deformation parameter introduces an energetically favoured *tilt* in neighbouring AB polarity vectors. Specifically, we modify $S_{ij}(A)$ according to

$$S_{ij}(A) = (\tilde{\mathbf{p}}_i \times \hat{\mathbf{r}}_{ij}) \cdot (\tilde{\mathbf{p}}_j \times \hat{\mathbf{r}}_{ij}), \quad (1.11)$$

where

$$\tilde{\mathbf{p}}_i = \mathbf{p}_i, \quad \text{for no wedging,} \quad (1.12)$$

$$\tilde{\mathbf{p}}_i \propto \mathbf{p}_i - \alpha \hat{\mathbf{r}}_{ij}, \quad \text{for isotropic wedging,} \quad (1.13)$$

$$\tilde{\mathbf{p}}_i \propto \mathbf{p}_i - \alpha(\mathbf{q}) \hat{\mathbf{r}}_{ij}, \quad \text{for anisotropic wedging.} \quad (1.14)$$

The proportionality sign \propto just indicates that we have suppressed a normalization factor in the equations. The introduction of the wedging parameter α thus causes the preferred angle between neighbouring polarities to deviate from 0 (i.e. parallel), as illustrated in Figure 2 and Figure 3. In the isotropic case (Figure 3B), the favoured *tilt* is the same towards all neighbouring cells. In the case of anisotropic wedging (Figure 3C), the wedging parameter α is modulated by the direction of PCP. When computing the interaction between cells i and j , $\alpha(\mathbf{q})$ is given by:

$$\alpha(\mathbf{q}) = \alpha_0 \langle \mathbf{q} \rangle_{ij} \cdot \hat{\mathbf{r}}_{ij}, \quad (1.15)$$

where $\langle \mathbf{q} \rangle_{ij}$ denotes the arithmetic mean of the PCP vectors of cells i and j . This ensures that wedging primarily happens towards (or away from, depending on the sign of α_0) the cells whose PCP is aligned along the line of sight between the two cells. The effect is for PCP to cause uniformly PCP-polarized cell sheets to “curl up” such that the PCP field ends up running *around* the resulting tube or groove.

1.1.2 RESULTS

Here we give a brief account of our results - further details can be found in [3].

Roles of Convergent Extension and Wedging in Budding Experimental results suggest that wedging and convergent extension (CE) both contribute to invagination in budding transitions [9, 8]. Computational models, on the other hand, have generally focused on *either* CE as a driver of tissue elongation or wedging as the driving force of invagination. We combine the two mechanisms in order to probe their roles in budding, separately and combined.

First, we find that apical constriction alone is not enough to initiate budding. A ring of *basally* constricting cells (as in Figure 4A), however, can facilitate invagination by themselves.

Our simulations of budding show that successful invagination and tube elongation can happen when both wedging and PCP (which drives CE) act in parallel, see Figure 4A-C (videos of these simulations can be found in [3]). This is not restricted to a planar geometry, and we have also simulated budding starting from a spherical shell (as in sea urchin gastrulation, see Figure 5).

In addition, we show that budding can proceed in the absence of wedging provided that noise is present (random fluctuations in cell position and polarity orientation). Even slight noise is sufficient to facilitate the symmetry breaking between the two sides of the cell layer, and allow CE-driven tubulation to take over. The limitation is that directional robustness of the process is destroyed, and tube formation may occur on either side of the local sheet. Furthermore, relying on noise to break the symmetry results in a higher proportion of failed invaginations at low noise levels. This leads us to suggest that the primary role of wedging in budding is to ensure correct orientation and consistent initial invagination.

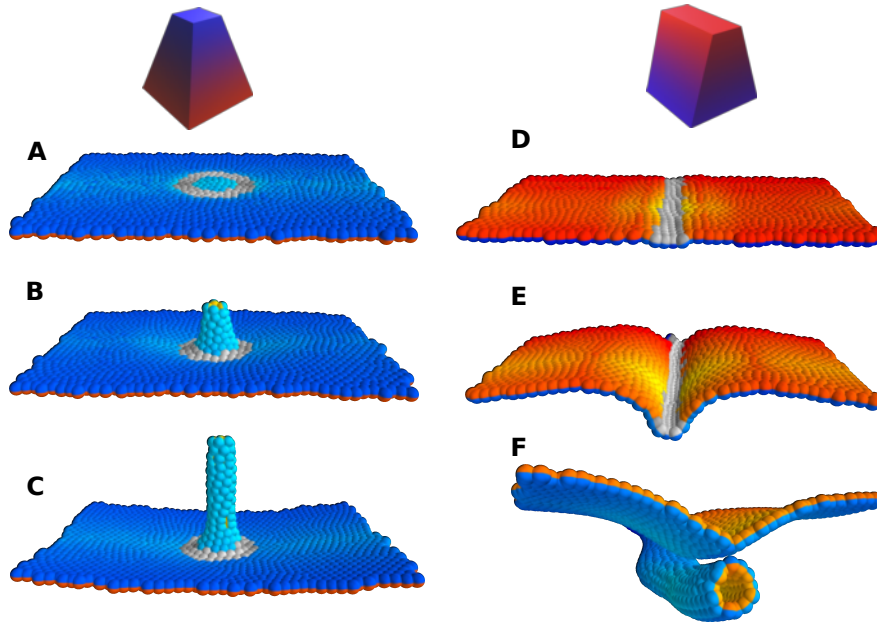


Figure 4: Simulations of tube formation by budding and wrapping. Panels A-C) Budding, where initial invagination is provided by a ring of isotropically wedging cells, constricting basally. Panels D-F) Wrapping. Here, wedging is anisotropic and modulated by planar cell polarity. Buckling is driven by differential proliferation. The polyhedra at the top of the figure illustrate isotropic vs. anisotropic wedging.

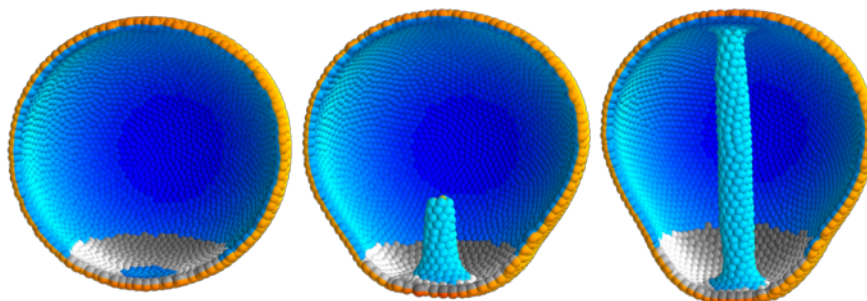


Figure 5: Simulation of budding in a spherical geometry, as is seen in Sea urchin gastrulation.

Anisotropic wedging and differential proliferation can drive wrapping The other sheet-to-tube transition we simulate is wrapping, as occurs in e.g. vertebrate neurulation. A stripe in the middle of the cell sheet represents the neuroepithelium (NE) and is assigned to anisotropically apically constrict (the grey cells in Figure 4B). The remaining cells then represent ectoderm (E). Initially, PCP is assumed to be uniform and pointing in the direction orthogonal to the axis of the future tube.

We find that this anisotropic wedging is required for wrapping. Replacing the anisotropically wedging cells by isotropic ones leads to a rounded, bulging invagination rather than a tube. With anisotropic wedging, a proto-tube groove forms. However, in order to obtain full closure of the tube, differential proliferation of cells at the NE-E boundary was found to be necessary (consistent with Ref. [10]). Concretely, we found that tube formation was possible within a rather broad range of cell cycle lengths (3-23 h, see Supporting Information of Ref. [3] for the dimensional analysis which allows translation of model timescales into hours). Longer *or* shorter cell cycles resulted in open-tube morphologies somewhat reminiscent of those seen in neural tube defects such as *spina bifida* (see Figure 6). The range of cycle durations seen here is consistent with the ≈ 4 hours seen in Ref. [10]. If proliferation is too fast, we observe that the sheet does not have time to equilibrate and that CE does not act fast enough to consistently narrow the tube. Due to this, sections of the tube fail to fuse properly. If proliferation is too slow, on the other hand, the out-of-equilibrium buckling effect is too weak and the folds never come close enough to fuse. Experimentally, slow proliferation

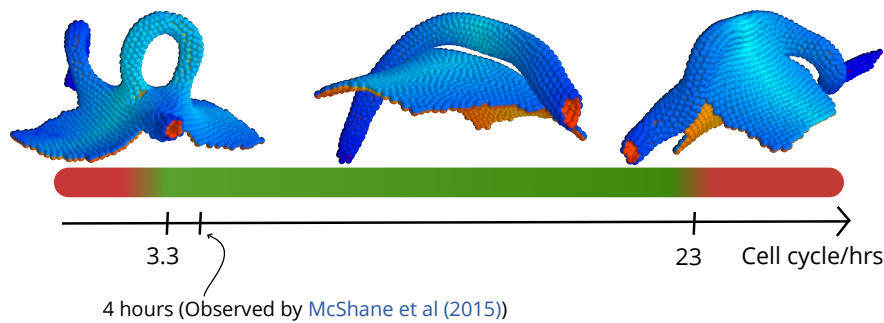


Figure 6: The fate of the neural tube as a function of cell cycle lengths (i.e. the inverse of proliferation speed). If cells divide too fast (cell cycle ≤ 3.3 hours), sections of the tube do not fuse properly while other parts bulge. If cells divide too slowly (cell cycle ≥ 23 hours) buckling is too weak and the two sides never come close enough to fuse.

has indeed been shown to lead to neural tube defects in mice [11]. In humans, mutations affecting the PAX3 transcription factor are known to be implicated in Waardenburg syndrome [12, 13] which is characterized by incomplete neural tube closure. This transcription factor is also essential in ensuring sufficient cell proliferation [14]. To our knowledge, overly fast proliferation has not been studied in this context and thus represents an interesting open question.

1.2 VASCULOGENESIS

In this section, we will treat the problem of vasculogenesis, i.e. the formation of the initial plexus of the vascular system. A remarkable property of this process is that it happens by aggregation of cells. In other words, a sparse collection of cells self-assemble into a vascular structure. Our model tackles two prominent features of this process, namely morphology-maintaining growth (i.e. proliferation without loss of topology or overall shape) and growth-induced buckling, the non-equilibrium phenomenon of proliferation generating curvature.

The remainder of this section is based on Ref. [15].

1.2.1 METHODS

The model described in the last section, and used in Ref. [3], treated polarities as unit vector quantities. It may seem that this is the natural choice for a quantity whose primary purpose is

to single out a direction. However, in biological processes of sheet formation it is not obvious that the *in-plane* polarity should not be flip-symmetric. The (out-of-plane) AB polarity of course distinguishes the apical and basal sides, and it makes sense that the theory is not symmetric¹ under the transformation $\mathbf{p}_i \rightarrow -\mathbf{p}_i$. However, the same requirement doesn't *a priori* exist for the in-plane polarity \mathbf{q} . In branching processes, which will play a central role in the next part of this section, this asymmetry is in fact counterproductive. The alternative is a nematic order where polarities are more accurately pictured as line elements rather than arrows. In Figure 7, which originally appeared in the supplementary material of Ref. [15], we show two simulations of branched structures, one with nematic PCP and one with vectorial PCP. Vectorial PCP evidently induces an asymmetry which is reflected on the macro scale. The origins of this are field defects (indicated by black arrows on the figure) which form at the branch point. To appreciate this, it is perhaps instructive to look at the flow field instead (i.e. the integral curves of the vectorial/nematic fields). These are shown in Figure 8.

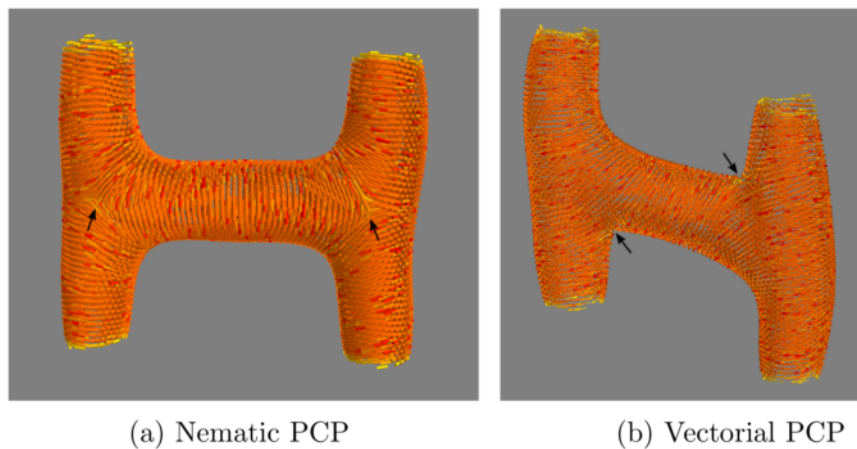


Figure 7: Branching structures with either (a) nematic or (b) vectorial planar cell polarity. The black arrows indicate defects in the PCP field. In the nematic case, four $-1/2$ defects appear symmetrically, while the vectorial case allows for just two -1 defects which induce an asymmetry in the branching.

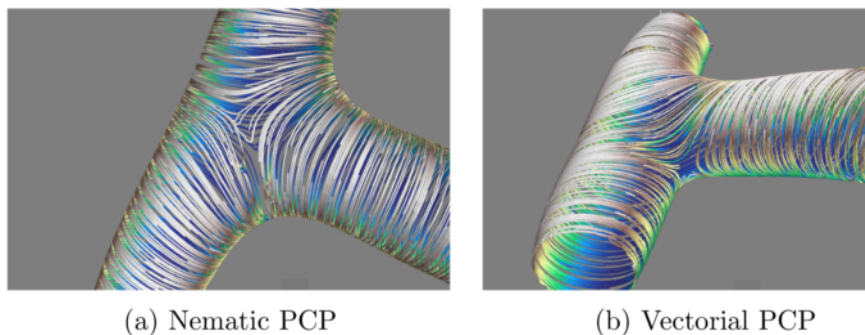


Figure 8: The same branching experiment as in the previous figure. Here, the integral curves – i.e. the stream lines – of the PCP fields are shown.

In both cases, the topological defect charges sum to the Euler characteristic $\chi = 2 - 2g = -2$ (with g the genus). However, nematic order allows for half-integer defects, enabling four $-1/2$ defects to appear symmetrically as opposed to the two -1 defects which appear in the vectorial case².

¹There is a subtlety here. The theory *is* symmetric under the transformation $\mathbf{p}_i \rightarrow -\mathbf{p}_i$ when performed for *all* cells. This simply corresponds to relabeling the apical and basal sides. It is, however, not symmetric under the flip of a single apical-basal polarity vector.

²One could object that the surfaces are not closed, and thus the Euler characteristic doesn't apply. However, the structures are such that one can close them by fusing the 'arms' pairwise, without introducing new defects or altering the genus.

In order to introduce nematic PCP, we take the polarity dependent part of the original model:

$$\tilde{S}_{ij} = \lambda_1 S_1^{ij} + \lambda_2 S_2^{ij} + \lambda_3 S_3^{ij}, \quad (1.16)$$

where

$$\begin{aligned} S_1^{ij} &= (\mathbf{p}_i \times \mathbf{r}_{ij}) \cdot (\mathbf{p}_j \times \mathbf{r}_{ij}), \\ S_2^{ij} &= (\mathbf{p}_i \times \mathbf{q}_j) \cdot (\mathbf{p}_j \times \mathbf{q}_i), \\ S_3^{ij} &= (\mathbf{q}_i \times \mathbf{r}_{ij}) \cdot (\mathbf{q}_j \times \mathbf{r}_{ij}). \end{aligned} \quad (1.17)$$

and modify it by taking the absolute value of the terms involving \mathbf{q} :

$$\tilde{S}_{ij} = \lambda_1 S_1^{ij} + \lambda_2 |S_2^{ij}| + \lambda_3 |S_3^{ij}|. \quad (1.18)$$

Furthermore, we introduce a spherically symmetric attractive term with a coupling constant λ_0 :

$$\tilde{S}_{ij} = \lambda_0 + \lambda_1 S_1^{ij} + \lambda_2 |S_2^{ij}| + \lambda_3 |S_3^{ij}|. \quad (1.19)$$

Our reason for introducing such a term is that we wish to study lumen formation, i.e. the transition from a bulk of cells to a structure with sheet-like walls. We continue to impose the normalization $\sum_{n=0}^3 \lambda_n = 1$. This allows the model to be run in the absence of any polarity ($\lambda_n = 0$ for all $n > 0$). Starting from a situation with $\lambda_0 = 1$ and then turning on AB polarity does produce a transition from a solid structure to one consisting of cell monolayers (compare panels a and b of Figure 9), but a consistent tubular structure is not formed. This also holds if PCP is turned on (by letting e.g. $\lambda_1 = 0.5$, $\lambda_2 = 0.42$, $\lambda_3 = 0.08$).

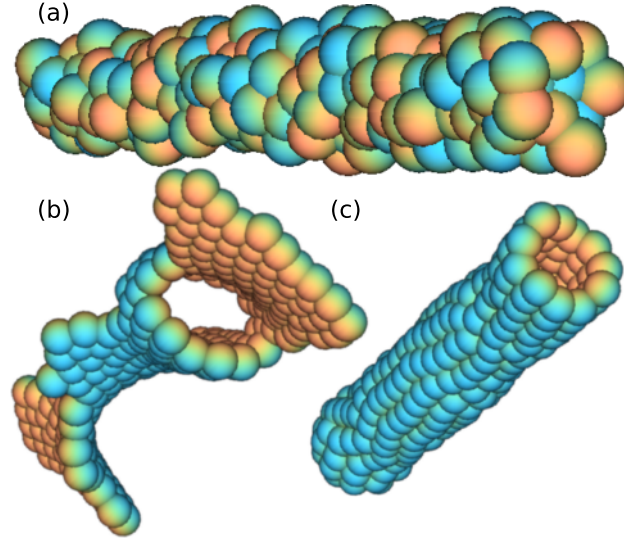


Figure 9: Transition from an initial solid structure (panel a) to a structure composed of cell monolayers. In the absence of the polarity-orienting term V_i , a structure similar to that of panel b is formed. When V_i is turned on, a tubular structure as seen in panel c can form.

To facilitate the consistent formation of AB polarity, we introduce a potential term:

$$V_i = \gamma \sum_j f(r_{ij}) \mathbf{p}_i \cdot \hat{\mathbf{r}}_{ij} \quad (1.20)$$

where $f(r) \sim \exp(-r^2/(2))$. The total potential is then $V = \sum_{ij} V_{ij} + \sum_i V_i$, where the first sum runs over all pairs of neighbouring cells. The V_i potential is inspired by experiments which have suggested that cell-cell contact orients AB polarity [16]. This potential precisely has the effect of locally aligning AB polarity towards regions of high cell density. With $\gamma = 5$ and the AB coupling

constant $\lambda_1 = 1$, this setup produces regular tubes as seen in Figure 9c, mimicking lumen formation by cord hollowing.

1.2.2 RESULTS

In Figure 10, we show the formation of a vascular network by aggregation in our model. In Figure 10a the initial, randomly positioned collection of cells are shown. Note that the color gradient in this plot does not indicate AB polarity, but is only added to aid depth perception. By $t = 2 \times 10^2$, a coarse sheet structure has started to emerge, signifying that AB polarities have mostly aligned, locally (Figure 10b). At a later stage, a vascular network has formed into a stable structure ($t = 10^4$, Figure 10c).

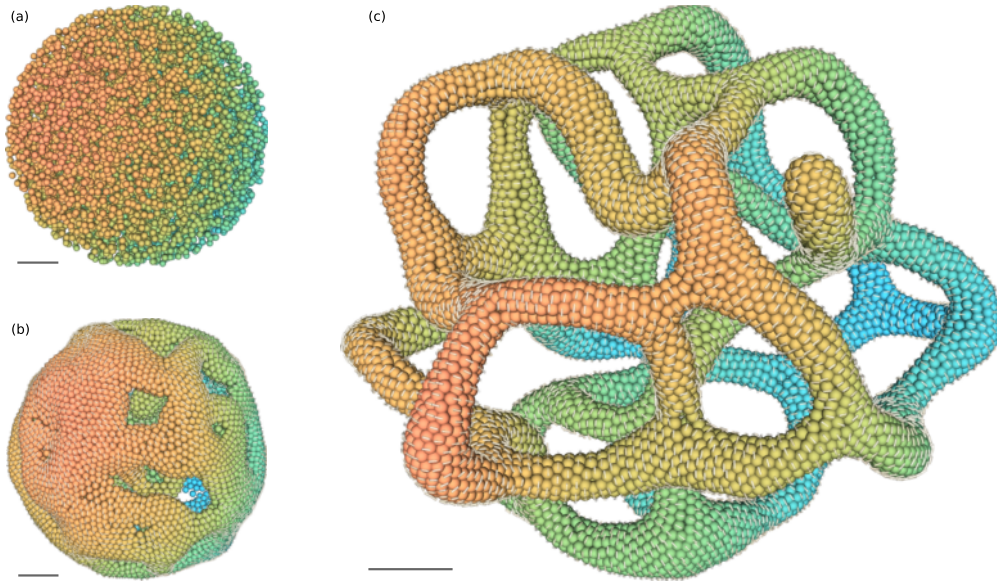


Figure 10: Formation of a vascular network by self-assembly from a disperse collection of cells. a) $t = 0$. b) $t = 2 \times 10^2$ c) $t = 10^4$. Note that color does not indicate polarity but is only added to aid depth perception.

The successful vasculogenesis seen in Figure 10 is not guaranteed, though. If the initial density of cells is too low, this process fails to take place. In fact, this process is very similar to the classic problem of percolation, and we find that it exhibits a phase transition from *disjointed lumina* to a *connected network*. In Figure 11, we plot the percolation probability as a function of the initial density of cells and find a transition occurring around a critical density of $\rho_c \approx 8.2 \times 10^{-3}$. Here, we define *percolation probability* as the probability for a cell to belong to the largest cluster. The inset of Figure 11 shows the formation of disjointed lumina at subcritical density.

After the initial vascular network is formed, it is of course biologically necessary for the system to be able to grow while maintaining its general morphology and density. Consider the growth of the vasculature of the islets of Langerhans over 44 weeks as shown in Figure 12 (image and data from Berclaz et al. [17]). Note that the vascular density remains approximately constant during this growth and that total vessel length certainly grows much more than the diameter. Another striking feature is the tortuosity of the vascular network, with vessels having undergone extensive buckling.

In order to replicate this type of growth, we induced cell division (modelled by a Poisson process as described in the last section). With $\lambda_3 = 0$, i.e. without the PCP-induced convergent extension (CE) term of the model, this results in density non-preserving growth with vessels primarily increasing their diameter, as seen in Figure 13.

However, if $\lambda_3 > 0$, the growth pattern is profoundly altered. λ_3 sets a preferred curvature of the tubes and this results in a tendency for vessels to grow in length rather than in diameter. In Figure 14, we begin with a small metastable vascular network (panel a) and then allow for proliferation until a certain effective radius has been reached. In panel b, the structure resulting from

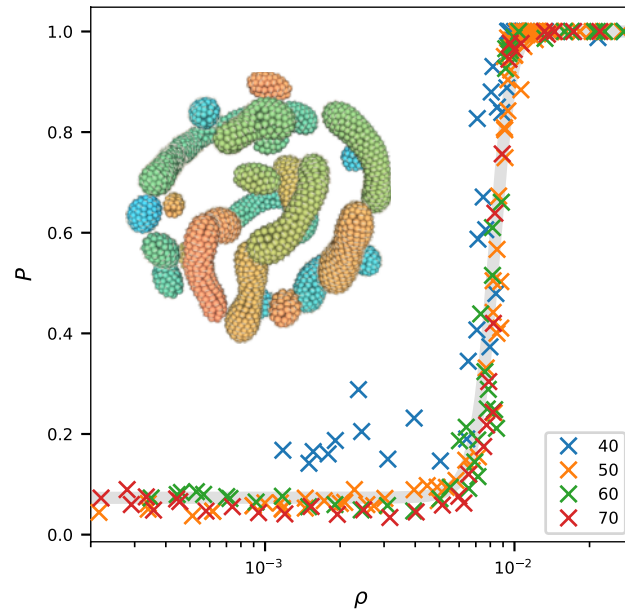


Figure 11: Transition between formation of disjointed lumina and connected networks is governed by initial density. Parameters: $\lambda_1 = 0.5$, $\lambda_2 = 0.45$, $\lambda_3 = 0.05$, $\gamma = 5.0$. The legend indicates the radius of the sphere within which cells are initiated. For radii $r \leq 40$, the percolation probability doesn't drop to zero below the threshold since the system is too small to allow for several well-formed lumina.

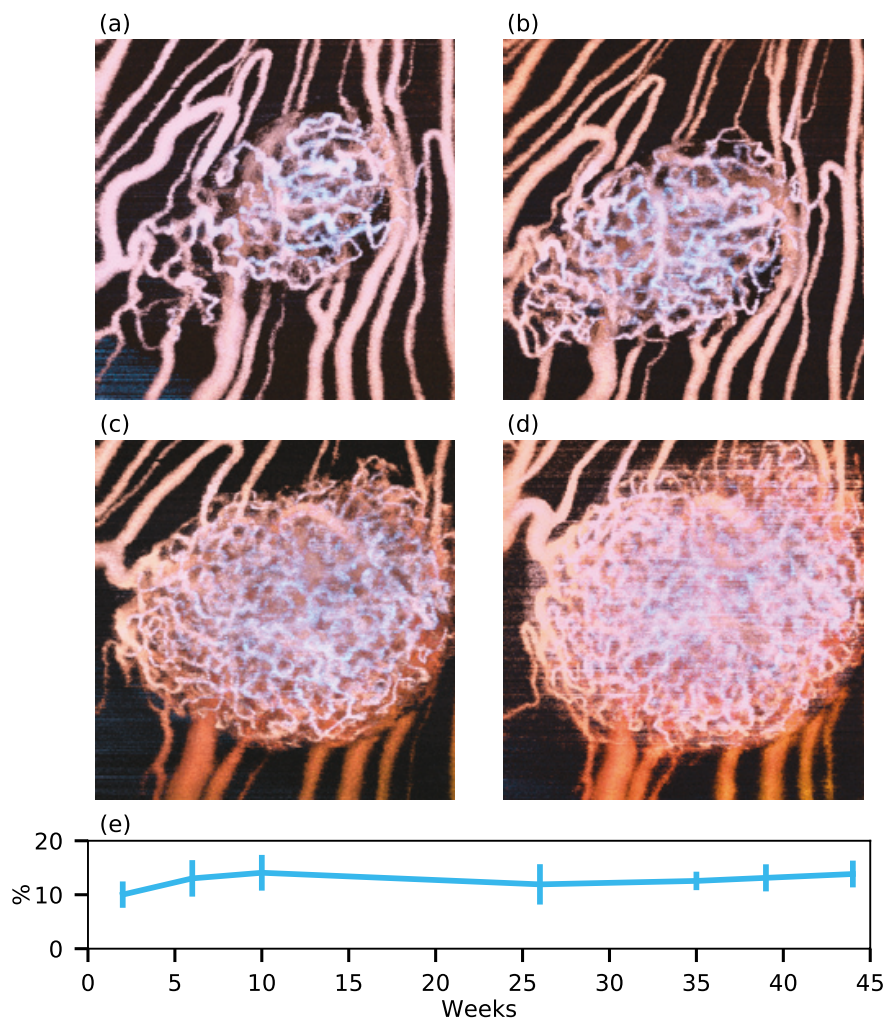


Figure 12: **a-d)** Growth of the vasculature of the islets of Langerhans over 44 weeks. **e)** Vascular density over time. Images and data from Ref. [17].

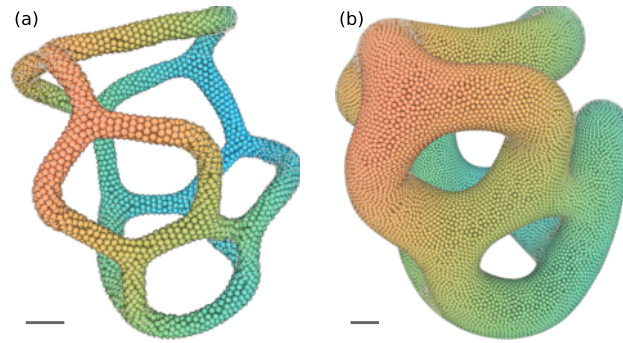


Figure 13: Vessel growth by proliferation in the absence of PCP ($\lambda_3 = 0$). In this case, growth occurs primarily by thickening of the vessels. Vascular density is clearly not preserved.

a growth rate of $\nu = 2.5 \times 10^{-6}$ is shown, having reached 15,000 cells. In panel c, we have allowed cells to divide at a growth rate of $\nu = 5.0 \times 10^{-5}$. This structure has thus reached 45,000 cells at a similar effective radius and thus much higher density. In panel d, we show the time evolution of the density under growth at different growth rates. Note that the time variable (horizontal axis) has been rescaled by the growth rate ν so as to be comparable across structures grown at different rates. This shows that rapidly increasing, decreasing or even approximately constant-density growth is possible with the same underlying mechanism. Furthermore, the figure clearly shows a certain growth-induced buckling which produces a tortuous structure consistent with what was shown in the islets of Langerhans in Figure 12.

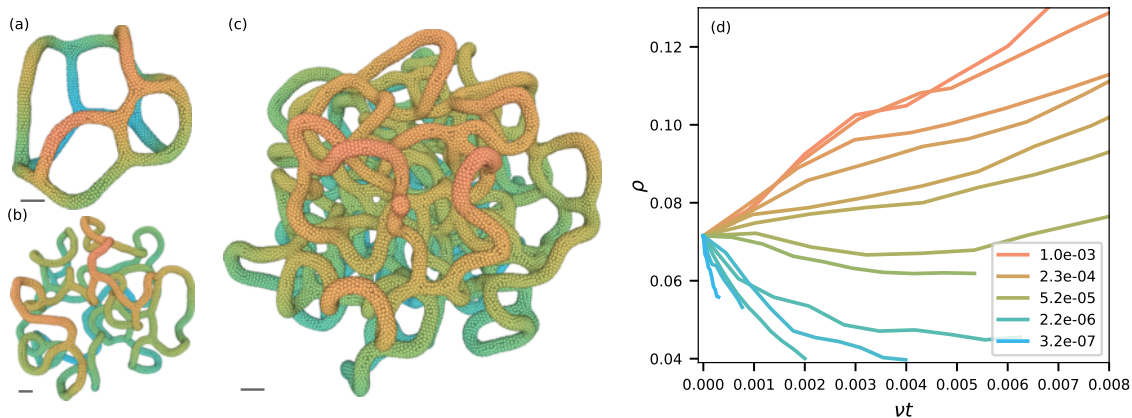


Figure 14: Growth of a vascular network *with* planar cell polarity. **a)** The initial, metastable structure. **b)** A structure resulting from a growth rate of $\nu = 2.5 \times 10^{-6}$, having reached 15,000 cells. **c)** A structure resulting from a growth rate of $\nu = 5.0 \times 10^{-5}$. The structure has reached 45,000 cells at a similar effective radius as that of panel c, but at a much higher density. **d)** Time evolution of the density under growth at different growth rates. The time variable has been rescaled by the growth rate ν so as to be comparable across structures grown at different rates.

1.3 DISCUSSION

This concludes the chapter on the generation of biological shapes. However, there are many open questions in this field. Most computational models in morphogenesis have been vertex or continuum models (implemented as e.g. finite-element simulations), and most have been two-dimensional. Some examples of three-dimensional vertex models exist [18] and a three-dimensional finite-element model of bud formation in lung tissue has also been published [19]. Perhaps the main advantage of a particle model, with the cells themselves (and their polarities) as the degrees of freedom, is how readily it lends itself to in-silico experiments. As we have seen, the relative simplicity of the approach means that it can incorporate multiple polarities, cell division and wedging, to name just a few elements. It is also straight-forward to extend it to couple to e.g. molecular concentration fields [15]. Furthermore, it can describe aggregation and self-assembly processes as well as dynamically

changing topologies – all of which are difficult to handle in continuum and vertex models.

A particularly interesting open question is whether there is a certain uniqueness in the mechanical polarized point-particle theories that one can write down for systems of polarized cells. One place to start would be with near-equilibrium configurations in which the deviations from the energetically optimal configuration can be treated as a small parameter in which to perform an expansion. Furthermore, symmetries severely restrict the possible terms that one could write down in a polarity-dependency factors such as the S of Equation (1.16). This classification is a work which we have started on, but it is not yet at the publication stage. If the space of reasonable theories is sufficiently restricted, one may hope to argue for a degree of universality in such polarized cell systems.

1.4 PUBLICATIONS FOR CHAPTER 1

The first chapter of this thesis builds on the following manuscripts. The papers were written during this degree and I have not submitted them for any other academic degree.

1. **B. F. Nielsen**, S. B. Nissen, K. Sneppen, J. Mathiesen, and A. Trusina, “Model to link cell shape and polarity with organogenesis,” *iScience* **23** (2020), no. 2, 100830.
2. J. B. Kirkegaard, **B. F. Nielsen**, A. Trusina, and K. Sneppen, “Self-assembly, buckling and density-invariant growth of three-dimensional vascular networks,” *Journal of the Royal Society Interface* **16** (2019), no. 159, 20190517.

MODEL TO LINK CELL SHAPE AND POLARITY WITH ORGANOGENESIS

Authors: Bjarke Frost Nielsen¹, Silas Boye Nissen¹, Kim Sneppen¹, Joachim Mathiesen¹ and Ala Trusina¹.

¹ The Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark.

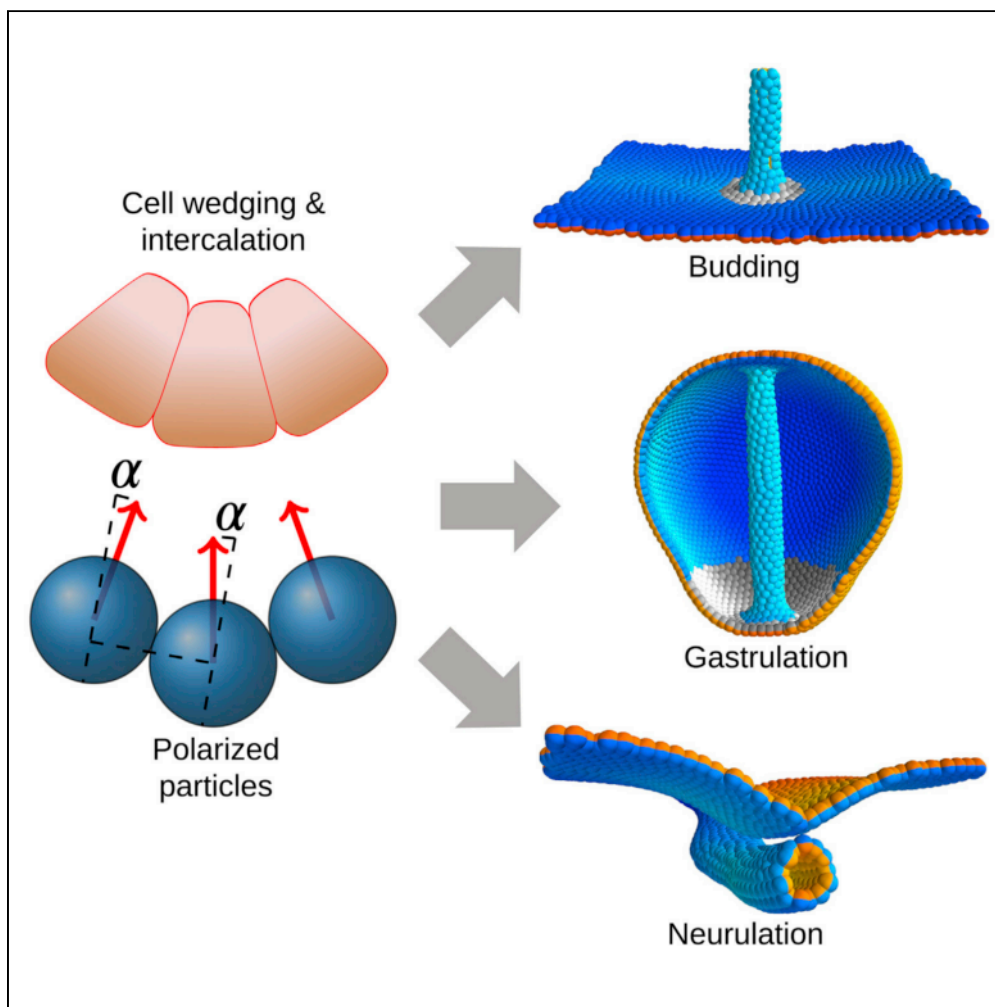
My contribution: Contributed to conceptualization and development, programming of the model, performing simulations and creating figures as well as writing of the manuscript.

Publication status: Published in *iScience* (2020).

Hyperlink(s): <https://doi.org/10.1016/j.isci.2020.100830> and [https://www.cell.com/iscience/fulltext/S2589-0042\(20\)30013-4](https://www.cell.com/iscience/fulltext/S2589-0042(20)30013-4)

Article

Model to Link Cell Shape and Polarity with Organogenesis



Bjarke Frost
Nielsen, Silas Boye
Nissen, Kim
Sneppen, Joachim
Mathiesen, Ala
Trusina

mathies@nbi.ku.dk (J.M.)
trusina@nbi.ku.dk (A.T.)

HIGHLIGHTS

Cell wedging and intercalation are modeled using a polarized point-particle approach

Cell intercalation is sufficient for tube budding

Tube budding is more robust when intercalation is complemented by wedging

Wedging and differential proliferation are sufficient for mammalian neurulation

Nielsen et al., iScience 23,
100830
February 21, 2020 © 2020 The
Authors.
[https://doi.org/10.1016/
j.isci.2020.100830](https://doi.org/10.1016/j.isci.2020.100830)



Article

Model to Link Cell Shape and Polarity with Organogenesis

Bjarke Frost Nielsen,¹ Silas Boye Nissen,¹ Kim Sneppen,¹ Joachim Mathiesen,^{1,*} and Ala Trusina^{1,2,*}

SUMMARY

How do flat sheets of cells form gut and neural tubes? Across systems, several mechanisms are at play: cells wedge, form actomyosin cables, or intercalate. As a result, the cell sheet bends, and the tube elongates. It is unclear to what extent each mechanism can drive tube formation on its own. To address this question, we computationally probe if one mechanism, either cell wedging or intercalation, may suffice for the entire sheet-to-tube transition. Using a physical model with epithelial cells represented by polarized point particles, we show that either cell intercalation or wedging alone can be sufficient and that each can both bend the sheet and extend the tube. When working in parallel, the two mechanisms increase the robustness of the tube formation. The successful simulations of the key features in *Drosophila* salivary gland budding, sea urchin gastrulation, and mammalian neurulation support the generality of our results.

INTRODUCTION

Early tubes in embryonic development—gut and neural tubes—form out of epithelial sheets. In mammalian neurulation and *Drosophila* gastrulation, the cell sheet wraps around the tube axis until the edges make contact and fuse. As a result of such *wrapping*, a tube is formed parallel to the cell layer. In sea urchin, the gut is formed orthogonal to the epithelial plane by *budding* out of the plane. Budding also appears to be a predominant form of tube formation in organ development (lungs and kidneys in vertebrates, salivary gland, and trachea in *Drosophila* [Andrew and Ewald, 2010]). The same key mechanisms drive both wrapping and budding sheet-to-tube transitions: changes in cell shape, contracting myosin cables spanning across cells, and convergent extension (CE) by directed cell intercalation (Andrew and Ewald, 2010; Chung et al., 2017). Cells change their shapes by adjusting their apical surfaces relative to their basal surfaces—apical constriction (AC) (Sawyer et al., 2010) or basal constriction (Gutzman et al., 2018; Visetsouk et al., 2018). In the following, we will refer to apical or basal constriction as *wedging* and directed cell intercalation as CE. In addition, oriented cell division and spatially restricted apoptosis (Andrew and Ewald, 2010) contribute to tubulogenesis in other systems.

Until recently, the consensus has been that wedging and CE each lead to distinct morphological transformations: wedging bends the sheet, and CE elongates the sheet and the eventual tube (Andrew and Ewald, 2010). Over decades, wedging was assumed to be a primary mechanism for invagination in budding (Paluch and Heisenberg, 2009). However, results by Sanchez-Corrales et al. (2018) show that wedging and radial CE are coupled, and both contribute to the invagination in *Drosophila* salivary gland. Furthermore, Nishimura et al. (2012) argue that in mammalian neurulation, CE and wedging are coupled through planar cell polarity (PCP). First, the direction of cell intercalations, orthogonal to the tube axis, is set by PCP. Second, wedging must be anisotropic—with a preferred direction parallel to PCP and the direction of intercalation—for the sheet to wrap into a tube and not a spherical lumen. This anisotropy may stem from the coupling between PCP and wedging, apical as well as basal constriction. This is supported by data at the molecular level (for neural tube closure [Nishimura et al., 2012; Ossipova et al., 2014], the midbrain-hindbrain boundary in zebrafish [Gutzman et al., 2018; Visetsouk et al., 2018], and gastrulation in *C. elegans* [Lee et al., 2006], sea urchin [Croce et al., 2006], and *Xenopus* [Choi and Sokol, 2009]). Although the role of anisotropic wedging has been well characterized in *Drosophila* gastrulation (Chanet et al., 2017; Guglielmi et al., 2015; Martin et al., 2010; Sweeton et al., 1991), the origins of the anisotropy are still being debated (Dobrovinski et al., 2018).

The recent developments open for new questions: *What are wedging and CE capable of on their own? Can invagination happen by CE alone? Is anisotropy in wedging essential for tubulogenesis, and, if so, when?*

¹Niels Bohr Institute, University of Copenhagen, Blegdamsvej 17, 2100 Copenhagen, Denmark

²Lead Contact

*Correspondence: mathies@nbi.ku.dk (J.M.), trusina@nbi.ku.dk (A.T.)

<https://doi.org/10.1016/j.isci.2020.100830>



In this paper, we introduce a theoretical model to address these questions. Theoretical models have been essential for understanding tubulogenesis. However, they are often limited to 2D and thus focus on either wedging or CE (Belmonte et al., 2016; Collinet et al., 2015; Spahn and Reuter, 2013). Although there are 3D models for budding and neurulation (Inoue et al., 2016; Kim et al., 2013), they lack the coupling between planar polarization, wedging, and CE and do not capture the entire sheet-to-tube transition. To close this gap, we introduce a model of polarized cell-cell interactions where cells are treated as point particles. As a starting point, we consider the model suggested in Nissen et al. (2018), which was used to study polarized adhesion. We use term *polarized adhesion* to refer to the cell-cell interaction where adhesion proteins are either apicobasally polarized (AB) or planar polarized by, e.g., PCP. The model parts describing PCP are not limited to the PCP pathway but can be applied to systems where planar polarity is induced by other pathways (e.g., polarized Baz/Par3 in *Drosophila* germband extension [Paré et al., 2014] or salivary gland budding [Sanchez-Corrales et al., 2018]). The model in Nissen et al. (2018), however, could not explicitly account for changes in cell shapes. Here, we show that the effect of cell wedging can be very simply modeled within a point-particle representation by modifying cell-cell forces to favor a tilt in AB polarities.

In line with the proposition by Chung et al. (2017), simulations show that, although CE alone can lead to a budding transition, it is less reliable, with frequent failure of invagination and even evagination. Our results suggest that isotropic wedging orients the budding process and allows for robust invagination. When applied to wrapping in neurulation, we find that anisotropic wedging alone was insufficient for final tube closure. However, closure as well as tube separation from the epithelium can be aided by differential proliferation. Furthermore, we find that anisotropic wedging on its own may be sufficient for tube elongation. Together, our results support the mutual complementarity of wedging and CE in bending and elongation.

RESULTS

To investigate the role of cell wedging in budding and wrapping, we aimed at capturing both isotropic and anisotropic (PCP-driven) wedging with as few parameters as possible.

Modeling Wedging of a Point Particle by Favoring Tilt in AB

Apical constriction leads to cell wedging and, as a consequence, the AB axes of neighboring cells become tilted toward the wedged cell (Figures 1B and 1C). In Nissen et al. (2018), a flat epithelial sheet was modeled by a cell-cell interaction force favoring parallel AB polarities in neighboring cells (Figure 1A, Equation S1 in the Transparent Methods). To model the effect of wedging, we modify the force to favor AB polarity vectors \mathbf{p}_i in neighbor cells to tilt toward the wedged cell (Figures 1B and 1C). That is, when the force is calculated, we replace \mathbf{p}_i by $\tilde{\mathbf{p}}_i$ (Equations 1–3).

$$\tilde{\mathbf{p}}_i = \mathbf{p}_i \quad (\text{for no wedging}), \quad (\text{Equation 1})$$

$$\tilde{\mathbf{p}}_i \propto \mathbf{p}_i - \alpha \hat{\mathbf{r}}_{ij} \quad (\text{for isotropic wedging}), \quad (\text{Equation 2})$$

$$\tilde{\mathbf{p}}_i \propto \mathbf{p}_i - \alpha (\hat{\mathbf{q}})_{ij} \quad (\text{for anisotropic wedging}). \quad (\text{Equation 3})$$

Here, $\hat{\mathbf{r}}_{ij}$ is the normalized displacement vector between cells i and j , whereas $(\hat{\mathbf{q}})_{ij}$ is the averaged PCP vector of the two interacting particles.

This change required only one parameter, α , setting the extent of the tilt (large α corresponds to pronounced wedging). If the wedging is isotropic, i.e., equally pronounced in all directions (Sanchez-Corrales et al., 2018), all neighbors to the wedged cell tend to tilt equally. In neurulation, the wedging is anisotropic: the wedging happens primarily parallel to the cell's PCP and perpendicular to the axis of the tube (Nishimura et al., 2012). To capture this PCP-directed anisotropy, we couple the direction of AB tilting to the orientation of the cell's PCP (Equation 3, Figure 1C). See the Transparent Methods section for details of the model and simulations.

Note that we aim only to capture the effects of wedging-PCP coupling and not the molecular mechanism. Also, in an attempt to generalize our results, we focus on a minimal set of conditions necessary for the outcome.

We first consider the complementary roles of CE and wedging in budding.

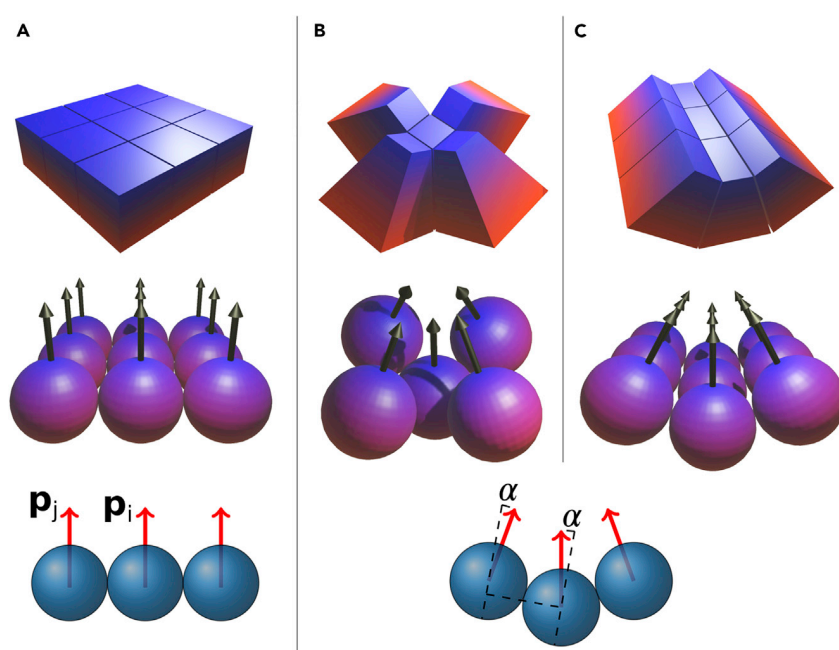


Figure 1. Wedging Is Introduced through a Cell-Cell Interaction that Favors Tilted AB Polarity Vectors

α is the extent of wedging. The blue-red gradient indicates the apical-basal axis.

(A) No wedging ($\alpha=0$), AB polarities (arrows) tend to be parallel.

(B) With isotropic wedging, the tilt α is the same in all directions.

(C) With anisotropic wedging, the tilt has a preferred direction. Blue and red signify, respectively, basal and apical surfaces. p_i and p_j are the AB polarities of cells i and j .

See also [Figure S3](#) and [S9](#), [Video S1](#).

Complementary and Unique Roles of CE and Wedging in Budding

Results by [Sanchez-Corrales et al. \(2018\)](#) and [Chung et al. \(2017\)](#) suggest that both wedging and CE contribute to invagination. However, computational models have generally focused on either wedging as a driver for invagination or CE as a driver of tissue elongation ([Belmonte et al., 2016](#); [Collinet et al., 2015](#); [Spahn and Reuter, 2013](#)). To date, no computational models have managed to combine both mechanisms or probe the role of CE in invagination.

We set out to reproduce these experimental observations. The aim is to only capture the budding, leaving out the finer details of the *Drosophila* salivary gland, such as off-center invagination. We start with a flat sheet of AB polarized cells. Motivated by the possible link between organizing signals (e.g., WNT), PCP, and wedging ([Habib et al., 2013](#); [Loh et al., 2016](#)), we define a region of “organizing signals” such that the cells within this region exhibit isotropic wedging and PCP. In *Drosophila* salivary glands, the apically constricting cells are distributed on a disk around the future center of the tube. With this configuration, we did not find parameters where both CE and wedging could act in parallel to form a well-defined tube [Figures S8A–S8C](#). However, a ring of basally constricting cells remedied this problem and allowed for wedging and CE to act in parallel. This was the case whether a disk of apically constricting cells was included ([Figures S8D–S8F](#)) or not ([Figure 2A](#)). Supporting this, the data by [Sanchez-Corrales et al. \(2018\)](#) suggest that there are basally constricting cells in the outer region of the placode. Furthermore, basal and apical constriction seems to be induced by the same organizing signal ([Gutzman et al., 2018](#)) through PCP pathways. Also, in sea urchin gastrulation, both types of wedging seem to be at play ([Kominami and Takata, 2004](#)). For simplicity, we limit our simulations to basal wedging, where basally constricting cells are distributed on a ring ([Figures 2A](#) and [S5](#)).

Our budding simulations thus show that successful invagination and tube elongation can proceed if both wedging and PCP (and thus CE) act in parallel ([Video S2](#), [Figures 2A–2C](#)). We have also succeeded in simulating sea urchin gastrulation where budding starts from a sphere of cells ([Figure 3](#), [Video S3](#), [Kimberly and](#)

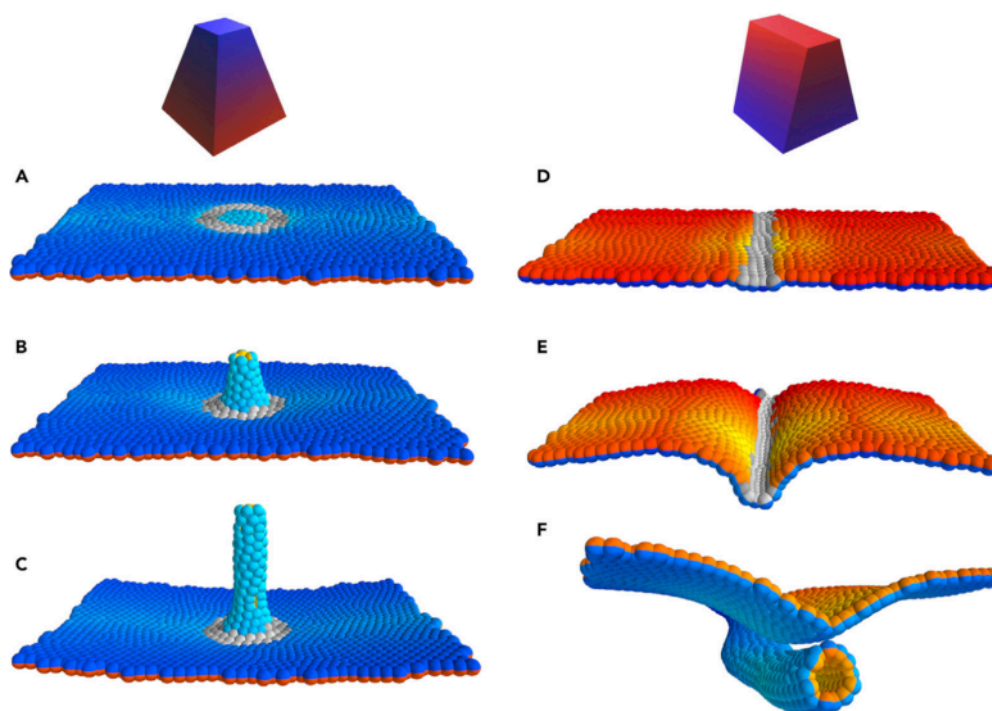


Figure 2. Isotropic and Anisotropic Wedging Drive Budding and Wrapping, Respectively

Wedging cells are labeled in gray, with a shading that indicates the PCP direction.

(A–C) Time evolution of budding simulation (similar to *Drosophila* salivary glands). Here, gray cells constrict basally and all cells on and inside the ring intercalate radially. The couplings are $(\lambda_1, \lambda_2, \lambda_3) = (0.5, 0.4, 0.1)$, the degree of wedging is $|\alpha| = 0.5$, and the annulus within which wedging occurs is given by the radii $r_0 = 5$ and $r_1 = 15$. See section *Modeling budding from a plane* for details, as well as [Figure S5](#). Total number of time steps was 6.25×10^4 at $dt = 0.2$. Snapshots correspond to times 5, 800, and 1.25×10^4 . The width of the Gaussian noise was $\sigma = 0.05$. See also [Figures S1, S7, and S8](#), [Video S2](#).

(D–F) Time evolution of wrapping simulation (similar to neurulation). Here, gray cells representing neuroepithelium constrict apically and constriction is anisotropic, follows the direction of PCP (Eq 3). Cells proliferate only at the gray/colored boundary (with 7-h doubling time), mimicking differential proliferation at the neuroepithelium/ectoderm boundary. The couplings are $(\lambda_1, \lambda_2, \lambda_3) = (0.6, 0.4, 0)$, the degree of wedging is $|\alpha| = 0.5$. See section *Modeling neurulation/wrapping* for details, as well as [Figure S4](#).

Total number of time steps was 3.9×10^4 at $dt = 0.1$, and snapshots were taken at times 5, 900, and 3.9×10^3 . The cell cycle length in simulation time units is 600. This simulation was run without added Gaussian noise, but noise is supplied by proliferation, which is implemented as a Poisson process. See also [Videos S4 and S5](#).

[Hardin, 1998](#); [Lyons et al., 2012](#)). This proceeds essentially like in the planar case (see [Transparent Methods](#) for details).

In addition, we find that budding can proceed without wedging if we allow for noise—random fluctuations in cell position and polarity orientation ([Figure S1](#), Equation S4 in the [Transparent Methods](#)). Even slight noise, with a width of less than a tenth of a cell radius, breaks the symmetry between the two sides of the plane and initiates the CE-driven tubulation in one of the two directions orthogonal to the plane. However, the robustness of the outcome decreases in two ways. First, the proportion of failed invaginations is higher ([Figure S1](#)). Second, the tube can form on either side of the epithelial plane.

Thus, it seems that the role of wedging is to aid in the initial invagination and ensure correct orientation. Interestingly, in the mutants where wedging is compromised, [Chung et al. \(2017\)](#) observe that, despite initial invagination in the right direction, the tubes form less reliably and sometimes reorient in the wrong direction. Our results, showing complementary roles of CE and wedging, are thus in line with the findings by [Sanchez-Corrales et al. \(2018\)](#) and [Chung et al. \(2017\)](#).

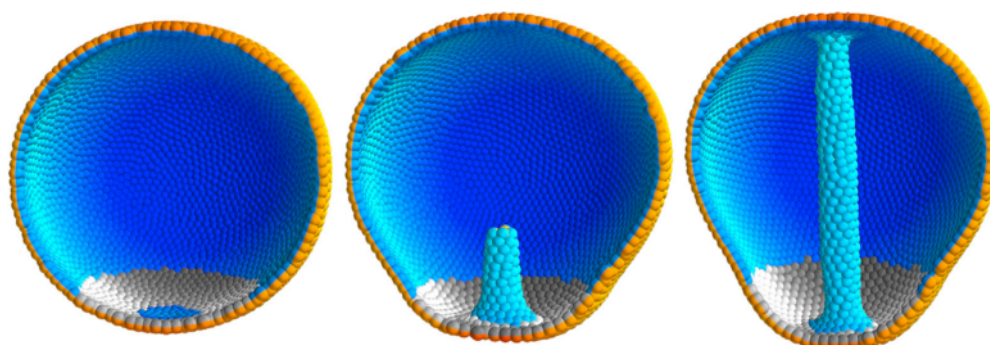


Figure 3. Isotropic Wedging in Conjunction with PCP Is Sufficient to Drive Sea Urchin Gastrulation without External Forcing

The gray ring shows cells with (isotropic) basal constriction, and the shading indicates the direction of planar cell polarity, which curls around the vertical axis in our simulation. The couplings are $(\lambda_1, \lambda_2, \lambda_3) = (0.5, 0.4, 0.1)$, the degree of wedging is $|\alpha| = 0.4$, and the annulus within which wedging occurs is given by the radii $r_0 = 7$ and $r_1 = 21$. See section *Modeling gastrulation* for details. Total number of time steps was 1.25×10^5 at $dt = 0.1$ and snapshots were taken at times 5, 1.5×10^3 , and 1.25×10^4 . The width of the Gaussian noise was $\sigma = 0.05$. See also [Video S3](#).

Cell shape change, intercalation, and tissue compression by supracellular myosin cables are also critical players in wrapping (Nishimura et al., 2012). The differences that cause some tubes to form parallel and others to form orthogonal to the epithelial plane appear to be encoded in the geometrical arrangement of the cells that participate in these three processes. In budding, such cells are arranged on a ring or a disk (circular symmetry), whereas in wrapping, they are arranged on a band (axial symmetry).

Anisotropic Wedging and Differential Proliferation Are Sufficient for Wrapping

To test if this difference in geometry alone is sufficient for wrapping, we choose a stripe of cells in the middle of the epithelial sheet to represent the neuroepithelium (NE) (shown as gray in [Figures 2D](#) and [2E](#)) and the remaining cells to represent ectoderm (E) (colored cells in [Figures 2D–2F](#)). The NE cells are then assigned anisotropic apical constriction and PCP pointing orthogonal to the future tube axis ([Figure S4](#)).

Wrapping Requires Anisotropy in Wedging

In the case of isotropic wedging, one would expect a collection of NE cells to eventually form a round invagination or spherical lumen—the minimum energy state ([Video S1](#)). If we impose isotropic wedging in our neurulation simulations, we obtain a bulging, rounded invagination, rather than a tube. See [Video S4](#).

Motivated by the results of Nishimura et al. (2012), showing that wedging is anisotropic ([Equation 3](#)) and cells wedge primarily in the direction orthogonal to the tube axis, we asked if anisotropic wedging can aid in tube closure. As expected, the tissue bends around the tube axis without capping at the ends of the tube ([Figures 1C](#) and [S2](#)).

Interestingly, anisotropic wedging also leads to cell intercalation by CE, narrowing, and elongating neuroepithelium (see [Figure S3](#)), thus supporting the link between PCP-driven wedging and cell intercalations. The simple, intuitive argument for this comes from how wedged cells pack in the tube. In the minimum energy state, the extent of wedging, α , determines how many cells can pack around the circumference of the tube ([Figures 1](#) and [S9](#)). If the cells do not change in size, fewer cells are needed to close the circumference as wedging increases. If there are more cells than the wedging can allow for, the “extra” cells will be displaced (to minimize energy). Because of the forces mediated by AB polarity (e.g., tight junctions), cells are constrained to move within the epithelium and are, as a result, displaced along the tube axis ([Figure S3](#)). CE-driven narrowing of the epithelium was proposed as necessary for tube closure (Wallingford et al., 2002). In our simulations, wedging and CE alone succeeded in bending the tissue in an axially symmetric fashion ([Figure S2](#)). However, we could not obtain successful tube closure even with maximal possible

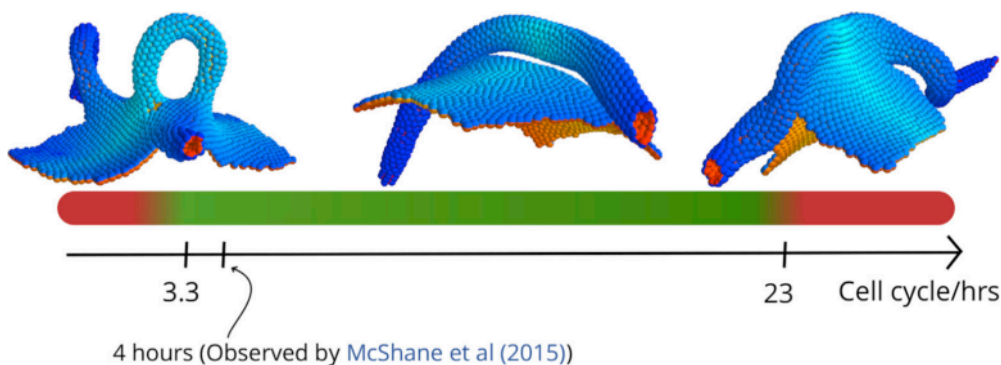


Figure 4. The Cell Cycle Length at the Neuroepithelial-Ectoderm Boundary Affects Tube Closure

For cell cycle lengths below 3.3 h and above 23 h the neural tube fails to close in our simulations. It should be noted that this broad interval also contains the cell cycle length of 4 h found for cells in the dorsolateral hinge points by [McShane et al. \(2015\)](#). The insets show outcomes of simulations run at short (2.6 h), intermediate (12 h), and long (26 h) cell cycle lengths. In simulation time, these correspond to 400, 1,800, and 4,000, respectively. See also [Figures S2 and S6](#).

CE and wedging (both tuned by the strength of α in [Equation 3](#)). This suggests that additional mechanisms are necessary for final tube closure.

Buckling by Proliferation at the NE Boundary Aids in Tube Closure

Images of neurulation cross-sections (see e.g., [Galea et al. \[2018\]](#)) show a sharp bending at the neuroepithelium-ectoderm (NE-E) boundary, with a curvature opposite to that inside the neuroepithelium (neural folds) ([Smith and Schoenwolf, 1997](#)). This is believed to be a result of combined forces from the ectoderm due to (1) change in cell shape (ectoderm cells become flatter and neuroepithelial cells become taller); (2) adhesion between basal surfaces of NE and E close to the neuroepithelium-ectoderm (NE-E) boundary ([Smith and Schoenwolf, 1997](#)), and (3) increase in cell density at this boundary either due to cell proliferation or intercalation ([McShane et al., 2015](#)).

Our goal was to test if the model can capture full tube closure with at least one of the mechanisms, so for simplicity, we focused on differential proliferation. When cells were set to proliferate only at the NE-E boundary ([McShane et al., 2015](#)), we found that the resulting buckling can lead to successful neural tube closure ([Video S5](#)). In the simulations, the out-of-equilibrium buckling created by rapid cell proliferation is necessary to create a narrow neck that allows epithelial folds to fuse. We find that tubulation is possible within a rather broad range of cell cycles (3–16 h). Shorter or longer cell cycles resulted in open-tube morphologies reminiscent of neural tube defects such as spina bifida ([Figure 4](#)). In both cases, the folds are too far apart to fuse, but for different reasons. If proliferation is too slow, the folds are far apart because the buckling is too weak.

On the other hand, when proliferation is too fast, the sheet does not have time to equilibrate, and CE does not catch up in narrowing it. Because of this, some sections of the tube become too wide to fuse. Interestingly this can sometimes lead to tube doubling/splitting ([Figure S6](#)).

The effect of slow proliferation in our simulations is in line with the experimental data. In [Copp et al. \(1988\)](#), it was shown that low proliferation rates could lead to neural tube defects in mice. In humans, mutations of the PAX3 transcription factor are implicated in Waardenburg syndrome ([Baldwin et al., 1994](#); [Tassabehji et al., 1993](#)) characterized by incomplete neural tube closure. The same transcription factor is essential in ensuring sufficient cell proliferation ([Wu et al., 2015](#)). The effect of increased (compared with wild-type) proliferation has not been addressed experimentally, and we hope that our predictions will motivate experiments in this direction.

DISCUSSION

Larger organisms rely on tubes for distributing nutrients across the body as well as for exocrine functions. How these tubes reliably form is an open question. A few recurrent mechanisms are known, e.g., directed

or differential proliferation, changes in cell shapes, supracellular myosin cables, polarized adhesion, and cell rearrangements. As evolution proceeds by tinkering rather than engineering, it is not surprising that these mechanisms have overlapping functions. Recently quantitative experiments (Chung et al., 2017; Nishimura et al., 2012; Sanchez-Corrales et al., 2018) enabled us to look beyond a “one mechanism, one function” relationship and toward a map of where mechanisms overlap and how they complement each other.

In this work, we have taken a step toward charting the functional overlap and complementarity among CE, wedging, and proliferation. A phenomenological point-particle representation allows us for the first time to combine PCP-driven cell intercalation (CE) and anisotropic wedging in thousands of cells in 3D and with a few free parameters. With this new tool we arrive at the following key results: First, our simulations show that CE can drive invagination in the absence of wedging, thus suggesting that this is a general mechanism that does not require forces from surrounding tissues. The invagination is, however, unreliable, and isotropic wedging plays a complementary role by setting the direction of invagination. The PCP pathway is not expressed in *Drosophila* salivary gland budding. One might therefore question why modeling the effects of planar polarity—and its role for CE—is valid in this system. However, despite differences at the molecular level, similarities emerge at the cellular level. At the cellular level, planar polarized adhesion is ubiquitous in systems undergoing CE: In mammalian neurulation, the adhesion protein Celsr is planar polarized by PCP (Nishimura et al., 2012); in early *Drosophila* development, Baz/Par3 is also planar polarized (by Toll receptors in gastrulation [Paré et al., 2014] and by unknown sources in salivary glands [Sanchez-Corrales et al., 2018]). Within our coarse-grained description of polar cell-cell interactions it is not necessary to differentiate whether the effects of planar polarization are due to PCP pathways or other sources, as long as polarized adhesion drives cell-cell intercalation. Also, we do not explicitly model the origins of planar polarity patterning, e.g., WNT signals orienting PCP (Humphries and Mlodzik, 2018) or Toll receptors orienting Baz/Par3 (Paré et al., 2014). Instead we pre-pattern the orientation of polarities directly. We can then either keep the orientation of planar polarities fixed, to simulate a global patterning by, e.g., Toll receptors, or let the global planar polarity pattern dynamically emerge from cell-cell interactions.

Second, our results predict that anisotropic, PCP-coupled wedging may play a role in tube formation and elongation. Our model predicts that anisotropy in wedging maintains axial symmetry of the tube during wrapping. Remarkably, anisotropic wedging can also lead to CE-like cell intercalation and, consequently, tube elongation. Although we have only tested the contribution of anisotropic wedging in wrapping, the same principle may apply in budding. In support of this, in budding, the initially isotropic wedging (Röper, 2012; Sanchez-Corrales et al., 2018) becomes anisotropic after the invagination, when the tube elongates (Pirraglia et al., 2010). Such an isotropic-to-anisotropic transition in wedging has been reported in *Drosophila* furrow formation (Leptin and Roth, 1994; Sweeton et al., 1991). Furthermore, visual inspection of tube cross-sections in the pancreas and kidneys suggests that cells are wedged. By analogy to neurulation, it is reasonable to expect wedging to be anisotropic in all tubes. It will be interesting to confirm this experimentally by, e.g., whole-mount 3D imaging of stained tubes.

Third, *differential proliferation* together with *anisotropic wedging* are sufficient for tube closure and separation in wrapping. Each of the mechanisms has to be spatially constrained. To buckle the cell sheet, proliferation must be faster at the neuroepithelium/ectoderm boundary than in the remaining tissue. Because only neuroepithelium forms the tube, anisotropic wedging must be localized to these cells. Differential proliferation has been proposed by McShane et al. (2015) as a mechanism for forming dorsolateral hinge points (DLHPs), regions where the tissue curvature has the same sign as at medial hinge points (MHPs). We find that modifying the extent of apical constriction or how it is distributed, i.e., throughout the entire neuroepithelium, or combinations of DLHPs and MHP, could not result in tube closure. Instead, our results highlight the importance of forming regions of opposite curvature at the boundaries. Our simulations suggest that differential proliferation buckles the boundaries and aids tube closure as it curves the epithelium oppositely to the curvature resulting from apical constriction.

Our simulations predict a wide range of proliferation rates capable of producing sufficient buckling for closure. These results call for testing for differential proliferation in systems without DLHPs (by accelerating or reducing proliferation rate in mutants or by molecular inhibitors [Li et al., 2017]). Although not

immediately feasible, it is also interesting to consider how to perturb the “opposite” curvature by interfering with differences in cell shapes or basal adhesion (Smith and Schoenwolf, 1997) of the neuroepithelium and ectoderm close to the boundary.

Models of tubulogenesis date back at least a few decades (Kerszberg and Changeux, 1998); however, most of them are limited to 2D and focus on either wedging or cell intercalation. Recently, Inoue et al. (2016) formulated a 3D vertex model of neurulation focusing on cell elongation, apical constriction, and active cell migration. The model does not include either cell proliferation or PCP but instead relies on active cell migration to pull the neural cells toward the midline. Although successful in bringing folds sufficiently close, it does not cover the separation of the tube from the sheet. In another system, the experimental and 3D modeling results by Osterfield et al. (2013) suggest that CE may be important in the early budding of the eggshell appendage. In their model, however, the initial invagination was driven by pre-patterned tension in the epithelium and neither cell polarity nor wedging were considered. Also, a recent 3D model of tube budding in the lung epithelium concluded that wedging can only result in rounded tubes and that it is insufficient to drive the entire process (Kim et al., 2013). Still, in that study, only isotropic wedging was considered. In our simulations, we see that anisotropy is necessary for tube formation.

We have demonstrated that cell wedging can be phenomenologically captured in a point-particle representation. This is not restricted to apical constriction but also covers, e.g., basal constriction, and can, in a similar spirit, be extended to capture changes in cell height and width. Also, adding oriented cell proliferation and local apoptosis is straight forward and could allow for modeling a wider range of tubulogenesis phenomena. Furthermore, we are now in a position to address tube branching in, e.g., lungs and vascularization, where cells forming the tubes also are the ones that secrete organizing signals that locally re-orient PCP polarities and may induce anisotropic changes in cell shapes.

Limitations of Study

A major limitation of our study is that we do not model the coupling of polarities to orienting morphogens (e.g., WNT, FGF, or BMP).

As a consequence, cell properties such as expression of apical-basal and planar cell polarity (and the orientation of polarities in individual cells) had to be assigned at the start of simulations. Furthermore, in the case of budding, the orientation of PCP had to be maintained fixed through the entire sheet-to-tube transition. We anticipate that, by including the morphogen-polarity coupling, the right distribution of cell types and polarity directions will emerge without externally imposed constraints.

METHODS

All methods can be found in the accompanying [Transparent Methods supplemental file](#).

DATA AND CODE AVAILABILITY

The source code for the simulations is available on GitHub (Nielsen, 2019).

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at <https://doi.org/10.1016/j.isci.2020.100830>.

ACKNOWLEDGMENTS

This research has received funding from the Danish National Research Foundation (grant number: DNRF116), the European Research Council under the European Union’s Seventh Framework Programme (FP/2007–2013)/ERC Grant Agreement number 740704, and VILLUM Foundation research grant 13168. The authors would like to thank Julius B. Kirkegaard for valuable discussions.

AUTHOR CONTRIBUTIONS

B.F.N. programmed and ran the model simulations and created figures; B.F.N., A.T., J.M., S.B.N., and K.S. outlined the paper, developed the model, contributed to discussions, and wrote the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: August 13, 2019

Revised: December 4, 2019

Accepted: January 7, 2020

Published: February 21, 2020

REFERENCES

- Andrew, D.J., and Ewald, A.J. (2010). Morphogenesis of epithelial tubes: Insights into tube formation, elongation, and elaboration. *Dev. Biol.* **341**, 34–55.
- Baldwin, C.T., Lipsky, N.R., Hoth, C.F., Cohen, T., Mamuya, W., and Milunsky, A. (1994). Mutations in *pax3* associated with waardenburg syndrome type i. *Hum. Mutat.* **3**, 205–211.
- Belmonte, J.M., Swat, M.H., and Glazier, J.A. (2016). Filopodial-tension model of convergent-extension of tissues. *PLoS Comput. Biol.* **12**, e1004952.
- Chanet, S., Miller, C.J., Vaishnav, E.D., Ermentrout, B., Davidson, L.A., and Martin, A.C. (2017). Actomyosin meshwork mechanosensing enables tissue shape to orient cell force. *Nat. Commun.* **8**, 15014.
- Choi, S.-C., and Sokol, S.Y. (2009). The involvement of lethal giant larvae and wnt signaling in bottle cell formation in xenopus embryos. *Dev. Biol.* **336**, 68–75.
- Chung, S., Kim, S., and Andrew, D.J. (2017). Uncoupling apical constriction from tissue invagination. *eLife* **6**, e22235.
- Collinet, C., Rauzi, M., Lenne, P.-F., and Lecuit, T. (2015). Local and tissue-scale forces drive oriented junction growth during tissue extension. *Nat. Cell Biol.* **17**, 1247.
- Copp, A.J., Brook, F.A., and Roberts, H.J. (1988). A cell-type-specific abnormality of cell proliferation in mutant (curly tail) mouse embryos developing spinal neural tube defects. *Development* **104**, 285–295.
- Croce, J., Duloquin, L., Lhomond, G., McClay, D.R., and Gache, C. (2006). Frizzled5/8 is required in secondary mesenchyme cells to initiate archenteron invagination during sea urchin development. *Development* **133**, 547–557.
- Dobrovinski, K., Tchoufag, J., and Mandadapu, K. (2018). A simplified mechanism for anisotropic constriction in *Drosophila* mesoderm. *Development* **145**, dev167387. <https://dev.biologists.org/content/145/24/dev167387>.
- Galea, G.L., Nychyk, O., Mole, M.A., Moulding, D., Savery, D., Nikolopoulou, E., Henderson, D.J., Greene, N.D., and Copp, A.J. (2018). *Vangl2* disruption alters the biomechanics of late spinal neurulation leading to spina bifida in mouse embryos. *Dis. Model. Mech.* **11**, dmm032219.
- Guglielmi, G., Barry, J.D., Huber, W., and De Renzis, S. (2015). An optogenetic method to modulate cell contractility during tissue morphogenesis. *Dev. Cell* **35**, 646–660.
- Gutzman, J.H., Graeden, E., Brachmann, I., Yamazoe, S., Chen, J.K., and Sive, H. (2018). Basal constriction during midbrain–hindbrain boundary morphogenesis is mediated by *wnt5b* and focal adhesion kinase. *Biol. Open* **7**, bio034520.
- Habib, S.J., Chen, B.-C., Tsai, F.-C., Anastasiadis, K., Meyer, T., Betzig, E., and Nusse, R. (2013). A localized wnt signal orients asymmetric stem cell division in vitro. *Science* **339**, 1445–1448.
- Humphries, A.C., and Mlodzik, M. (2018). From instruction to output: *wnt/pcp* signaling in development and cancer. *Curr. Opin. Cell Biol.* **51**, 110–116.
- Inoue, Y., Suzuki, M., Watanabe, T., Yasue, N., Tateo, I., Adachi, T., and Ueno, N. (2016). Mechanical roles of apical constriction, cell elongation, and cell migration during neural tube formation in xenopus. *Biomech. Model. Mechanobiol.* **15**, 1733–1746.
- Kerszberg, M., and Changeux, J.-P. (1998). A simple molecular model of neurulation. *BioEssays* **20**, 758–770.
- Kim, H.-Y., Varner, V.D., and Nelson, C.M. (2013). Apical constriction initiates new bud formation during monopodial branching of the embryonic chicken lung. *Development* **140**, 3146–3155.
- Kimberly, E.L., and Hardin, J. (1998). Bottle cells are required for the initiation of primary invagination in the sea urchin embryo. *Dev. Biol.* **204**, 235–250.
- Kominami, T., and Takata, H. (2004). Gastrulation in the sea urchin embryo: a model system for analyzing the morphogenesis of a monolayered epithelium. *Dev. Growth Differ.* **46**, 309–326.
- Lee, J.-Y., Marston, D.J., Walston, T., Hardin, J., Halberstadt, A., and Goldstein, B. (2006). *Wnt/frizzled* signaling controls *C. elegans* gastrulation by activating actomyosin contractility. *Curr. Biol.* **16**, 1986–1997.
- Leptin, M., and Roth, S. (1994). Autonomy and non-autonomy in *Drosophila* mesoderm determination and morphogenesis. *Development* **120**, 853–859.
- Li, Y., Muffat, J., Omer, A., Bosch, I., Lancaster, M.A., Sur, M., Gehrke, L., Knoblich, J.A., and Jaenisch, R. (2017). Induction of expansion and folding in human cerebral organoids. *Cell Stem Cell* **20**, 385–396.
- Loh, K.M., van Amerongen, R., and Nusse, R. (2016). Generating cellular diversity and spatial form: wnt signaling and the evolution of multicellular animals. *Dev. Cell* **38**, 643–655.
- Lyons, D.C., Kaltenbach, S.L., and McClay, D.R. (2012). Morphogenesis in sea urchin embryos: linking cellular events to gene regulatory network states. *Wiley Interdiscip. Rev. Dev. Biol.* **1**, 231–252.
- Martin, A.C., Gelbart, M., Fernandez-Gonzalez, R., Kaschube, M., and Wieschaus, E.F. (2010). Integration of contractile forces during tissue invagination. *J. Cell Biol.* **188**, 735–749.
- McShane, S.G., Molè, M.A., Savery, D., Greene, N.D., Tam, P.P., and Copp, A.J. (2015). Cellular basis of neuroepithelial bending during mouse spinal neural tube closure. *Dev. Biol.* **404**, 113–124.
- Nielsen, B.F. (2019). OrganogenesisPCP. <https://github.com/BjarkeFN/OrganogenesisPCP>.
- Nishimura, T., Honda, H., and Takeichi, M. (2012). Planar cell polarity links axes of spatial dynamics in neural-tube closure. *Cell* **149**, 1084–1097.
- Nissen, S.B., Rønild, S., Trusina, A., and Sneppen, K. (2018). Theoretical tool bridging cell polarities with development of robust morphologies. *eLife* **7**, e38407.
- Ossipova, O., Kim, K., Lake, B.B., Itoh, K., Ioannou, A., and Sokol, S.Y. (2014). Role of *rab11* in planar cell polarity and apical constriction during vertebrate neural tube closure. *Nat. Commun.* **5**, 3734.
- Osterfield, M., Du, X., Schüpbach, T., Wieschaus, E., and Shvartsman, S.Y. (2013). Three-dimensional epithelial morphogenesis in the developing *Drosophila* egg. *Dev. Cell* **24**, 400–410.
- Paluch, E., and Heisenberg, C.-P. (2009). Biology and physics of cell shape changes in development. *Curr. Biol.* **19**, R790–R799.
- Paré, A.C., Vichas, A., Fincher, C.T., Mirman, Z., Farrell, D.L., Mainieri, A., and Zallen, J.A. (2014). A positional toll receptor code directs convergent extension in *Drosophila*. *Nature* **515**, 523.
- Pirraglia, C., Walters, J., and Myat, M.M. (2010). Pak1 control of e-cadherin endocytosis regulates salivary gland lumen size and shape. *Development* **137**, 4177–4189.
- Röper, K. (2012). Anisotropy of crumbs and *apkc* drives myosin cable assembly during tube formation. *Dev. Cell* **23**, 939–953.
- Sanchez-Corrales, Y.E., Blanchard, G.B., and Röper, K. (2018). Radially patterned cell behaviours during tube budding from an epithelium. *eLife* **7**, e35717.

Sawyer, J.M., Harrell, J.R., Shemer, G., Sullivan-Brown, J., Roh-Johnson, M., and Goldstein, B. (2010). Apical constriction: a cell shape change that can drive morphogenesis. *Dev. Biol.* **341**, 5–19.

Smith, J.L., and Schoenwolf, G.C. (1997). Neurulation: coming to closure. *Trends Neurosci.* **20**, 510–517.

Spahn, P., and Reuter, R. (2013). A vertex model of drosophila ventral furrow formation. *PLoS One* **8**, e75051.

Sweeton, D., Parks, S., Costa, M., and Wieschaus, E. (1991). Gastrulation in *Drosophila*: the formation of the ventral furrow and posterior midgut invaginations. *Development* **112**, 775–789.

Tassabehji, M., Read, A.P., Newton, V.E., Patton, M., Gruss, P., Harris, R., and Strachan, T. (1993). Mutations in the *pax3* gene causing waardenburg syndrome type 1 and type 2. *Nat. Genet.* **3**, 26.

Visetsook, M.R., Falat, E.J., Garde, R.J., Wendlick, J.L., and Gutzman, J.H. (2018). Basal epithelial tissue folding is mediated by differential

regulation of microtubules. *Development* **145**, dev167031.

Wallingford, J.B., Fraser, S.E., and Harland, R.M. (2002). Convergent extension: the molecular control of polarized cell movement during embryonic development. *Dev. Cell* **2**, 695–706.

Wu, T.-F., Yao, Y.-L., Lai, I.-L., Lai, C.-C., Lin, P.-L., and Yang, W.-M. (2015). Loading of *pax3* to mitotic chromosomes is mediated by arginine methylation and associated with Waardenburg syndrome. *J. Biol. Chem.* **290**, 20556–20564.

iScience, Volume 23


Supplemental Information

Model to Link Cell Shape and Polarity with Organogenesis

Bjarke Frost Nielsen, Silas Boye Nissen, Kim Sneppen, Joachim Mathiesen, and Ala Trusina

Supplementary Figures

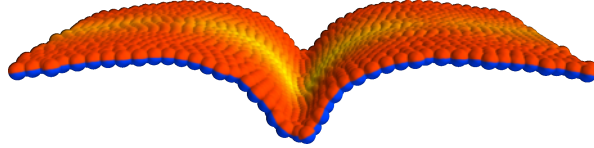
Figure S1 Budding outcomes in the absence of wedging. Related to Fig 2.



	Normal	Failed	Misoriented
High Noise $\sigma = 0.1$ $N = 50$	40%	0%	60%
Low Noise $\sigma = 0.002$ $N = 50$	20%	50%	30%
No Noise $\sigma = 0$ $N = 50$	0%	100%	0%

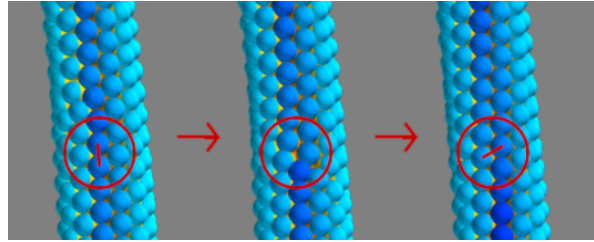
Budding outcomes without wedging at high and low noise as well as in the absence of noise. The first column shows the proportion of normal initiations of tubulation, the middle column shows failed invaginations while the last column shows evaginations. σ is the width of the Gaussian noise, while N is the number of simulations run at the given noise level. See the Methods section for details on the implementation of noise. In all cases $dt = 0.1$. The couplings were kept at $(\lambda_1, \lambda_2, \lambda_3) = (0.4, 0.5, 0.1)$ and the annulus within which wedging occurs is given by the radii $r_0 = 5$ and $r_1 = 10$. Since wedging is absent, $\alpha = 0$.

Figure S2 Lack of proliferation. Related to Fig 4.



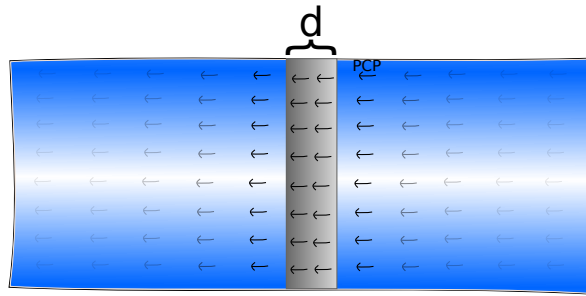
The fate of the neural sheet in our simulations in the absence of proliferation. Here the couplings are $(\lambda_1, \lambda_2, \lambda_3) = (0.6, 0.4, 0)$, the degree of wedging is $|\alpha| = 0.5$. See the section *Modeling neurulation/wrapping* for details. Total number of time steps was 1.4×10^5 at $dt = 0.1$. The simulation was run without noise.

Figure S3 T1 transition induced by wedging. Related to Fig 1.



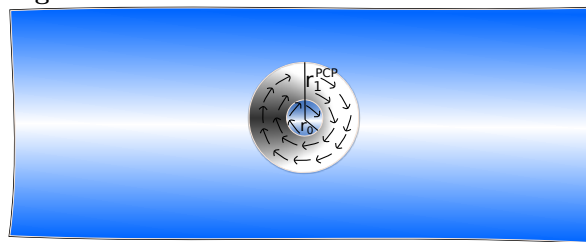
The T1 transition was induced by starting with a tube which was stabilized with anisotropic wedging of strength $|\alpha| = 0.3$ and then increasing the extent of wedging to $|\alpha| = 0.5$, causing the structure to tighten and elongate by intercalation. The couplings are $(\lambda_1, \lambda_2, \lambda_3) = (0.55, 0.45, 0)$ and the width of the Gaussian noise is 0.1 with time step size $dt = 0.2$.

Figure S4 The initial configuration of the cell sheet for neurulation. Related to Fig 2.



The initial configurations of the cell sheet for neurulation. Wedging is turned on in a band of width d (gray) with PCP running orthogonal to this band.

Figure S5 The initial configuration of the cell sheet for budding. Related to Fig 2.



The initial configurations of the cell sheet for budding. Wedging is turned on in an annulus (gray) where PCP curls around tangentially.

Figure S6 Tube splitting observed with excessive proliferation rate. Related to Fig 4. The proliferation rate corresponds to a cell cycle length of 1.5h for cells at the neuroepithelium/ectoderm boundary. The remaining parameters are as in the main neurulation simulation, as described in Fig 2

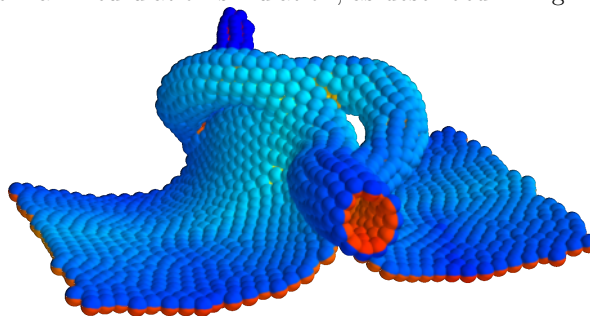
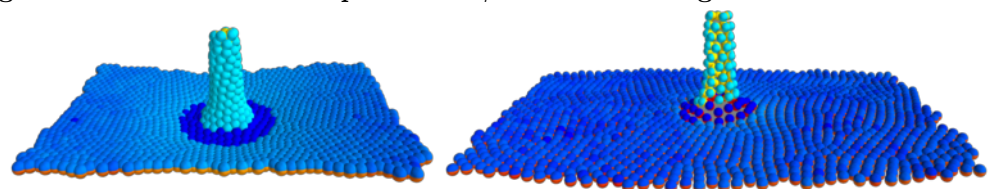


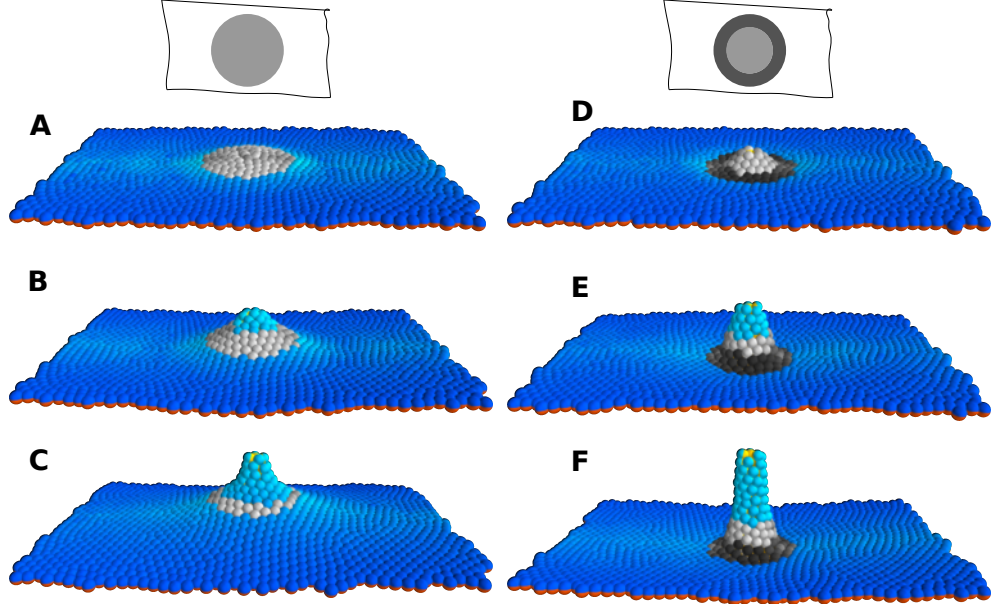
Figure S7 Influence of the parameter β . Related to Fig 2.



Budding simulations run with $\beta = 2.5$ (left) and $\beta = 10$ (right). This affects the equilibrium distance so that cells are closer together resp. further apart (and thus come across as larger resp. smaller) but budding progresses in a qualitatively similar manner.

The remaining simulations in this paper were all run with $\beta = 5$ ensuring an equilibrium distance of $d_{eq} = 2$.

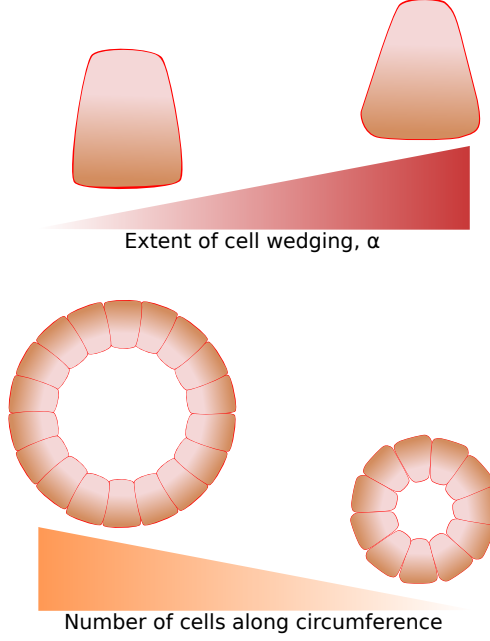
Figure S8 Apical constriction in budding. Related to Fig 2.



(**A-C**) Time evolution of budding simulation when only a disk of apically constricting cells (light gray) are assigned, and no basally constricting cells. The couplings are $(\lambda_1, \lambda_2, \lambda_3) = (0.5, 0.4, 0.1)$, the degree of wedging is $|\alpha| = 0.3$. The radius of the disk of apically constricting cells is given by $r_0 = 10$. Total number of time steps was 6.8×10^4 at $dt = 0.1$. Snapshots correspond to times 175, 600 and 6800.

(**D-F**) Time evolution of budding simulation when a disk of apically constricting cells (light gray) as well as a ring of basally constricting cells (dark gray) are assigned. The couplings are $(\lambda_1, \lambda_2, \lambda_3) = (0.5, 0.4, 0.1)$, the degree of wedging is $|\alpha| = 0.3$. The outer radius of the ring for which basal constriction occurs is given by $r_1 = 10$ while the radius of the disk of apically constricting cells is given by $r_0 = 5$. Total number of time steps was 2.2×10^4 at $dt = 0.1$. Snapshots correspond to times 25, 400 and 2200.

Figure S9 The degree of wedging affects the circumference of the tube. Related to Fig 1.



Transparent Methods

Model

Following Nissen et al. (2018), cells are treated as point particles interacting with neighboring cells through a pair-potential V_{ij} . The potential has a rotationally symmetric repulsive term and a polarity-dependent attractive term. In terms of r_{ij} (the distance between two cells i and j), the dimensionless potential can be formulated as

$$V_{ij} = e^{r_{ij}} - [\lambda_1 S_{ij}(A) + \lambda_2 S_{ij}(AP) + \lambda_3 S_{ij}(P)] e^{-r_{ij}/\beta}. \quad (S1)$$

The parameter β has the fixed value $\beta = 5$, since this ensures that the equilibrium distance is always 2, corresponding to 2 cell radii. In Figure S7 we have shown that one can obtain qualitatively similar results at other values of β . The parameters λ_i are coupling constants which define the strength of polar interactions in the model. $S_{ij}(A)$ gives the form of the interaction between AB polarity and position, whereas $S_{ij}(AP)$ and $S_{ij}(P)$ give the coupling of PCP with AB and position, respectively, as described in Nissen et al. (2018). These couplings are formulated in terms of AB vectors \mathbf{p}_i , PCP vectors \mathbf{q}_i and a unit vector $\hat{\mathbf{r}}_{ij}$ from cell i to j . The coupling $S_{ij}(AP) = (\mathbf{p}_i \times \mathbf{q}_i) \cdot (\mathbf{p}_j \times \mathbf{q}_j)$ dynamically maintains the orthogonality of the PCP unit vectors \mathbf{q}_i and \mathbf{q}_j to their corresponding AB polarity vectors while lateral organization is favored by $S_{ij}(P) = (\hat{\mathbf{r}}_{ij} \times \mathbf{q}_i) \cdot (\hat{\mathbf{r}}_{ij} \times \mathbf{q}_j)$. In the absence of any cell shape effects, the coupling between AB and position is given by $S_{ij}(A) = (\hat{\mathbf{r}}_{ij} \times \mathbf{p}_i) \cdot (\hat{\mathbf{r}}_{ij} \times \mathbf{p}_j)$, which favors a flat cell sheet. Wedging of cells is introduced into our model by a single deformation parameter α , which describes an attractive interaction between the AB polarity unit vectors \mathbf{p}_i and \mathbf{p}_j :

$$S_{ij}(A) = (\hat{\mathbf{r}}_{ij} \times \tilde{\mathbf{p}}_i) \cdot (\hat{\mathbf{r}}_{ij} \times \tilde{\mathbf{p}}_j), \quad (S2)$$

where $\tilde{\mathbf{p}}_i$ is given by

$$\begin{aligned}\tilde{\mathbf{p}}_i &= \mathbf{p}_i \quad (\text{for no wedging}), \\ \tilde{\mathbf{p}}_i &= \frac{\mathbf{p}_i - \alpha \hat{\mathbf{r}}_{ij}}{|\mathbf{p}_i - \alpha \hat{\mathbf{r}}_{ij}|} \quad (\text{for isotropic wedging}), \\ \tilde{\mathbf{p}}_i &= \frac{\mathbf{p}_i - \alpha \langle \hat{\mathbf{q}} \rangle_{ij}}{|\mathbf{p}_i - \alpha \langle \hat{\mathbf{q}} \rangle_{ij}|} \quad (\text{for anisotropic wedging}).\end{aligned}\tag{S3}$$

Here, $\langle \hat{\mathbf{q}} \rangle_{ij}$ denotes the mean of PCP vectors \mathbf{q}_i and \mathbf{q}_j belonging to the two interacting cells. The above substitution, $\mathbf{p}_i \rightarrow \tilde{\mathbf{p}}_i$, is only performed in $S_{ij}(A)$, so as to only affect the coupling between AB polarity and position.

Setting $\alpha = 0$ favors a flat sheet (see Fig 1A–B) whereas a non-zero α favors bending of AB polarity vectors towards (or away from) one another and induces curvature in a sheet of cells (Fig 1C–D).

The time development is simulated by overdamped (relaxational) dynamics along the gradient of the above potential, Eq (S1):

$$\begin{aligned}\frac{\partial \mathbf{r}_i}{\partial t} &= -\frac{\partial V_i}{\partial \mathbf{r}_i} + \eta, \\ \frac{\partial \mathbf{p}_i}{\partial t} &= -\frac{\partial V_i}{\partial \mathbf{p}_i} + \eta, \\ \frac{\partial \mathbf{q}_i}{\partial t} &= -\frac{\partial V_i}{\partial \mathbf{q}_i} + \eta,\end{aligned}\tag{S4}$$

where the potential energy function for the i 'th cell is $V_i = \sum_j V_{ij}$. The sum runs over those cells j which are within direct line of sight of the i 'th cell as described in Nissen et al. (2018). η is a noise term corresponding to Gaussian white noise with vanishing mean. This noise term provides a degree of randomness to cell position as well as the orientation of polarities. Cell division (when present) is modeled as a Poisson process with daughter cells being placed randomly around the mother cell at a distance of one cell radius.

The model was implemented in Python using PyTorch for automatic differentiation (Paszke et al. 2017). Numerical integration of the equations of motion is implemented through the Euler method, usually with $dt = 0.1$. We have checked that the model converges to similar results (tested for budding) with $dt = 10^{-4}$. The source code for the simulations is available on GitHub (Nielsen 2019).

Parameter estimation and robustness

We have tested the robustness of our approach on a number of model cases and find that, for example, *budding* can be reproduced with a broad range of wedging parameters, $\alpha \in [0.1, 0.6]$ and for diverse PCP coupling strengths $\lambda_3 \in [0.8, 0.14]$. For these intervals, the budding is qualitatively similar to that illustrated in Fig 2A. Our typical values of wedging used in simulations, $\alpha \in [0.3, 0.5]$ are comparable with the wedging strains reported in Sanchez-Corrales et al. (2018), e.g. $0.03pp/\mu\text{m}$, corresponding to $\alpha = 0.4$ (assuming a cell diameter of $13\mu\text{m}$) (Brown & Bron 1987).

We further explore our model by re-instating dimensions in the formulation of the potential and the equation of motion and estimating dimensionful quantities. With dimensions reinstated, the pair-potential takes the form

$$V_{ij} = V_0 [\exp(-r/\ell) - S \exp(-r/(\beta\ell))].\tag{S5}$$

The overdamped equation of motion (without noise) becomes

$$0 = \gamma \mathbf{v}_i + \frac{\partial V_{ij}}{\partial \mathbf{r}_i}, \quad (\text{S6})$$

where $\mathbf{v}_i = \partial \mathbf{r}_i / \partial t$. We now introduce dimensionless (tilded) parameters

$$V_{ij} = V_0 \tilde{V}_{ij}, \quad \mathbf{r}_i = \ell \tilde{\mathbf{r}}_i, \quad v_i = v_0 \tilde{v}_i = \frac{\ell}{t_0} \tilde{v}_i. \quad (\text{S7})$$

and insert the dimensionless parameters in our equation of motion

$$\tilde{\mathbf{v}}_i = -\frac{V_0}{\ell \gamma v_0} \frac{\partial \tilde{V}_{ij}}{\partial \tilde{\mathbf{r}}_i}. \quad (\text{S8})$$

Inserting the dimensionless equation of motion, this reduces to $V_0 = \ell \gamma v_0$. In Eskandari & Salcudean (2008), a typical value for the dynamical viscosity μ was reported to be on the order of 250 Pa.s. This can be related to the coefficient γ by Stokes' Law of viscous drag, $\gamma = 6\pi\mu\ell$. We now compare our model with epithelial cell extrusion and use the typical cell speed reported in Yamada et al. (2017), $v_0 \approx 1 \text{ mm h}^{-1}$ and use the typical cell size reported in Brown & Bron (1987), $2\ell = 13 \mu\text{m}$. With these numbers, our model predicts a typical extrusion energy on the order of

$$12V_0 \approx 12 \times 6\pi\mu\ell^2 v_0 \approx 2 \times 10^{-13} \text{ J}. \quad (\text{S9})$$

The factor of $12 = 2 \times 6$ is due to the hexagonal structure of the cell sheet. Note that our estimate of the extrusion energy is consistent with the finding in Yamada et al. (2017) for epithelial cell extrusion. Here, an actomyosin ring is measured to exhibit a contraction force of the order of 1 kPa, which results in an extrusion energy of the order $1 \text{ kPa} \times \ell^3 \approx 3 \times 10^{-13} \text{ J}$.

With these identifications of parameters, it is possible to extract dimensionful quantities from our simulations. This is what allows for e.g. the computation of cell cycle lengths in Fig 4.

We anticipate that the values of the couplings λ_i can be estimated from the extent and speed of CE (e.g in our model these would be determined by the values of λ_3 relative to λ_1).

Modeling neurulation/wrapping

The starting point for our simulation of neurulation is a planar sheet of cells where a line with a width of six cell radii is given non-zero wedging strength $|\alpha| = \alpha_0 > 0$ and all other cells have $\alpha = 0$. The line is centered at $x = 0$ and PCP is initialized orthogonally to this line, along the x direction ($\mathbf{q}|_{t=0} = \hat{\mathbf{x}}$). See Figure S4.

Cell proliferation is simulated as a Poisson process by choosing a rate Γ for *each cell* to divide in each time unit. Only cells at the neuroepithelium-ectoderm boundary (defined as cells with $|\alpha| > 0$ who are neighbours of cells with $\alpha = 0$) proliferate (with rate $\Gamma = \Gamma_0 > 0$) while the rest have $\Gamma = 0$. Daughter cells inherit all properties of their mother cell and are initiated randomly in a distance of one cell radius from their mother cell.

It should be noted that the initial width of the strip is not particularly important, since wedging will ensure the correct tube width given sufficient proliferation.

All cells in the simulation have the same coupling constants, typically $\lambda = (0.6, 0.4, 0)$. Typical values for Γ_0 and α_0 are 2.8×10^{-4} and 0.5. respectively.

Modeling gastrulation

In our gastrulation simulation, the assignment of PCP and cell wedging is characterized by two radii, describing an annulus (see Figure S5):

$$r_0 = 7, \quad (\text{S10})$$

$$r_1 = 3r_0 = 21. \quad (\text{S11})$$

PCP is assigned within the disk Ω_1 given by

$$\Omega_1 = \left\{ (x, y, z) \mid \sqrt{x^2 + y^2} < r_1 \right\}. \quad (\text{S12})$$

The PCP coupling strength λ is taken to be

$$\lambda = \begin{cases} (0.5, 0.5 - \lambda_3, \lambda_3) & \text{inside } \Omega_1, \\ (1, 0, 0) & \text{everywhere else.} \end{cases} \quad (\text{S13})$$

where a typical value for λ_3 is between 0.08 and 0.12.

The PCP vector field \mathbf{q} is initially assigned so that it spirals around the axis of tube formation (the z -axis):

$$\mathbf{q}|_{t=0} = \hat{\mathbf{z}} \times \mathbf{r}, \quad (\text{S14})$$

In the gastrulation simulations, the PCP vector field is fixed on a per-cell basis.

Nonzero apical constriction parameter α is assigned in an annulus Ω_2 , which shares its outer radius with the disk Ω_1 :

$$\Omega_2 = \left\{ (x, y, z) \mid r_0 < \sqrt{x^2 + y^2} < r_1 \right\}. \quad (\text{S15})$$

The magnitude of α for the cells in Ω_2 is taken as 0.4:

$$|\alpha| = \begin{cases} 0.4 & \text{inside } \Omega_2, \\ 0 & \text{everywhere else.} \end{cases} \quad (\text{S16})$$

The regions Ω_1 and Ω_2 are fixed in space and not on a particle basis. The number of particles in this simulation is $N = 4000$.

Modeling budding from plane

The budding simulation is, apart from global topology, very similar to the gastrulation simulation.

The relevant length parameters are r_0 and r_1 with $r_0 < r_1$. Typically we take

$$r_0 = 5, \quad (\text{S17})$$

$$r_1 = 2r_0 \text{ or } r_1 = 3r_0. \quad (\text{S18})$$

Two regions are correspondingly defined – the disk Ω_1 and the annulus Ω_2 :

$$\Omega_1 := \left\{ (x, y, z) \mid \sqrt{x^2 + y^2} < r_1 \right\}, \quad (\text{S19})$$

$$\Omega_2 := \left\{ (x, y, z) \mid r_0 < \sqrt{x^2 + y^2} < r_1 \right\}. \quad (\text{S20})$$

The PCP coupling strength λ is taken to be

$$\lambda = \begin{cases} (0.5, 0.5 - \lambda_3, \lambda_3) & \text{inside } \Omega_1, \\ (1, 0, 0) & \text{everywhere else.} \end{cases} \quad (\text{S21})$$

where a typical value for λ_3 is between 0.08 and 0.12.

The PCP vector field \mathbf{q} is initially assigned so that it spirals around the center of invagination (the origin of coordinates):

$$\mathbf{q}|_{t=0} = \hat{\mathbf{z}} \times \mathbf{r}, \quad (\text{S22})$$

In the gastrulation simulations, the PCP vector field is fixed on a per-cell basis.

Nonzero apical constriction parameter α is assigned in the annulus Ω_2 with magnitude 0.5:

$$|\alpha| = \begin{cases} 0.5 & \text{inside } \Omega_2, \\ 0 & \text{everywhere else.} \end{cases} \quad (\text{S23})$$

The total number of particles in the simulation is 1384.

References

- Brown, N. & Bron, A. J. (1987), ‘An estimate of the human lens epithelial cell size in vivo’, *Experimental Eye Research* **44**(6), 899–906.
- Eskandari, H. & Salcudean, S. E. (2008), Characterization of the viscosity and elasticity in soft tissue using dynamic finite elements, *in* ‘2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society’, IEEE, pp. 5573–5576.
- Nielsen, B. F. (2019), ‘OrganogenesisPCP’, <https://github.com/BjarkeFN/OrganogenesisPCP>.
- Nissen, S. B., Rønild, S., Trusina, A. & Sneppen, K. (2018), ‘Theoretical tool bridging cell polarities with development of robust morphologies’, *eLife* **7**, e38407.
- Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L. & Lerer, A. (2017), ‘Automatic differentiation in pytorch’.
- Sanchez-Corrales, Y. E., Blanchard, G. B. & Röper, K. (2018), ‘Radially patterned cell behaviours during tube budding from an epithelium’, *eLife* **7**, e35717.
- Yamada, S., Iino, T., Bessho, Y., Hosokawa, Y. & Matsui, T. (2017), ‘Quantitative analysis of mechanical force required for cell extrusion in zebrafish embryonic epithelia’, *Biology Open* **6**(10), 1575–1580.

SELF-ASSEMBLY, BUCKLING AND DENSITY-INVARIANT GROWTH OF THREE-DIMENSIONAL VASCULAR NETWORKS

Authors: J. B. Kirkegaard¹, B. F. Nielsen¹, A. Trusina¹, and K. Sneppen¹.

¹ The Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark.

My contribution: Contributed to conceptualization and writing of the manuscript.

Publication status: Published in *Journal of the Royal Society Interface* (2019).

Hyperlink(s): <https://doi.org/10.1098/rsif.2019.0517>

Research



Cite this article: Kirkegaard JB, Nielsen BF, Trusina A, Sneppen K. 2019 Self-assembly, buckling and density-invariant growth of three-dimensional vascular networks. *J. R. Soc. Interface* **16**: 20190517.
<http://dx.doi.org/10.1098/rsif.2019.0517>

Received: 23 July 2019

Accepted: 2 October 2019

Subject Category:

Life Sciences—Physics interface

Subject Areas:

biophysics, computational biology

Keywords:

polarity, vasculogenesis, buckling, pancreatic islets, self-organization, blood vessels

Author for correspondence:

Julius B. Kirkegaard

e-mail: juliusbierk@gmail.com

Electronic supplementary material is available online at <https://doi.org/10.6084/m9.figshare.c.4695161>.

Self-assembly, buckling and density-invariant growth of three-dimensional vascular networks

Julius B. Kirkegaard, Bjarke F. Nielsen, Ala Trusina and Kim Sneppen

Niels Bohr Institute, University of Copenhagen, 2100 Copenhagen, Denmark

JBK, 0000-0003-0799-3829

The experimental actualization of organoids modelling organs from brains to pancreases has revealed that much of the diverse morphologies of organs are emergent properties of simple intercellular ‘rules’ and not the result of top-down orchestration. In contrast to other organs, the initial plexus of the vascular system is formed by aggregation of cells in the process known as vasculogenesis. Here we study this self-assembling process of blood vessels in three dimensions through a set of simple rules that align intercellular apical–basal and planar cell polarity. We demonstrate that a fully connected network of tubes emerges above a critical initial density of cells. Through planar cell polarity, our model demonstrates convergent extension, and this polarity furthermore allows for both morphology-maintaining growth and growth-induced buckling. We compare this buckling with the special vasculature of the islets of Langerhans in the pancreas and suggest that the mechanism behind the vascular density-maintaining growth of these islets could be the result of growth-induced buckling.

1. Introduction

Tubes are ubiquitous features of numerous biological systems. In humans, they form the gastrointestinal tract, the ductal network of the pancreas, the fallopian tubes, the urinary tract and so on, with the most obvious example being the entire vascular network of blood vessels. On the relevant time scales of multicellular energy consumption, diffusion is limited to delivering metabolites over length scales smaller than $\sim 100 \mu\text{m}$. Instead, on larger length scales, tissue needs some form of directed transport [1]. In vertebrates, this active transport is provided by the beating heart through the vascular network, which in turn has to branch into every part of the organism to nourish tissue and remove waste.

The development of the vascular network involves mainly two processes: vasculogenesis and angiogenesis [2,3]. During vasculogenesis, individual endothelial cells coalesce and *de novo* form functional vessels [4–6]. Studies of vasculogenesis *in vitro* have been mainly restricted to two dimensions [7], but recently three-dimensional vascular organoids have been produced [8]. Vasculogenesis results in a randomly connected vascular plexus, which is subsequently remodelled by pruning or branching [9–12] to a mature vascular network, e.g. with a hierarchical tree-like structure. In angiogenesis, the tree-like structure is formed by branching processes involving either splitting (intussusception) or sprouting dynamics from already formed blood vessels [2,13,14]. This remodelling can be guided by blood flow, pressure and vessel wall stresses [15]. We shall be interested in modelling blood vessel organoids and will thus not consider this latter reorganization, which becomes relevant only in connection with certain organs (such as a pumping heart).

From a theoretical and computational viewpoint, the most intriguing feature of vasculogenesis is its three-dimensional self-assembly of tubular networks. Of equal importance is whether these self-assembled networks *percolate* across the tissue, i.e. whether a fully connected network of tubes is formed. What densities of endothelial cells are needed in three dimensions to ensure these

criteria? We will additionally be interested in questions of growth. Once a network is formed, can this network undergo stable growth? And what are the possible mechanisms for such networks to grow while maintaining a constant space-to-vessel density?

Understanding blood vessel formation computationally has received much attention [16]. Continuum models enable descriptions of density fields of chemotaxing endothelial cells during vasculogenesis [7,17–19]. Likewise, cellular Potts models [20] and models of individual cells [21] have been employed. These studies of vasculogenesis have focused mainly on two-dimensional systems. In this paper, we introduce a coarse-grained description of tubes in three dimensions using a formulation that resolves features of both single cells and full organs (vessel network). In particular, we are able to simulate vascular networks comprising up to hundreds of thousands of particles.

Our focus will be on emergent features of the model such as vasculogenesis and buckling during growth, but we note that our model also has the ability to describe angiogenic sprouting [22] (*budding*) and intussusception, which on the cell level resembles gastrulation [23]. While lumen formation in blood vessels is its own research field [5,24–26], we introduce a simple mechanism for lumen formation by describing the evolution of the apical–basal polarity of cells, thus yielding a fully emergent approach to tubulogenesis.

The paper is organized as follows. In §2, we introduce the methodology and mathematics of the model and demonstrate lumen formation. Section 3 is devoted to vasculogenesis and the percolation of the vascular network. In §4, we study the growth of vascular networks for various parameters of the model and show that both morphology-maintaining and buckling growth patterns can arise. Lastly, in §5, we describe the vasculature of the islets of Langerhans in the pancreas and describe how their tortuous features could be the result of buckling during growth. In particular, we show that the vascular density during the growth of these islets could be maintained simply by a buckling mechanism without the need for angiogenic processes.

2. Model, polarities and lumen formation

Contrary to two-dimensional models, in three dimensions, cell polarity is crucial to model cell sheets. Our coarse-grained model describes a collection of particles/cells each defined by their position x , their apical–basal polarity (AB) p , and their planar cell polarity (PCP) q , illustrated in figure 1. Our model is coarse grained in the sense that a collection of particles model a cell, and as such, even though each particle is a sphere, shape deformations are possible in a collection of particles. In a tube, such as a blood vessel, the AB polarity of cells will define the inside versus the outside of the vessel, while PCP defines the direction around the tube versus the direction along the tube.

To model cell behaviour, we use a slightly modified version of the model of Nissen *et al.* [23]. In this model, particles interact pairwise only if they are line-of-sight Voronoi neighbours and their mutual potential energy is

$$V_{ij} = \exp(-r_{ij}) - S_{ij} \exp(-r_{ij}/\beta), \quad (2.1)$$

for which

$$S_{ij} = \lambda_0 + \lambda_1 S_1^{ij} + \lambda_2 |S_2^{ij}| + \lambda_3 |S_3^{ij}|, \quad (2.2)$$

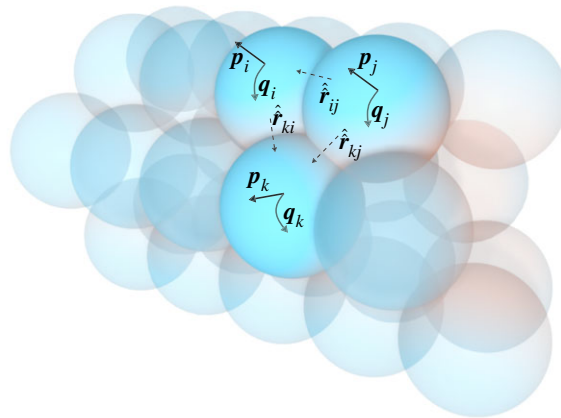


Figure 1. The polarity of each particle, which stems from a distribution of proteins, is modelled as vectors. AB polarity is indicated by p and PCP by q . For a tube, as illustrated, p points away from the tube (distinguishing outside from inside) and q curls around the tube (distinguishing along versus around the tube). Interactions between cells depend on their orientation between each other, \hat{r}_{ij} . In equations (2.3), S_1 favours p_i and p_j parallel and both orthogonal to \hat{r}_{ij} , S_2 favours p_i 's and q_j 's orthogonal and S_3 favours q_i and q_j parallel and both orthogonal to \hat{r}_{ij} . Note that, in reality, cells can deform based on the polarities, but we model them as point particles. Shape deformation is captured by collections of multiple particles. (Online version in colour.)

where

$$\left. \begin{aligned} S_1^{ij} &= (p_i \times \hat{r}_{ij}) \cdot (p_j \times \hat{r}_{ij}), \\ S_2^{ij} &= (p_i \times q_i) \cdot (p_j \times q_j), \\ S_3^{ij} &= (q_i \times \hat{r}_{ij}) \cdot (q_j \times \hat{r}_{ij}) \end{aligned} \right\} \quad (2.3)$$

and

$$\hat{r}_{ij} = \frac{r_{ij}}{r_{ij}} = \frac{x_i - x_j}{|x_i - x_j|}. \quad (2.4)$$

We keep $\beta = 5$, which sets the inter-particle spacing to ≈ 2 units, and furthermore enforce $\lambda_0 + \lambda_1 + \lambda_2 + \lambda_3 = 1$ with $\lambda_1 \geq \lambda_3$. This strikingly simple model can describe a plethora of phenomena related to polarity-driven morphogenesis in organoids [23]. Naturally, real interactions between polarized cells will be much more complex than the model used here; indeed, these depend on the precise distribution of the surface proteins that make up the polarity of the cells. The interactions used here can be thought of as the first relevant and symmetry-obeying terms that give rise to polarity-aligning cells. The simplicity of the present model is thus agnostic towards the underlying microscopic details. Here we introduce a small extension to this model that permits the *de novo* formation of tube-like structures.

The dynamics of the model follows from taking all mobilities to be equal. Hence,

$$\frac{\partial x_i}{\partial t} = -\frac{\partial V}{\partial x_i}, \quad \frac{\partial p_i}{\partial t} = -\frac{\partial V}{\partial p_i}, \quad \frac{\partial q_i}{\partial t} = -\frac{\partial V}{\partial q_i}, \quad (2.5)$$

with the norms of p and q kept at unity. For this study, the model was implemented using PYTORCH and run with CUDA-acceleration.

With only spherical interactions, i.e. $\lambda_0 = 1.0$, a solid tube is a meta-stable structure of this model, as shown in figure 2*a*. Lumen formation corresponds to the formation of AB polarity, i.e. the discrimination of the inside and outside of the tube. Various methods for lumen formation exist, e.g.

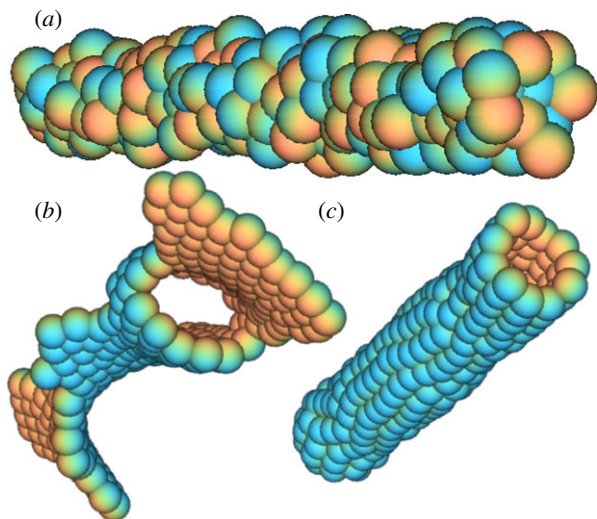


Figure 2. Hollowing of a solid tube with self-organizing AB polarity. (a) Initial solid-tube particles having only spherical interactions ($\lambda_0 = 1$) and random AB polarity. (b) Self-organization of AB polarity with $\gamma = 0$. Similar behaviour is observed with $\lambda_1 = 1.0$ and with $\lambda_1 = 0.5$, $\lambda_2 = 0.42$, $\lambda_3 = 0.08$ with PCP already organized. (c) Self-organization of AB polarity with $\lambda_1 = 1.0$ and $\gamma = 5.0$. Tubes are enclosed in simulations, but cut open for illustration. Colours indicate AB polarity. (Online version in colour.)

extracellular cord hollowing and lumen ensheathment or intracellular vacuole fusion [27]. While the specifics of these mechanisms vary, they all establish the AB polarity of the tubes. If we turn on AB polarity in the present model, that is, we let $\lambda_0 \rightarrow 0.0$ and $\lambda_1 \rightarrow 1.0$, the solid structure tube of figure 2a opens up into a sheet-like structure as shown in figure 2b. This behaviour occurs because of the random initialization of the AB polarity (as illustrated in figure 2a).

To allow AB polarity to form properly we introduce the potential

$$V_i = \gamma \sum_j f(r_{ij}) \mathbf{p}_i \cdot \hat{\mathbf{r}}_{ij}, \quad (2.6)$$

where $f(r) \simeq e^{-r^2/2\ell^2}$, such that the total potential is

$$V = \sum_{ij} V_{ij} + \sum_i V_i. \quad (2.7)$$

This potential aligns AB polarity against local areas of high density, in correspondence with experiments suggesting cell–cell contact directs AB polarity [28]. It can also be thought of as alignment along a gradient field c ,

$$V_i = \gamma \mathbf{p}_i \cdot \nabla c, \quad (2.8)$$

where c is a molecular, diffusing field of particles nucleated at cell locations,

$$D\nabla^2 c = \kappa c - \sum_i \delta(x_i). \quad (2.9)$$

This formulation assumes the existence of such a molecular field. Although many molecular gradients are set up during blood vessel growth, such as vascular endothelial growth factor (VEGF), which elongates and reorganizes cells [6], it is unclear if these interact with and orient the polarity of cells. It is thus easier to think of the interaction as a direct cell–cell interaction as described by equation (2.8).

With $\gamma > 0$, lumen formations occur and the solid tube becomes hollow and fully enclosed, as shown in figure 2c. While network formation and lumen formation in reality are separate processes, in this study we will consider the simplified system of them occurring simultaneously. While $\gamma > 0$ ensures enclosed structures, PCP with $\lambda_3 > 0$ is needed to control the tube thickness. That is, $\lambda_3 > 0$ creates a preference for length-wise alignment of particles, and thus establishes *convergent extension*, which in turn happens through cell intercalation events. Mathematically, the only difference between AB (p) and PCP (q) is the fact that $\lambda_1 > \lambda_3$. The λ_2 term keeps p and q approximately orthogonal, and the magnitude of λ_3 thus determines how much AB alignment is favoured over PCP alignment, which in turn controls the thickness of the tubes since thicker tubes will have better aligned PCP.

In equation (2.2) we include vectorial interactions of AB polarity (S_1), but only nematic interactions of PCP (S_2, S_3), since we do not want to impose a handedness to the vascular tubes. At branch points of the vascular network, there must be defects in PCP alignments, since, in a similar fashion to how you cannot perfectly comb the hair on a sphere, you cannot have a smooth surface vector field at a tube branch point. Taking the absolute value in equation (2.2) turns vectorial -1 charge defects into two $-1/2$ defects, which establishes symmetric branch points (see electronic supplementary material).

Finally, we note that we only model the endothelial cells themselves; in the jargon of active matter research, our model is ‘dry’. In organoid experiments, there will naturally also be culture medium, extracellular matrix, pericytes, etc., present, and the system will perhaps be embedded in, for example, matrigel and collagen [29]. These components mitigate their own interactions between one another and with the cells and could be explicitly modelled in a similar manner to our cell–cell interactions. Such interactions would complicate our model a lot and make interpretations harder, but one should keep in mind that the parameters we use effectively include these interactions and are not solely the result of pure cell–cell interactions. For instance, the effects of the viscosity of the culture medium would effectively introduce mobilities in equation (2.5). Likewise, the effects of shaking could be modelled effectively by including external noise in equation (2.5). We have tested such effects and our results remain qualitatively unchanged.

3. Vasculogenesis

During vasculogenesis blood vessels form from aggregating endothelial cells [2]. Figure 3 shows the self-assembly of three-dimensional vessels in our model. From an initial random distribution (figure 3a) the cells start aggregating (figure 3b) and form a tubular network (figure 3c). In figure 3, cells are initially sampled from a uniform distribution within a sphere but any distribution works.

Naturally, a major concern in vasculogenesis is to form a network of blood vessels that is fully connected. It has previously been shown how this *percolation* condition, i.e. whether all particles connect to one another, depends on the density of endothelial cells [7]. In our model, cells have a preferred distance to one another and can attract over long distances. At first glance, therefore, it seems that initial density might not be an important quantity. However, particles only attract if their polarizations match, and as soon

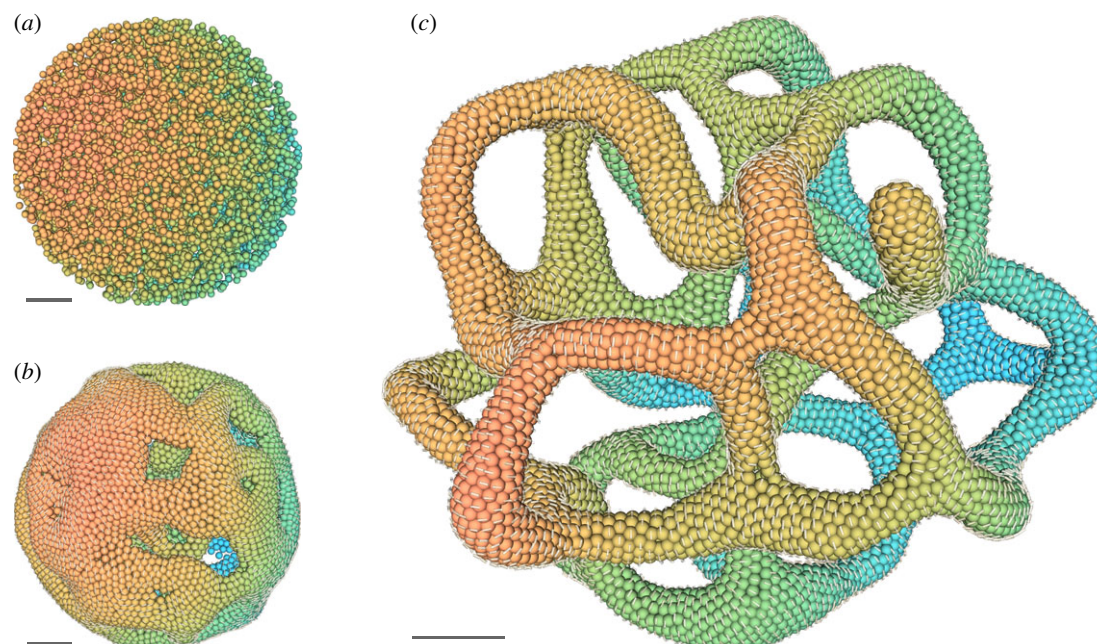


Figure 3. Self-assembly of a vascular network from $\sim 10^4$ particles initialized uniformly randomly within a sphere (q). (a) Initial conditions. (b) Early equilibration of polarization at $t = 2 \times 10^2$, in simulation units. (c) Vascular network formed at $t = 10^4$. Parameters: $\lambda_1 = 0.5$, $\lambda_2 = 0.42$, $\lambda_3 = 0.08$, $\gamma = 5.0$. Scale bar: five particle diameters. Lines indicate the direction of PCP. Colours for visualization purposes only. (Online version in colour.)

as vessel structures have formed, enclosed vessels will not attract one another, since two vessels nearing each other will have opposing AB polarity on their adjacent surfaces. Because of this polarization, the vasculogenesis process in our model is also density dependent.

The density-dependent percolation behaviour is visualized in figure 4, which shows the probability for a particle to be part of the largest cluster P as a function of the initial density ρ . This is shown for various initial radii, or, in other words, for various numbers of particles ranging from ~ 250 to $\sim 30\,000$. As is clear, the vessel network percolates at around $\rho_c \sim 8.2 \times 10^{-3}$, i.e. at an initial length scale of $\rho_c^{-1/3} \sim 5$ —the same order of magnitude as the inter-particle spacing = 2. Figure 4 also shows some finite-size effects, since for a small number of particles, even those far below the transition point, the largest cluster, albeit small, will constitute a significant fraction of the whole system.

4. Growth and buckling

As organisms grow, their networks of blood vessels also need to grow. The vascular system needs to grow in two distinct ways: first, blood vessels need to increase their diameter in order to deliver increased amounts of blood. However, as vessels grow, their surface area to volume fraction decreases and so does their effectiveness. Thus, they also need to grow their network structure to maintain a space-filling network with small diameter vessels, capillaries, as the ‘leaves’ of the network [30]. This latter version of growth is called angiogenesis and, as mentioned, is not the focus of our study. In this section, we introduce the growth of the blood vessel and consider the effect of PCP strength λ_3 , which creates a preference for growth in tube length rather than in tube diameter. The next section will demonstrate a less considered alternative to space-filling growth exploiting the buckling phenomenon demonstrated in this section.

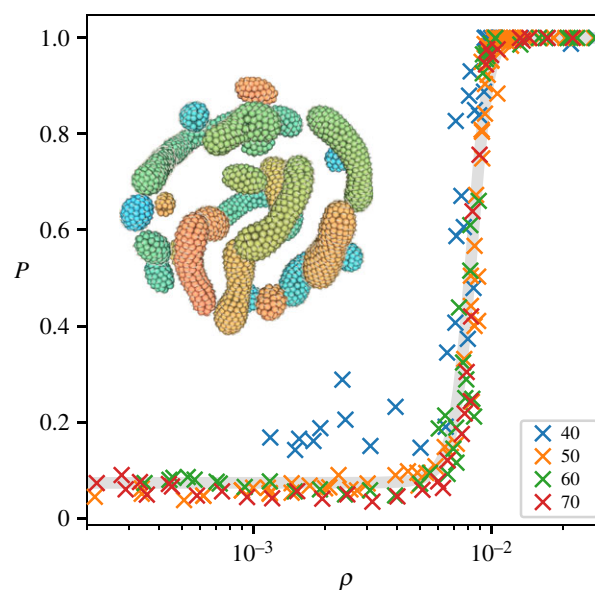


Figure 4. Percolation of vascular networks. Graphs show the probability P for particles to belong to the largest cluster as a function of the initial density $\rho = n/(4/3 \pi r^3)$, where n is the number of particles and r is the radius of the initialization sphere, its value indicated by the legend. The critical density is found to be $\rho_c \sim 8.2 \times 10^{-3}$. Inset shows the small clusters for $n = 3300$ and $r = 50$. Parameters: $\lambda_0 = 0.0$, $\lambda_1 = 0.5$, $\lambda_2 = 0.45$, $\lambda_3 = 0.05$, $\gamma = 5.0$. The critical percolation density ρ_c depends on λ_3 , since thinner structures percolate more easily. (Online version in colour.)

First, we demonstrate that the growth of vessel diameters follows naturally when $\lambda_3 = 0$. Cell division is implemented as a Poisson process in the sense that each cell has a constant rate of division ν . When a cell divides, a new cell is created with the same polarities p and q , but placed at a random position next to its mother cell in the plane orthogonal to p , meaning that cells divide within the cell sheet.

Figure 5 shows the results of growth dynamics with $\lambda_3 = 0$. Figure 5a is the steady-state result of a self-assembly with $\lambda_3 = 0.08$. We then let $\lambda_3 \rightarrow 0$ and $\nu \rightarrow 3 \times 10^{-5}$. Figure 5b

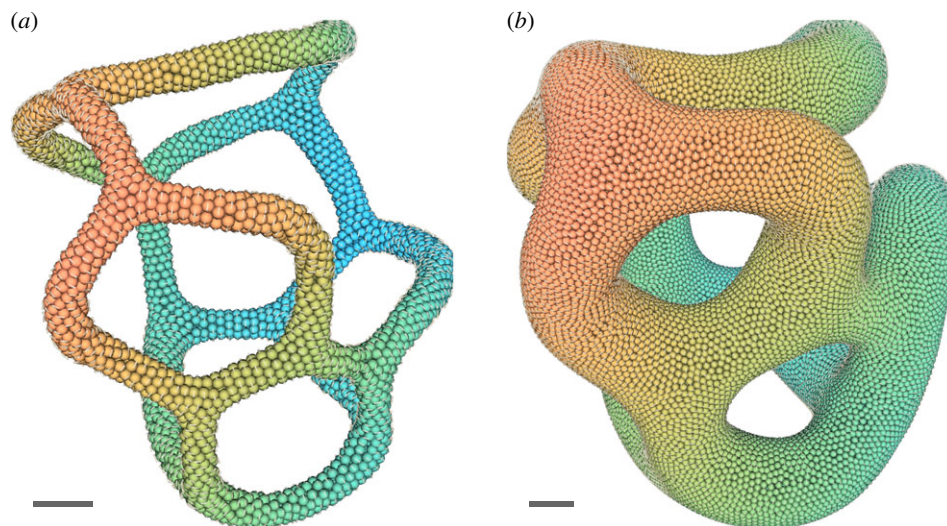


Figure 5. Growth of vascular network with $\lambda_3 = 0$. (a) Steady structure formed of $n = 3500$ particles with $\lambda_3 = 0.08$. (b) After cell division to $n = 25\,000$ particles with $\lambda_3 = 0$. Remaining parameters: $\lambda_0 = 0.0$, $\lambda_1 = 0.5$, $\lambda_2 = 0.42$, $\gamma = 5.0$. Scale bar: five particle diameters. (Online version in colour.)

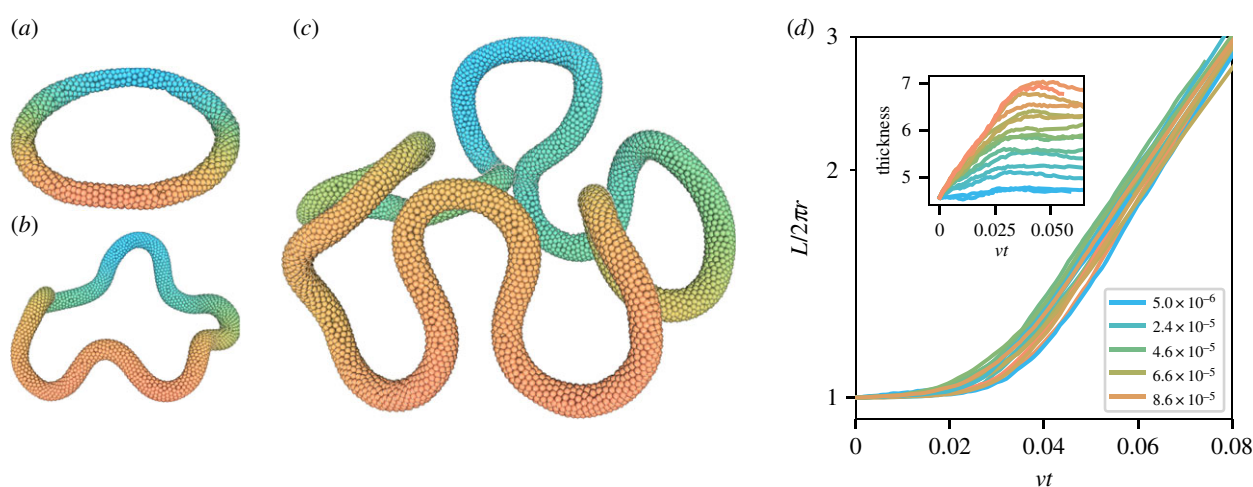


Figure 6. Buckling of vessels during cell growth with $\lambda_3 > 0$. (a) Initial condition of a torus of 1000 cells. (b,c) Buckling during growth. (d) Buckling measured as the vessel contour length L over the effective radius r as a function of time re-scaled by division rate v , which is shown by colour; its value is given in the legend. $vt \sim 0.03$ corresponding to ~ 2100 cells. Inset shows tube thickness (N/L) during buckling. (Online version in colour.)

shows the result of the growth from 3500 particles to 25 000 particles. As is evident, the vascular network can grow uniformly under this model. However, the structure as a whole does not grow much, and, in fact, the density (vascular volume to free space) grows as well. This happens because we only model the cell division of the vascular network. In reality, the tissue between the vessels, which are cells we are not modelling, will also be dividing and in turn grow the structure as a whole.

If we instead consider the case of $\lambda_3 > 0$ with cell division, this leads to completely different growth. Since λ_3 induces a preferred diameter, this creates a tendency to grow more in length than thickness. Figure 6a–c shows how this leads to a growth-induced buckling instability [31], i.e. growth that does not retain the structure's shape. The buckling occurs because cell division is faster than the time to relax shape perturbations on long length scales.

A simple measure of buckling of a growing ring is to compare the curve length L with the structure's effective radius r . If no buckling occurs $L/r = 2\pi$. This definition works well as long as the structure remains a simple curve. Figure 6d shows this buckling as a function of time for various division rates v . Clearly, the buckling occurs even for exceedingly small

division rates at approximately the same value of vt , i.e. approximately the same number of cells. After the transition, the buckling degree grows exponentially in time, or, in other words, linearly with the number of cells. In this regime, this can be understood as the structure growing mostly in structure length L and not in effective size r . This continues until one side of the structure meets the other.

The inset of figure 6d shows the thickness of tubes calculated as N/L , where N is the number of cells in the structure. As is clear, a high growth rate leads to thicker tubes as the time scale for growth severely outpaces relaxation. The thickness of the tubes grows until buckling starts to occur, after which the thickness decreases as there is suddenly room to grow in length instead of thickness and the thickness stabilizes. The peak in thickness occurs at slightly later vt for larger v .

5. Islets of Langerhans

In the pancreas, the so-called islets of Langerhans are responsible for the production of hormones such as insulin. While these islets constitute only about 1% of the pancreatic volume they contain about 10% of the blood vasculature. This dense vessel network is needed to provide energy to

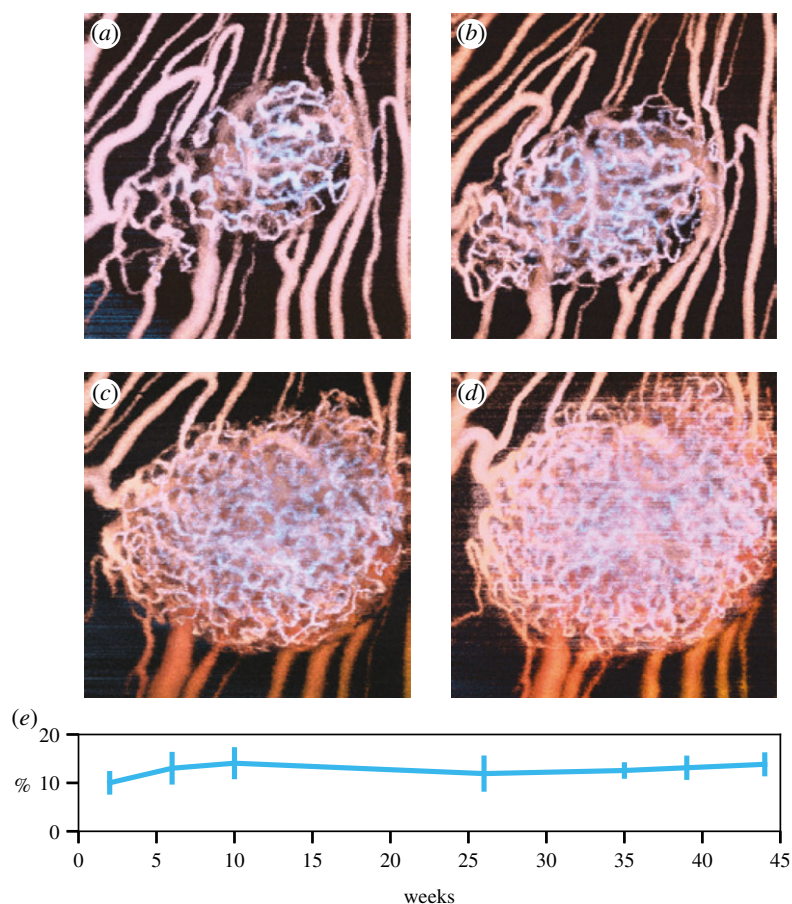


Figure 7. Growth of the vasculature of islets of Langerhans over 44 weeks. Images show the vasculature at weeks 2 (*a*), 6 (*b*), 35 (*c*) and 44 (*d*). The islet is growing, but its vascular density remains approximately constant, as shown in (*e*). The *y*-axis shows the percentage of the vascular volume of the total islet volume. During these 44 weeks, the islet more than triples its volume. Data and images from Berclaz *et al.* [33]. (Online version in colour.)

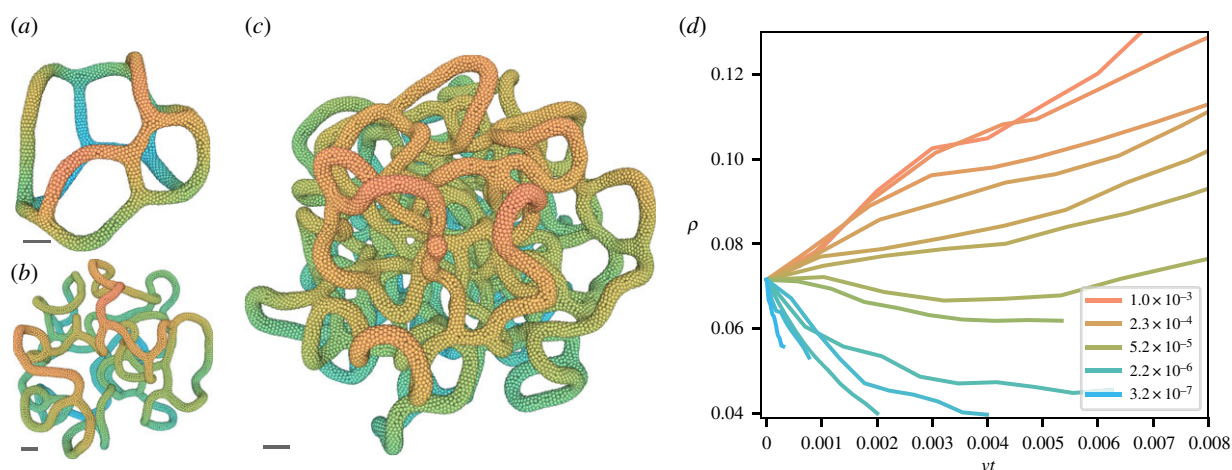


Figure 8. Vasculature growth. (*a*) Initial meta-stable structure of 3500 particles. (*b*) Structure at $\sim 15\,000$ particles under $\nu \approx 2.5 \times 10^{-6}$ growth. (*c*) Growth with $\nu \approx 5.0 \times 10^{-5}$ having reached the same effective radius as (*b*), but at $\sim 45\,000$ particles and thus at a much higher density. (*d*) Vascular density ρ of the structure as a function of time (rescaled by division rate ν , as given in the legend). Scale bar: five particle diameters. (Online version in colour.)

the islets; it is also needed for intercellular communication [32] and to measure and regulate the production and injection of hormones into the bloodstream.

The dense vasculature of pancreatic islets is shown in figure 7. The figure also shows the growth of these islets over 44 weeks. A crucial observation is that the islets' vascular density, despite the overall growth of the islets, remains almost constant [33]. Furthermore, the thickness of the blood vessels also does not change much over this course

(figure 7; variations in thickness are linked to age [34]). In other words, the growth of the vessel network is unlike that of figure 5. Angiogenesis is the canonical explanation for the growth of vascular network structure and this is indeed a factor for pancreatic islets [32].

What is also clear from figure 7 is the tortuous structure of the vasculature. This is in contrast to normal vasculature, which is regular and structured, as can be seen, for example, in the surrounding vessels in figure 7. Considering only

angiogenesis as a growth mechanism, the tortuous structure and the space-filling growth are independent observations.

Both observations can, however, be simultaneously explained by growth-induced buckling, as we demonstrate in figure 8. Figure 8*b,c* shows the different structures that can be attained in our model with different growth rates. As is clear, growing vessel structures with $\lambda_3 > 0$ leads to very tortuous structures. Furthermore, as shown in figure 8*d*, the division rate ν also determines the vascular density. With a large division rate, the structure does not have time to relax under the division-induced stress and the vascular density increases. Conversely, the vascular density is decreased for small division rates. Hence, a simple feedback mechanism where the proliferation rate inversely depends on the density can keep vascular density constant during growth. While this effect would definitely be co-occurring with angiogenesis, it is intriguing that it simultaneously gives an explanation for the tortuosity of the vascular network. Although the AB polarity-aligning parameter γ is only used for self-assembly, during this sort of growth $\gamma > 0$ can allow for anastomosis, the fusing of separate blood vessels.

6. Conclusion

We have demonstrated three-dimensional vasculogenesis in a simple model that aligns cell polarities through cell–cell interactions. The initial self-assembly of enclosed structures is ensured through the alignment of AB polarity against cell density. This is the key driver of lumen formation and enables the *de novo* formation of tubes. This interaction also allows for fusing of vessels (anastomosis). While PCP is not needed for the formation of enclosed structures, this polarity ensures thin tubular structures and convergent extension.

The self-assembly of the vascular networks results in fully connected vessels if the particles are initialized above a critical density, in accordance with previous two-dimensional experiments [7]. Enclosed, non-connected structures appear at lower densities, and these do not interact because of their opposing AB polarities.

References

- Goldstein RE, van de Meent J-W. 2015 A physical perspective on cytoplasmic streaming. *Interface Focus* **5**, 20150030. (doi:10.1098/rsfs.2015.0030)
- Udan RS, Culver JC, Dickinson ME. 2013 Understanding vascular development. *WIREs* **2**, 327–346. (doi:10.1002/wdev.91)
- Cleaver O, Krieg PA. 2010 *Vascular development*, vol. 1. Amsterdam, The Netherlands: Elsevier Inc.
- Davis GE, Bayless KJ, Mavila A. 2002 Molecular basis of endothelial cell morphogenesis in three-dimensional extracellular matrices. *Anat. Rec.* **268**, 252–275. (doi:10.1002/(ISSN)1097-0185)
- Hogan BLM, Kolodziej PA. 2002 Molecular mechanisms of tubulogenesis. *Nat. Rev. Genet.* **3**, 513–523. (doi:10.1038/nrg840)
- Helmlinger G, Endo M, Errara N, Hlatky L, Jain RK. 2002 Formation of endothelial cell networks. *Nature* **405**, 139–141. (doi:10.1038/35012132)
- Gamba A, Ambrosi D, Coniglio A, de Candia A, Di Talia S, Giraudo E, Serini G, Preziosi L, Bussolino F. 2003 Percolation, morphogenesis, and Burgers dynamics in blood vessels formation. *Phys. Rev. Lett.* **90**, 118101. (doi:10.1103/PhysRevLett.90.118101)
- Chen Ya *et al.* 2017 A three-dimensional model of human lung development and disease from pluripotent stem cells. *Nat. Cell Biol.* **19**, 542–549. (doi:10.1038/ncb3510)
- Kurz H, Go R. 2001 Structural and biophysical simulation of angiogenesis and vascular remodeling. *Dev. Dyn.* **401**, 387–401. (doi:10.1002/dvdy.1118)
- Ochoa-Espinosa A, Affolter M. 2015 Branching morphogenesis from cells to organs and back. *Cold Spring Harbor Perspect. Biol.* **4**, 1–14.
- Dodds PS. 2010 Optimal form of branching supply and collection networks. *Phys. Rev. Lett.* **104**, 048702. (doi:10.1103/PhysRevLett.104.048702)
- Restrepo JG, Ott E, Hunt BR. 2006 Scale dependence of branching in arterial and bronchial trees. *Phys. Rev. Lett.* **96**, 128101. (doi:10.1103/PhysRevLett.96.128101)
- Preziosi L, Astanin S. 2006 Modelling the formation of capillaries. In *Complex systems in biomedicine* (eds A Quareroni, L Formaggia, A Veneziani), pp. 109–145. Milan, Italy: Springer.
- Kurz H, Burri PH, Djonov VG. 2015 Angiogenesis and vascular remodeling by intussusception: from form to function. *Physiology* **18**, 65–70. (doi:10.1152/nips.01417.2002)
- Pries AR, Reglin B, Secomb TW. 2005 Remodeling of blood vessels: responses of diameter and wall thickness to hemodynamic and metabolic stimuli. *Hypertension* **46**, 725–731. (doi:10.1161/01.HYP.0000184428.16429.be)

Introducing cell proliferation in our model leads to distinct behaviour depending on the strength λ_3 of PCP, which controls the preference of tube diameter. With $\lambda_3 = 0$, we have shown that uniform growth is possible. With $\lambda_3 > 0$ the tubes buckle under growth. While blood vessel buckling is typically associated with high blood pressure [35], this shows that such behaviour can also stem from cell proliferation.

Considering this buckling mode of growth, we compared it with the vasculature of islets of Langerhans, which show a large degree of tortuosity. The vessel density of these pancreatic islets furthermore remains constant during their growth. We suggest that a simple explanation for this behaviour is growth-induced buckling in which the cell division rate is coupled to the vascular density. This rate, in turn, may be controlled by negative feedback from the blood supply to the tissue.

Tortuous blood vessels are in general also found in cancerous tumours [36]. Cancer growth is often associated with angiogenesis in nearby tissue, a process that we did not explore here, but could easily be introduced via local cell shape changes [22]. The abnormality of tumour vessels is thought to be, in part, due to over-expression of VEGF-A [37]. We have shown how tortuosity in blood vessels can be linked to the growth rate. This could thus also play a major role in the abnormal morphology of tumour vascularization.

Data accessibility. This article does not contain any additional data.

Authors' contributions. J.B.K. conceptualized the study, designed the model, wrote the model code, ran the experiments and analysis and drafted the manuscript. B.F.N. participated in model discussions, was involved in writing the model code and revised the manuscript. A.T. provided biological insight and revised the manuscript. K.S. conceptualized the study, participated in model and analysis discussions and revised the manuscript.

Competing interests. We declare we have no competing interests.

Funding. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme, grant agreement no. 740704, and the Danish National Research Foundation, grant no. DNRF116.

Acknowledgements. We thank Anne Grapin-Botton for helpful discussions.

16. Ambrosi D, Bussolino F, Preziosi L. 2005 A review of vasculogenesis models. *J. Theor. Med.* **6**, 1–19. (doi:10.1080/1027366042000327098)
17. Manoussaki D, Lubkin SR, Vernon RB, Murray JD. 1996 A mechanical model for the formation of vascular networks *in vitro*. *Acta Biotheor.* **44**, 271–282. (doi:10.1007/BF00046533)
18. Murray JD. 2003 On the mechanochemical theory of biological pattern formation with application to vasculogenesis. *C. R. Biol.* **326**, 239–252. (doi:10.1016/S1631-0691(03)00065-9)
19. Namy P, Ohayon J, Tracqui P. 2004 Critical conditions for pattern formation and *in vitro* tubulogenesis driven by cellular traction fields. *J. Theor. Biol.* **227**, 103–120. (doi:10.1016/j.jtbi.2003.10.015)
20. Merks RMH, Glazier JA. 2006 Dynamic mechanisms of blood vessel growth. *Nonlinearity* **19**, C1–C10. (doi:10.1088/0951-7715/19/1/000)
21. Merks RMH, Perryn ED, Shirinifard A, Glazier JA. 2008 Contact-inhibited chemotaxis in de novo and sprouting blood-vessel growth. *PLoS Comput. Biol.* **4**, e1000163. (doi:10.1371/journal.pcbi.1000163)
22. Nielsen BF, Nissen SB, Sneppen K, Trusina A, Mathiesen J. 2019 Model to link cell shape and polarity with organogenesis. *bioRxiv*. (doi:10.1101/699413)
23. Nissen SB, Rønild S, Trusina A, Sneppen K. 2018 Theoretical tool bridging cell polarities with development of robust morphologies. *eLife* **7**, e38407. (doi:10.7554/eLife.38407)
24. Lubarsky B, Krasnow MA. 2003 Tube morphogenesis: making and shaping biological tubes. *Cell* **112**, 19–28. (doi:10.1016/S0092-8674(02)01283-7)
25. Xu K, Cleaver O. 2011 Tubulogenesis during blood vessel formation. *Semin. Cell Dev. Biol.* **22**, 993–1004. (doi:10.1016/j.semcdb.2011.05.001)
26. Iruela-Arispe ML, Beitel GJ. 2013 Tubulogenesis. *Development* **140**, 2851–2855. (doi:10.1242/dev.070680)
27. Schuermann A, Helker CSM, Herzog W. 2014 Angiogenesis in zebrafish. *Semin. Cell Dev. Biol.* **31**, 106–114. (doi:10.1016/j.semcdb.2014.04.037)
28. Strilić B, Kučera T, Eglinger J, Hughes MR, McNagny KM, Tsukita S, Dejana E, Ferrara N, Lammert E. 2009 The molecular basis of vascular lumen formation in the developing mouse aorta. *Dev. Cell* **17**, 505–515. (doi:10.1016/j.devcel.2009.08.011)
29. Wimmer RA *et al.* 2019 Human blood vessel organoids as a model of diabetic vasculopathy. *Nature* **565**, 505–510. (doi:10.1038/s41586-018-0858-8)
30. West GB, Brown JH, Enquist BJ. 1997 A general model for the origin of allometric scaling laws in biology. *Science* **276**, 122–126. (doi:10.1126/science.276.5309.122)
31. Drasdo D. 2000 Buckling instabilities of one-layered growing tissues. *Phys. Rev. Lett.* **84**, 4244–4247. (doi:10.1103/PhysRevLett.84.4244)
32. Ballian N, Brunicardi FC. 2007 Islet vasculature as a regulator of endocrine pancreas function. *World J. Surg.* **31**, 705–714. (doi:10.1007/s00268-006-0719-8)
33. Berclaz C *et al.* 2016 Label-free fast 3D coherent imaging reveals pancreatic islet micro-vascularization and dynamic blood flow. *Biomed. Opt. Express* **7**, 4569. (doi:10.1364/BOE.7.004569)
34. Almagá J, Molina J, Arrojo R, Abdulreda MH, Jeon WB, Berggren P-O, Caicedo A, Nam HG. 2014 Young capillary vessels rejuvenate aged pancreatic islets. *Proc. Natl Acad. Sci. USA* **111**, 17 612–17 617. (doi:10.1073/pnas.1414053111)
35. Han HC, Chesnutt JKW, Garcia JR, Liu Q, Wen Q. 2013 Artery buckling: new phenotypes, models, and applications. *Ann. Biomed. Eng.* **41**, 1399–1410. (doi:10.1007/s10439-012-0707-0)
36. Goel S, Duda DG, Xu L, Munn LL, Boucher Y, Fukumura D, Jain RK. 2012 Normalization of the vasculature for treatment of cancer and other diseases. *Physiol. Rev.* **91**, 1071–1121. (doi:10.1152/physrev.00038.2010)
37. Nagy JA, Chang S-H, Dvorak AM, Dvorak HF. 2009 Why are tumour blood vessels abnormal and why is it important to know? *Br. J. Cancer* **100**, 865–869. (doi:10.1038/sj.bjc.6604929)

CHAPTER 2

CHEMICAL SELF-ORGANIZATION: BLOCK COPOLYMERS

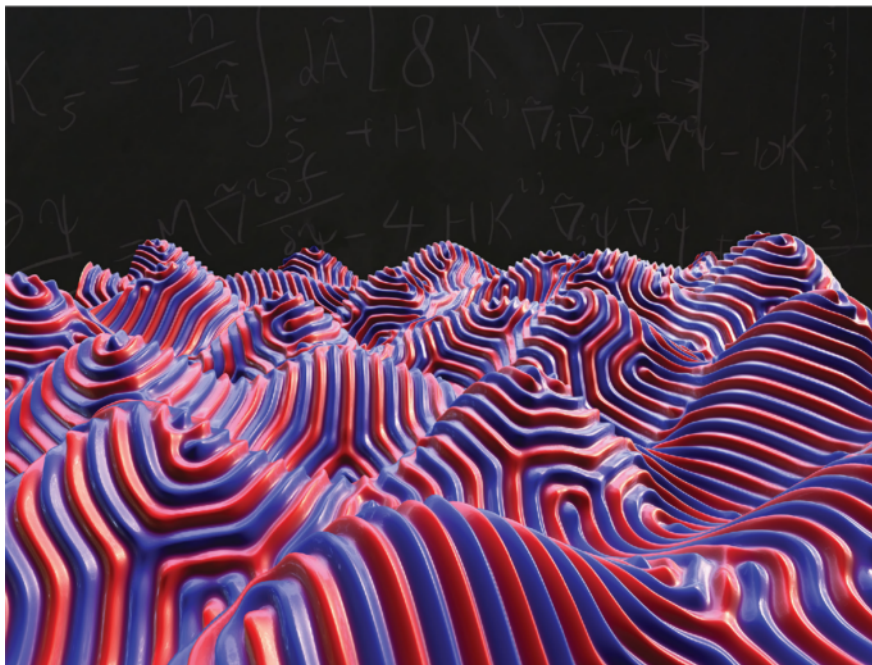


Figure 15: Simulation of crystallization of lamellar phase block copolymers on a curved substrate. Figure from Ref. [20].

2.1 INTRODUCTION

We now turn to a non-living system, which does however bear similarity with the one discussed in the previous chapter in one respect. In that chapter, we saw how *cells* can self-organize into elongated structures such as the tubes of the vascular system. In this chapter, we will be looking at *molecules* which can self-organize into patterns of elongated structures, namely stripes. This class of molecules is called diblock copolymers. The ‘copolymer’ part refers to polymers which consist of more than one species of monomer. The ‘block’ part then refers to the fact that the monomers occur in patterns such as A-A-A-A-B-B-B-B-A-A-A-A, as opposed to e.g. stochastic or random copolymers. Finally the ‘di’ part refers to the fact that only two types of blocks are involved in these particular polymers.

In this chapter, which is based wholly on Ref. [20], we develop a finite-thickness model of (thin) layers of diblock copolymers. This is done by starting from a flat 2D description, covariantizing it and extending the theory into the third dimension by means of a small thickness parameter in

which we can expand the free energy functional. We then develop a novel numerical framework for solving these equations on curved manifolds. The solutions reveal how curvature can be used as a guiding field to control the pattern formation. The space of applications is vast, ranging from microelectronics to microfluidics.

2.2 METHODS

The theoretical development of our model starts from the Brazovskii free energy, which has previously been used to describe block copolymers [21, 22, 23, 24, 25]. This free energy, denoted $F(\psi)$, is a Ginzburg-Landau expansion in the order parameter $\psi(\mathbf{x})$. We work with the free energy *density* functional, which is just the free energy per volume, $f = F/V$:

$$f(\psi) = \frac{1}{V} \int dV \left[2(\nabla^2 \psi)^2 - 2|\nabla \psi|^2 + \frac{\tau}{2} \psi^2 + \frac{1}{4} \psi^4 \right]. \quad (2.1)$$

The order parameter ψ is the ‘‘relative composition’’, i.e. it measures the over-density of the ‘A’ or ‘B’ monomers. Formally it is the deviation $\psi(\mathbf{x}) = \phi(\mathbf{x}) - \phi_0$ from the average composition ϕ_0 at the critical temperature T_c . The model thus has a single tunable parameter, namely the reduced temperature $\tau = (T - T_c)/T_c$. We assume $\phi_0 = 0$ throughout our study, since we’re interested in the compositionally symmetric lamellar phase. To see that this free energy favours the emergence of periodic patterns with a certain wavelength, we insert a ‘‘stripe field’’ (which is periodic in only one direction, which we take to be the x direction without loss of generality), $\psi(\mathbf{x}) = \psi_0 \sin(q_0 x)$. The free energy then becomes³

$$f(\psi) = (q_0^4 - q_0^2) \psi_0^2 \quad (2.2)$$

This is of course analogous to studying the Fourier transformed equations of motion, with q_0 the wave number. Minimizing the free energy gives $q_0 = 1/\sqrt{2}$ corresponding to a characteristic wavelength of $\lambda = 2\pi\sqrt{2}$. Any deviation from this wavelength is energetically penalized, ensuring that a stripe pattern of a well-defined spacing is preferred.

Our goal is to develop a theory for thin films of such block copolymers on curved surfaces. The natural first step is to take the two-dimensional free energy and adapt it for a curved surface. This entails replacing partial derivatives with their covariant counterparts, i.e. $\partial \rightarrow \nabla$, to make sure that derivatives transform tensorially even on a general curved background. The simple flat-space volume element $dx dy$ must also be replaced by its invariant counterpart. We write:

$$f(\psi) = \frac{1}{A} \int d\tilde{A} \left[2(\tilde{\nabla}^2 \psi)^2 - 2|\tilde{\nabla} \psi|^2 + \frac{\tau}{2} \psi^2 + \frac{1}{4} \psi^4 \right]. \quad (2.3)$$

Here the tilde denotes a surface quantity. It is also important to understand that the shorthand $|\tilde{\nabla} \psi|^2$ no longer stands for the Euclidean norm $(\nabla_x \psi)^2 + (\nabla_y \psi)^2 + \dots$ but rather for the appropriate inner product, formed by contraction with the surface metric g_{ij} , i.e. $|\tilde{\nabla} \psi|^2 = g_{ij} \tilde{\nabla}_i \psi \tilde{\nabla}_j \psi$. We use the Einstein summation convention, where a repeated index is implicitly summed over. The indices i, j, \dots take on the values $\{1, 2\}$.

This programme of replacing the bulk free energy with a covariantized surface version has been used in related models [26, 27, 28, 25, 29]. It introduces a coupling between the pattern and the underlying geometry through the intrinsic curvature of the surface. If the surface has intrinsic curvature along some particular direction, the pattern will either be compressed or stretched in this direction. This necessarily carries an energy penalty and the pattern will respond by reorienting. The effect of this is thus to align the stripes along the direction of largest intrinsic curvature.

³The ‘trick’ to obtain simple expressions not involving higher powers of ψ_0 is to compute the average energy over an entire wavelength of the pattern.

However, this covariantization approach is not enough to describe a film that has a non-zero thickness. It does not take the extension of the pattern into the third dimension into account. This, as we shall see more clearly later, is equivalent to saying that it does not take the *extrinsic* curvature of the surface into account. Consider the cylindrical substrate illustrated in Figure 16. Such a geometry has no *intrinsic* curvature whatsoever, only extrinsic curvature. Hence every orientation of the stripe pattern is equally energetically favourable according to (2.3). However, due to the finite thickness of the film, the orientations in panels a and b clearly place the innermost parts of the stripe pattern under compression while the outermost part is diluted, meaning that neither will obtain the most favourable spacing. Only the configuration in panel c allows for uniform, optimal spacing throughout. This is a clear argument for the importance of not only intrinsic, but also extrinsic curvature as a guiding field for pattern formation in block copolymers.

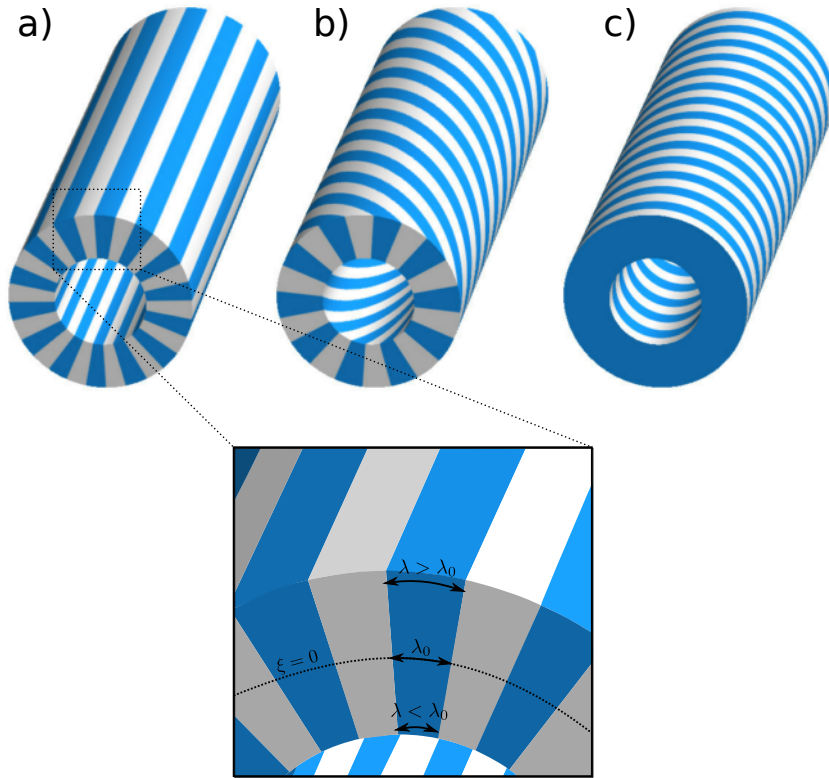


Figure 16: Three possible orientations of a striped pattern of definite wavelength on a cylindrical surface. Note that the cylindrical surface is extrinsically curved but carries no intrinsic curvature. The orientation in panel a is maximally energetically unfavourable, since the stripes are forced to deviate from their preferred spacing (λ_0), being compressed ($\lambda < \lambda_0$) near the inner surface and diluted ($\lambda > \lambda_0$) near the outer surface. The pattern shown in panel b is intermediate, while the one shown in panel c allows the pattern to obtain its preferred spacing throughout.

Geometric setup. Our scheme for including the third dimension in a thin-film approximation is to consider the film as a three-dimensional region Ω of thickness h defined symmetrically around the midplane \tilde{S} , which is then a regular two-dimensional surface. The setup is sketched in Figure 17. Note that we use a tilde to denote midsurface quantities. The normal coordinate is denoted by $\xi \in [-h/2, h/2]$ and the surface \tilde{S} is parametrized by the coordinates u and w . As such, a position $\mathbf{p}(u, w, \xi)$ in the region Ω can be reached by choosing a point $\tilde{\mathbf{p}}(u, w)$ in \tilde{S} (the projection of \mathbf{p} onto the surface) and moving a distance ξ along the surface normal $\tilde{\mathbf{n}}(u, w)$:

$$\mathbf{p}(u, w, \xi) = \tilde{\mathbf{p}}(u, w) + \xi \tilde{\mathbf{n}}(u, w). \quad (2.4)$$

In order to compute the surface metric \tilde{g}_{ij} we need the tangent vectors $\tilde{\mathbf{a}}_i$. Note that these comprise a family of three-dimensional vectors which also carry surface indices, since a three-dimensional

tangent vector exists for each coordinate direction on the surface. In this case, it is cleaner to switch to index notation with separate sets of indices for the three-dimensional manifold Ω and the two-dimensional submanifold \tilde{S} . We thus denote the tangent vectors \tilde{a}_i^μ , with the latin indices i, j, k, \dots referring to the surface and the Greek indices μ, ν, ρ, \dots referring to Ω . These tangent

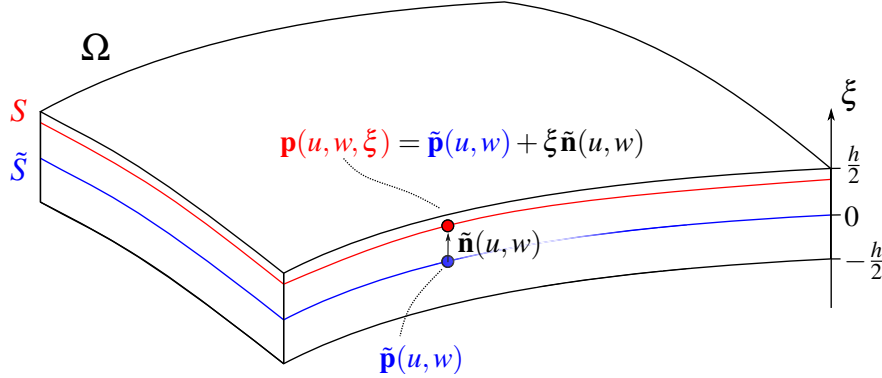


Figure 17: Schematic of the geometric setup used in the derivation of the effective free energy. The film is modeled as a thin three-dimensional region Ω with thickness h distributed uniformly around the midplane \tilde{S} . The coordinates on the two-dimensional surface \tilde{S} are taken as u and w . Each point $\mathbf{p} \in \Omega$ can then be reached from a point $\tilde{\mathbf{p}} \in \tilde{S}$ by moving some distance ξ (the normal coordinate) along the normal vector $\tilde{\mathbf{n}}(u, w)$. This is our chosen parametrization of Ω .

vector components are also themselves elements of the Jacobian or change-of-coordinate matrix, and are also referred to as *projectors*. The reason for this nomenclature is that they are the objects which allow us to *pull back* ambient quantities onto the surface. For instance, a co-vector ω_μ may be projected according to:

$$\tilde{\omega}_i = \tilde{a}_i^\mu \omega_\mu. \quad (2.5)$$

Recall the tensor transformation law under change of coordinates from e.g. x^μ to X^i :

$$T_j^i = \frac{\partial x^\rho}{\partial X^j} \frac{\partial X^i}{\partial x^\sigma} T_\rho^\sigma. \quad (2.6)$$

The projectors (or tangent vectors) and their inverses⁴ can thus be identified as

$$\tilde{a}_i^\mu = \frac{\partial x^\mu}{\partial X^i}, \quad (2.7)$$

$$\tilde{a}^i_\mu = \frac{\partial X^i}{\partial x^\mu}, \quad (2.8)$$

where we have denoted the coordinates on the surface by X^i (with $(X^1, X^2) = (u, w)$) and the three-dimensional coordinates by x^μ (with $(x^1, x^2, x^3) = (x, y, z)$).

The induced metric on the surface can then be computed by a pullback as described above:

$$\tilde{g}_{ij} = \tilde{a}_i^\mu \tilde{a}_j^\nu \delta_{\mu\nu} = \tilde{\mathbf{a}}_i \cdot \tilde{\mathbf{a}}_j, \quad (2.9)$$

where we used the fact that the three-dimensional background of course has the Euclidean metric $\delta_{\mu\nu}$. The inverse metric \tilde{g}^{ij} can be obtained by a pullback like the above or by solving the equation $\tilde{g}^{ik} \tilde{g}_{kj} = \delta^i_j$ (i.e. by matrix inversion).

The curvature tensor can then be computed as

$$K_{ij} = n_\mu \tilde{\nabla}_i \tilde{a}_j^\mu. \quad (2.10)$$

⁴One may object that since \tilde{a}_i^μ is not in general represented by a quadratic matrix, it doesn't have a true inverse. The inverses used here are in fact *projective* inverses.

The curvature tensor as defined above should not be conflated with either the Riemann or Ricci tensors, which are completely intrinsic. K_{ij} , on the other hand, contains information about the extrinsic curvature as well.

This curvature tensor also allows for the identification of the principal curvatures at a point $\tilde{\mathbf{p}}$ on the surface, which we shall denote $\kappa_1(\tilde{\mathbf{p}})$ and $\kappa_2(\tilde{\mathbf{p}})$. This allows us to more precisely define what it means to work in the thin film limit. We first define a local curvature length scale $l(\tilde{\mathbf{p}})$ by:

$$l(\tilde{\mathbf{p}}) = \min \left[\frac{1}{\kappa_1(\tilde{\mathbf{p}})}, \frac{1}{\kappa_2(\tilde{\mathbf{p}})} \right]. \quad (2.11)$$

$l(\tilde{\mathbf{p}})$ is thus the shortest curvature radius at the given point, corresponding to the greatest curvature. The global curvature scale ℓ is then defined as the shortest local curvature length on the surface:

$$\ell = \min_{\tilde{\mathbf{p}} \text{ in } \tilde{S}} \{l(\tilde{\mathbf{p}})\}. \quad (2.12)$$

The thin film limit is then defined by the condition

$$\left(\frac{h}{\ell} \right)^2 \ll 1. \quad (2.13)$$

Before we turn to the systematic thickness expansion, we will introduce a central assumption. We assume that the order parameter ψ can be considered as uniform throughout the thickness of the film, i.e. that $\psi(\tilde{\mathbf{p}} + \xi \tilde{\mathbf{n}}) = \psi(\tilde{\mathbf{p}})$ for all points $\tilde{\mathbf{p}}$ in \tilde{S} and all $\xi \in [-h/2; h/2]$. This assumption could be violated by polymers which change their configuration appreciably in response to compression or dilation.

Energy density thickness expansion. Our goal is to develop an effective two-dimensional theory which nonetheless takes into account extrinsic curvature effects in the thin film limit. We start with the three-dimensional free energy density functional of (2.1) and write it as an expansion in the normal coordinate ξ . This will allow us to integrate out this parameter and obtain an effective 2D free energy functional to lowest order in the small dimensionless parameter (h/ℓ) .

Due to orthogonality of tangent and normal vectors, the curvature tensor can be written in terms of derivatives of either:

$$K_{ij} = \tilde{n}_\mu \tilde{\nabla}_i \tilde{a}_j^\mu = -\tilde{a}_j^\mu \tilde{\nabla}_i \tilde{n}_\mu. \quad (2.14)$$

In the latter form it is clear that the curvature tensor measures the (projected) rate of change of the normal vector as one moves across the surface. In general, we will assume the curvature to be slowly varying, so that its derivatives can be neglected in the equations of motion.

In the end, we wish to write the equations in terms of surface quantities. Now, due to the definition of the surface tangent vector, $\tilde{\mathbf{a}}_i = \partial_i \tilde{\mathbf{p}}$ and the occurrence of the normal vector in (2.4), the metric g_{ij} of a subsurface inside Ω (at some given ξ) will depend explicitly on the curvature. The exact expression is:

$$g_{ij} = \tilde{g}_{ij} - 2\xi K_{ij} + \xi^2 K^k{}_i K_{jk}. \quad (2.15)$$

While that expression is exact, we compute the inverse metric to second order in the small quantity ξ/ℓ . In these expansions, the curvature has order ℓ^{-2} .

$$g^{ij} = (1 - 2\xi^2 K) \tilde{g}^{ij} + 2(\xi + 2\xi^2 H) K^{ij} + \mathcal{O}(\xi/\ell)^3. \quad (2.16)$$

Here, H denotes the mean curvature $H = (1/2)K^i{}_i$ and K is the Gaussian curvature $K = \det(K^i{}_j)$. While the latter is intrinsic, the former encodes extrinsic curvature as well.

The above expression for the metric also immediately shows that the curvature will affect the $|\nabla\psi|^2$

and $(\nabla^2\psi)^2$ terms of the free energy, something which becomes clearer when writing them in their full index form:

$$|\nabla\psi|^2 = g^{ij}(\partial_i\psi)(\partial_j\psi), \quad (\nabla^2\psi)^2 = (g^{ij}\nabla_i\nabla_j\psi)^2. \quad (2.17)$$

The remaining terms of the free energy are not free from curvature effects, however. The invariant volume element $\sqrt{g}d^d x$ which multiplies all terms of the free energy involves the determinant of the metric. Here we use g as short hand for the metric determinant. When written in terms of ξ and midsurface quantities, it becomes:

$$\sqrt{g} = J_\xi \sqrt{\tilde{g}}, \quad \text{where } J_\xi = 1 - 2H\xi + K\xi^2. \quad (2.18)$$

Splitting the volume element into a ξ part and a two-dimensional part as $dV = d\xi dA$ we can write the total volume, which occurs in the denominator of the free energy, as:

$$V = \int_\Omega dV = \int_{-h/2}^{h/2} d\xi \int d\tilde{A} J_\xi = \tilde{A}h + \frac{h^3}{12}\chi \quad (2.19)$$

Combining these terms and integrating out the ξ dependence, we can write the free energy as

$$f = \frac{h\tilde{A}}{V} (f_{\tilde{S}} + k_{\tilde{S}}) \quad (2.20)$$

where $f_{\tilde{S}}$ is the covariantized Brazovskii surface free energy given in (2.3) and the finite-thickness correction $k_{\tilde{S}}$ is given by:

$$k_{\tilde{S}} = \frac{h^2}{12\tilde{A}} \int_{\tilde{S}} d\tilde{A} \left[8 \left((K^{ij}\tilde{\nabla}_i\nabla_j\psi)^2 + H(K^{ij}\tilde{\nabla}_i\tilde{\nabla}_j\psi)(\tilde{\nabla}^2\psi) \right) \right. \\ \left. - 10K(\tilde{\nabla}^2\psi)^2 - 4 \left(HK^{ij} - K\tilde{g}^{ij} \right) (\tilde{\nabla}_i\psi)(\tilde{\nabla}_j\psi) + K \left(\frac{\tau}{2}\psi^2 + \frac{1}{4}\psi^4 \right) \right]. \quad (2.21)$$

The only ingredient missing before we can begin to implement this equation in a numerical framework is a time evolution equation. This is given by the relaxational dynamics of a conserved order parameter with free energy functional f :

$$\frac{\partial\psi}{\partial t} = M\tilde{\nabla}^2 \frac{\delta f}{\delta\psi}, \quad (2.22)$$

where $\frac{\delta}{\delta\psi}$ is a functional derivative operator and M is a mobility coefficient which merely sets the time scale and plays no interesting role for our results.

Numerical implementation. In order to solve the equation of motion derived above, we developed a numerical framework by the name of *Surfaise* as a layer on top of the FEniCS/Dolfin system for solving partial differential equations (PDEs) using the finite element method. The framework we developed is in fact very general, and allows one to input any analytically parametrizable curved surface and then solve a wide range of PDEs on said surface. Surface derivatives, curvature tensors and related quantities are then automatically computed using the SymPy Python package for symbolic manipulations. The software developed for this study is open source and available at https://github.com/gautelinga/surface_pfc/SoftMatter2019. For a detailed description of our implementation of the equations of motion, see the appendix of Ref. [20].

2.3 RESULTS

We wish to gain a systematic understanding of the effects of curvature (intrinsic and extrinsic) on the formation of lamellar patterns. Before we do so, we simulate the system on a randomly generated substrate of sinusoidal bumps and ridges to develop some intuition for the solutions. In Figure 18, we plot the end state of crystallization for three different values of the film thickness (panels a-c) along with the mean and Gaussian curvatures of the underlying surface (panels d-e). As the thickness increases and curvature effects become stronger, it is clear that the stripes have a

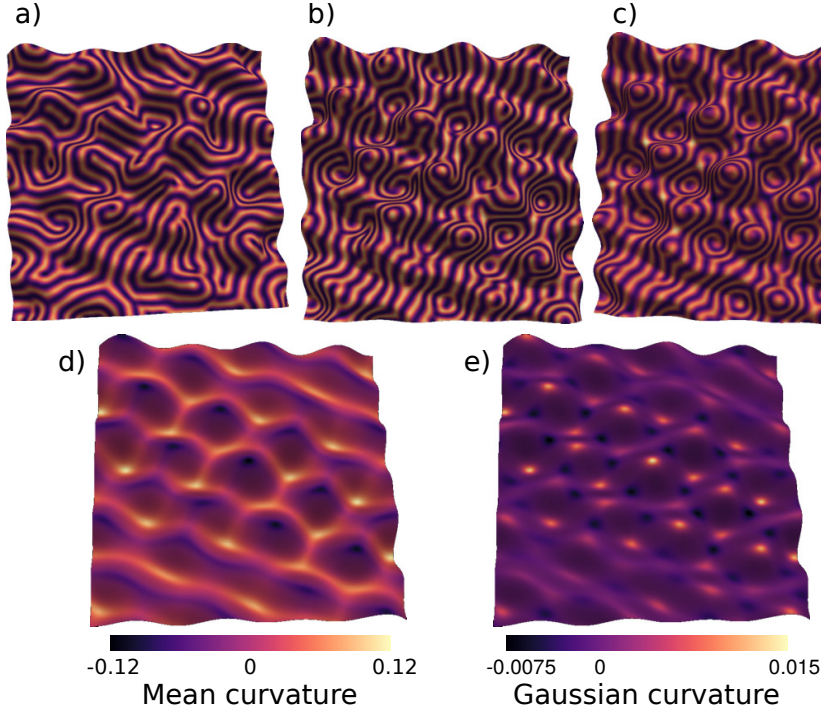


Figure 18: Crystallization at different values of the film thickness. **a)** $(h/\ell)^2 = 0$. **b)** $(h/\ell)^2 = 0.49$. **c)** $(h/\ell)^2 = 1.25$ (formally outside the thin film approximation). **d)** Mean curvature of the substrate. **e)** Gaussian curvature of the substrate.

tendency to encircle the bumps and run perpendicular to the ridges. For the sake of intuition, recall that ridges are similar to cylinders in that they are dominated by extrinsic curvature, whereas the bumps have intrinsic curvature as well.

In order to separately study these effects, we study the crystallization on a few select, simple geometries. Assume that the surface can be locally parametrized in terms of two coordinates $x^1 = u$ and $x^2 = w$ and that these are chosen such that the coordinate curves are lines of curvature. This particular choice results in a diagonal curvature tensor which can be written entirely in terms of the mean and Gaussian curvatures:

$$K^i_j = \text{diag} \left(H \pm \sqrt{H^2 - K}, H \mp \sqrt{H^2 - K} \right). \quad (2.23)$$

In order to delineate the effects of intrinsic and extrinsic curvature, we study two geometries which are extremal in the sense that they have *either* Gaussian (K) or mean curvature (H). These geometries are the cylinder ($K = 0, H \neq 0$) and the saddle ($K \neq 0, H = 0$).

For **the cylinder**, we find analytically that the preferred angle of the stripes is $\pi/2$ relative to the axis of the cylinder, causing stripes to run around the cylinder, consistent with what we observed for the ridges in Figure 18. Simulating the crystallization process on the cylinder, we find that the angle $\pi/2$ is strongly preferred when film thickness is taken into account, while letting $h = 0$ results in no preferred direction arising (Figure 19). In order to reach the minimum-energy state, we employed an annealing procedure where the system was repeatedly heated, then cooled, to prevent it from getting stuck in a local “false” minimum.

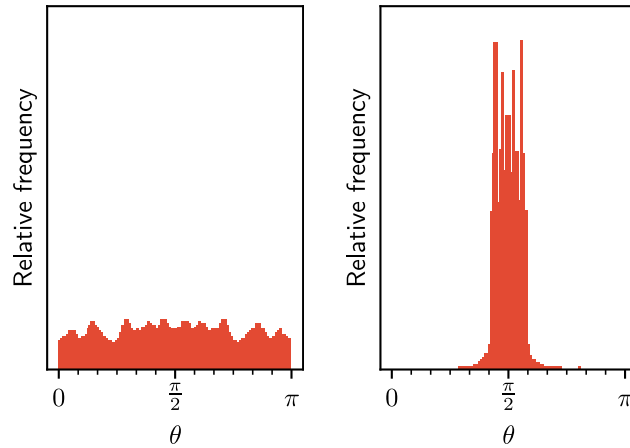


Figure 19: The frequency of pattern orientations on the cylinder. **Left:** With $h = 0$, no particular preferred orientation is observed. **Right:** With $h > 0$, an orientation of $\pi/2$ is strongly preferred, corresponding to stripes running *around* the cylinder.

We now turn to the **saddle geometry**, which has intrinsic curvature ($K < 0$) but no extrinsic curvature ($H = 0$). Analytically, we can again insert a sinusoidal stripe pattern with amplitude ψ_0 (although the expression is more involved, due to the nontrivial intrinsic geometry, see [20] for details). We find that the finite thickness correction to the free energy reduces to

$$k_{\tilde{S}} = \frac{h^2}{3A} \psi_0^2 \cos(4\alpha) + \text{const.} \quad (2.24)$$

Here, α is the angle with one of the principal directions on the saddle. The minimum-energy configuration thus occurs when the *director* (which is orthogonal to the stripes) makes a 45° angle. Simulations very clearly corroborate this (Figure 20). At the top of the figure, we show an example of a configuration reached by annealing, while the bottom shows a plot of the average distribution of angles (from entirely radial to azimuthal), clearly showing how curvature directs the pattern formation.

Our last example is **the Gaussian bump**. The central part of the Gaussian bump has *positive* nonzero intrinsic as well as extrinsic curvature. However, the Gaussian curvature changes as one moves radially outwards from the top. At a certain radial coordinate, $r = \sigma$, the Gaussian curvature locally vanishes and then changes sign. The bump of course has two principal curvatures, a radial and an azimuthal curvature which we denote by κ_r and κ_ϕ , respectively. To gain a better understanding of the relative strength of these curvatures, we plot their ratio in Figure 21 for five different values of the reduced height \tilde{h} , which is the ratio h_0/σ between the height and the width of the bump. Note that these variables should not be confused with the film *thickness* denoted by h . Likewise, the radial coordinate plotted on the horizontal axis is the *reduced* radius, $\tilde{r} = r/\sigma$. We see that the azimuthal curvature is generally stronger (when $r \leq \sigma$) and that the ratio quickly drops as a function of distance when the bump is tall.

In Figure 22a we show the outcome of a simulated crystallization on the Gaussian bump with finite thickness effects taken into account. The pattern clearly confirms the finding of Figure 18, where we observed a tendency for stripes to run around bumps. In Figure 22b, we plot the ensemble averaged stripe orientation, with a black circle showing the radius at which the intrinsic curvature vanishes ($r = \sigma$) and a red circle showing the radius at which the strength of the two curvatures are equal. From panels b and c, we can clearly see that the tendency for stripes to run around the bump (at $h/\ell > 0$) is very strong inside the red circle, but that long-range order quickly vanishes beyond this point.

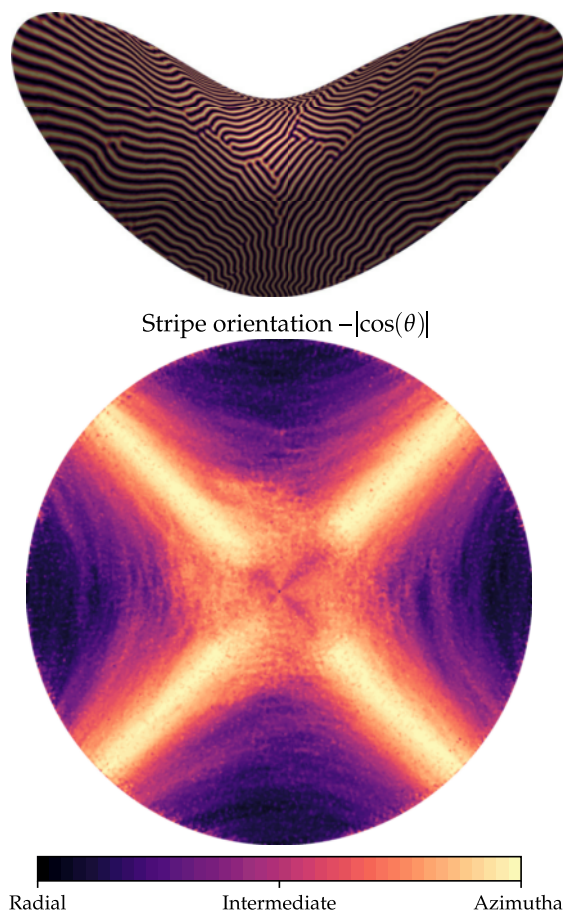


Figure 20: **Top:** A crystallized pattern on a saddle geometry. **Bottom:** The mean orientation of the director (perpendicular to the stripes) based on 12 separate runs. The orientation is measured as $|\cos(\theta)|$ where $\theta = \angle(\hat{\mathbf{r}}, \tilde{\nabla}\psi)$. In this plot, the directions of principal curvature run in the horizontal and vertical directions. The director is clearly seen to orient itself along the radial vector at an azimuthal angle of $\pi/4$, as predicted by the theory. In these simulations, $(h/\ell)^2 = 0.23$.

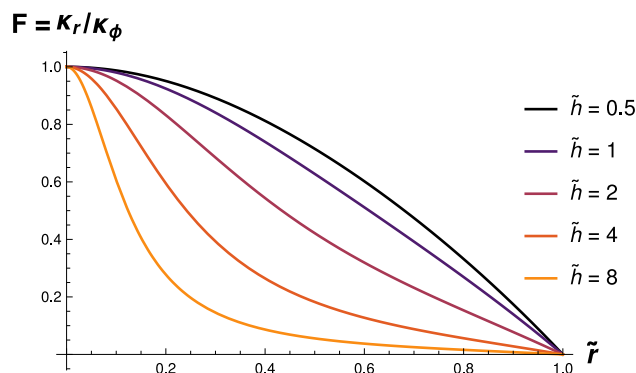


Figure 21: The relative strengths of the two principal curvatures on the Gaussian bump - the radial curvature κ_r and the azimuthal curvature κ_ϕ . The colours indicate different (reduced) bump heights, as indicated by the legend.

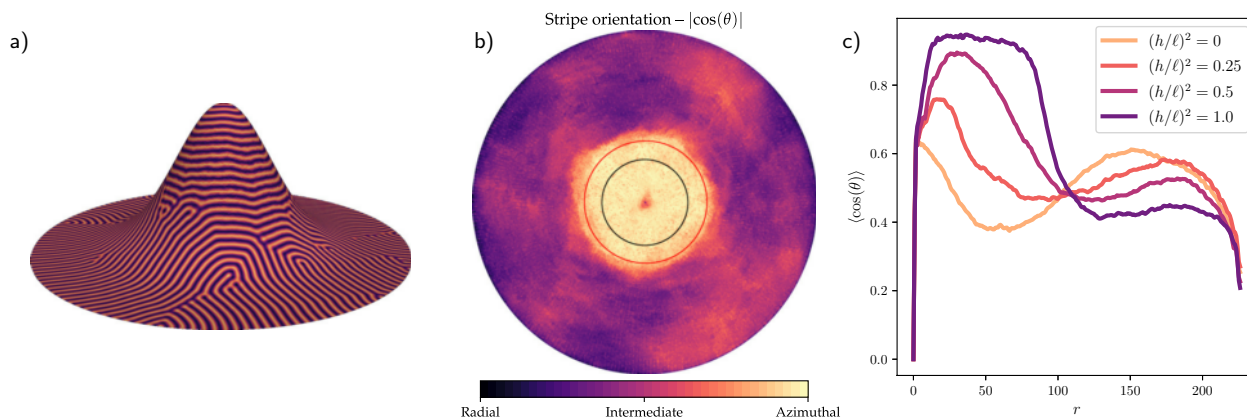


Figure 22: **a)** Outcome of a simulated crystallization on a Gaussian bump. **b)** The average orientation of the director field, relative to the radial vector. **c)** Again, the average orientation of the director, but only as a function of radial coordinate. The azimuthal angle has been averaged over. The tendency for stripes to run *around* the bump is seen to depend strongly on film thickness.

2.4 DISCUSSION

This concludes the chapter on chemical self-assembly. In closing, it deserves mention that there are several relevant possible extensions of this model. For instance, we have assumed that the surface is perfectly rigid, but it would be fascinating to see what happens if one allows for some feedback between the pattern and the underlying geometry in this finite thickness model, by e.g. promoting the metric tensor or the thickness itself to a dynamical variable. A 2015 paper by Matsumoto et al. [30] did study how defects can give rise to deformations of the underlying surface, but did so in a simple covariantized Brazovskii model and so could not take extrinsic curvature effects into account.

2.5 PUBLICATIONS FOR CHAPTER 2

The second chapter of this thesis builds on the following manuscript. The paper was written during this degree and I have not submitted it for any other academic degree.

1. **B. F. Nielsen**, G. Linga, A. Christensen, and J. Mathiesen, “Substrate curvature governs texture orientation in thin films of smectic block copolymers”, *Soft Matter* **16** (2020), no. 14, 3395–3406.

SUBSTRATE CURVATURE GOVERNS TEXTURE ORIENTATION IN THIN FILMS OF SMECTIC BLOCK COPOLYMERS

Authors: B. F. Nielsen¹, G. Linga², A. Christensen³, and J. Mathiesen¹.

¹ The Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark.

² PoreLab, The Njord Centre, Department of Physics, University of Oslo, Oslo, Norway.

³ Danmarks Nationalbank, Copenhagen, Denmark

My contribution: Contributed to development of computational framework (*Surfaise*), performing simulations, analytical calculations, interpreting data, creating figures and journal cover art and writing of the manuscript.

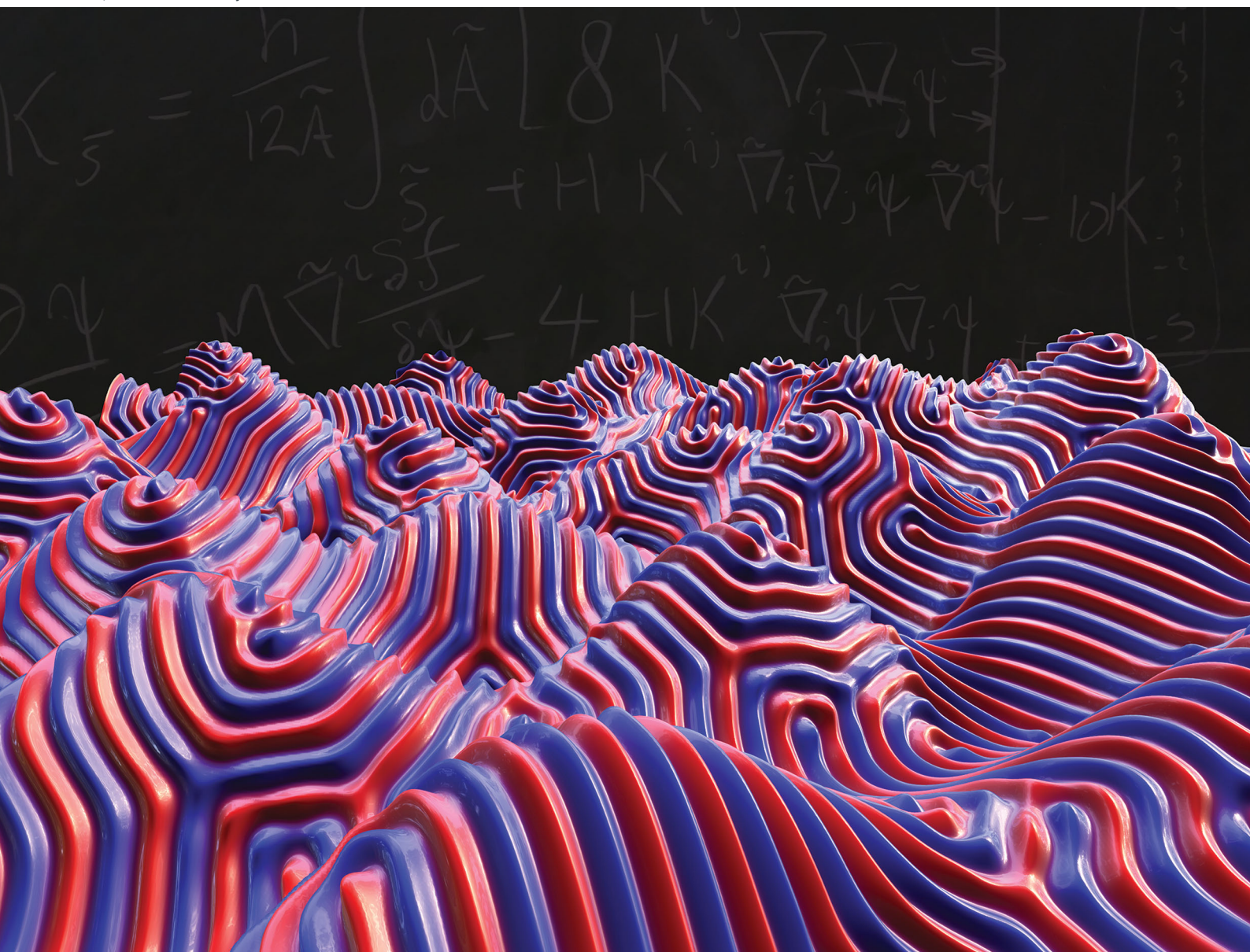
Publication status: Published in *Soft Matter* (2020).

Hyperlink(s): <https://doi.org/10.1039/C9SM02389E>

Volume 16
Number 14
14 April 2020
Pages 3353–3588

Soft Matter

rsc.li/soft-matter-journal



ISSN 1744-6848



ROYAL SOCIETY
OF CHEMISTRY

PAPER

Joachim Mathiesen *et al.*
Substrate curvature governs texture orientation in thin films
of smectic block copolymers



Cite this: *Soft Matter*, 2020,
16, 3395

Substrate curvature governs texture orientation in thin films of smectic block copolymers

Bjarke Frost Nielsen,^a Gaute Linga,^b Amalie Christensen^{ac} and Joachim Mathiesen^{ib*}

Self-assembly of ordered nanometer-scale patterns is interesting in itself, but its practical value depends on the ability to predict and control pattern formation. In this paper we demonstrate theoretically and numerically that engineering of extrinsic as well as intrinsic substrate geometry may provide such a controllable ordering mechanism for block copolymers films. We develop an effective two-dimensional model of thin films of striped-phase diblock copolymers on general curved substrates. The model is obtained as an expansion in the film thickness and thus takes the third dimension into account, which crucially allows us to predict the preferred orientations even in the absence of intrinsic curvature. We determine the minimum-energy textures on several curved surfaces and arrive at a general principle for using substrate curvature as an ordering field, namely that the stripes will tend to align along directions of maximal curvature.

Received 4th December 2019,
Accepted 3rd March 2020

DOI: 10.1039/c9sm02389e

rsc.li/soft-matter-journal

1 Introduction

Thin films of block-copolymers have received strong attention in the last two decades due to their diverse nanometer-scale self-assembly properties. Their ability to form regular hexagonal and cylindrical as well as lamellar patterns makes them promising candidates for applications in microelectronics and optics as well as nanofluidics. In microelectronics and the semiconductor industry, where feature-size is of the essence, much of the appeal comes from the use of thin block copolymer films as etch masks for fabrication of ultra-small circuitry elements and memory devices. “Bottom-up” self-organization of block co-polymers promise to continue the miniaturization to length scales where traditional “top-down” lithography ceases to be feasible.^{1,2} Cylindrical phase block copolymers allow for manufacture of nanoporous membranes for ultrafiltration and molecular sieves^{3–7} as well as superhydrophobic materials in nanofluidics.⁸

Lamellar and cylindrical phase block copolymer films have been demonstrated as viable templates for microelectronic circuitry and polarizing grids as well.^{9–13}

For most of these applications, a high degree of long-range order and control over macroscopic patterning is desirable. In practice, this is complicated by the formation of defects and

microdomains. Different experimental techniques have been developed in attempts to avoid defects and obtain a macroscopic order. One such method is chemoepitaxy, where the substrate is pretreated with another chemical species, thus using the interfacial energy to facilitate the formation of long-range ordered patterns.^{14–16} Shearing flow^{11,12,17} as well as applied electric fields^{13,18} have also been used with some success. Perhaps the most obvious approach to annihilating defects is annealing–heating to near the order–disorder transition temperature and subsequently cooling. A more sophisticated version of this is the sweeping temperature gradient method, which has also proven relatively effective.^{19,20} Our work focuses on using curvature as an ordering field – *i.e.* using substrate topography to control the macroscopic order of lamellar patterns, analogously to an external field. Experimental studies have already shown this graphoepitaxy technique to be a viable method to control microdomain formation.^{21–28} However, for this technique to be generally applicable, we must understand how to design substrates to favour the formation of specific patterns and – conversely – which types of pattern formation to expect as a function of substrate geometry. Our focus is on the smectic-symmetry stripe patterns obtained from compositionally symmetric diblock copolymers.

In this paper we consider a free energy which is dominated by the deviation of the stripe spacing from its preferred value, in accord with the approach of Pezzutti *et al.*²⁹ Our strategy is to formulate a free energy which takes into account not only the intrinsic geometry, but also the extrinsic geometry. The latter comes into play due to the fact that the co-polymer film has a non-zero thickness. While the film is thin, the extension into

^a Niels Bohr Institute, University of Copenhagen, 2100 Copenhagen, Denmark.
E-mail: bjarkenielsen@nbi.ku.dk, mathies@nbi.ku.dk

^b PoreLab, The Njord Centre, Department of Physics, University of Oslo,
P. O. Box 1048, 0316 Oslo, Norway. E-mail: gaute.linga@mn.uio.no

^c Danmarks Nationalbank, DK-1093 Copenhagen K, Denmark

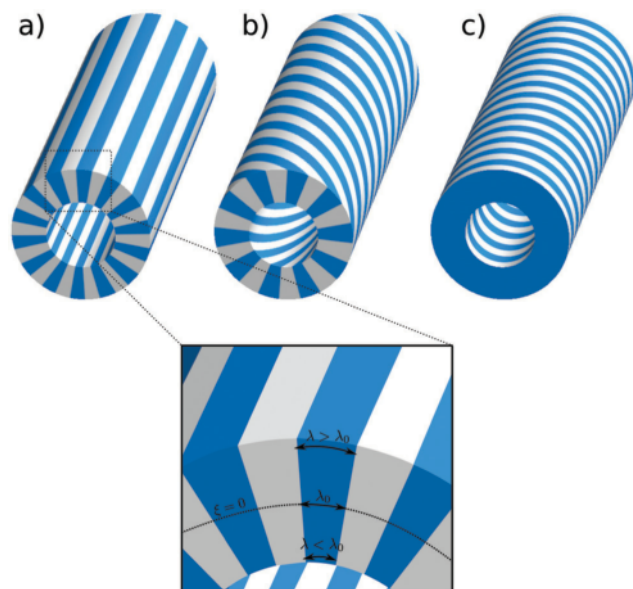


Fig. 1 Stripe textures on a cylindrical surface, with different texture orientations, given by the angle α between the stripe direction and the axial direction. (a) When the stripe texture runs along the cylinder axis $\alpha = 0$, the finite thickness of the layer and the curvature of the cylinder results in a slight increase of the stripe wavelength λ with the radial coordinate, see inset. (b) Also the stripes with orientation $\alpha = \pi/4$ experience an increase of the wavelength in the radial direction, although the effect is smaller. (c) When the stripes run around the cylinder, $\alpha = \pi/2$ they are not affected by the curvature. This figure is inspired by ref. 29.

the third dimension nonetheless has implications for the minimum-energy pattern on any given curved surface. Consider, as an example, the thin striped layers in Fig. 1(a and b). Due to bending of the surface, the stripe spacing is forced to vary across the thickness of the film, leading to the layer being simultaneously under compression and dilation. Hence, even though the film is thin, the third dimension cannot be neglected as it effectively couples the free energy to the extrinsic curvature of the surface.

By performing a systematic expansion in the thickness of the film, we obtain a two-dimensional effective theory which leads to an explicit coupling between extrinsic geometry and pattern formation. We perform computer simulations of this system, assuming the dynamics to consist of the relaxation towards equilibrium of a conserved order parameter. We find that in situations where the intrinsic curvature vanishes, such as on cylinders, on ridges and in trenches, the extrinsic curvature can serve to orient the pattern in a controllable fashion. Consequently, the coupling to extrinsic geometry is indispensable in such situations. In cases where there is only appreciable Gaussian curvature, such as the saddle geometry, the intrinsic geometry picks out a preferred direction. In situations where both types of curvature are present we find that the extrinsic and intrinsic features may work together in orienting the pattern. As such, we show that there are situations where extrinsic curvature is merely a contributing factor and others where it is the crucial component in determining the macroscopic order.

There have been previous attempts at modeling the effect of curvature on lamellar phase block co-polymer assembly, but none which correctly incorporate the effects of film thickness and coupling to extrinsic curvature for general surfaces. Several authors have developed models,^{30,31} where both intrinsic and extrinsic bending of the stripes themselves is energetically penalized. Intrinsic bending occurs when the stripes deviate from geodesics of the surface. Extrinsic bending, on the other hand, occurs when the stripes bend in three-dimensional space. As an example, consider the two-dimensional top layers of the cylindrical films in Fig. 1. The stripes running along the cylinder (Fig. 1a) have neither intrinsic nor extrinsic bending, whereas stripes running around the cylinder (Fig. 1c) have no intrinsic bending but do have extrinsic bending because they are curved in three-dimensional space. The type of model employed in ref. 30 and 31 implies that stripes prefer to be straight in 3D and that running along the cylinder as in Fig. 1a is preferred.

The starting point for our expansion is a Brazovskii-type free energy. This type of free energy has been used before to model the effects of curvature on block copolymer configurations.^{29,32,33} Pezzutti *et al.*²⁹ employ a covariantized Brazovskii surface free energy and perform a finite-thickness expansion specifically in the case of the cylinder and thus arrive at the conclusion that the preferred stripe direction is around the cylinder, as in Fig. 1c. However they do not derive a general finite-thickness model for arbitrary surfaces and, furthermore, investigate only surfaces of vanishing Gaussian curvature. Matsumoto *et al.*³² employ the same type of block copolymer free energy and couple it to a Canham–Helfrich membrane model of the substrate. By covariantizing the free energy, the metric – and thus the intrinsic geometry – naturally couples to the copolymer pattern. This leads to a model that predicts stripe patterns running perpendicular to substrate wrinkles for nonzero Gaussian curvature. However, since vanishing thickness is assumed, the coupling of the phase field to the extrinsic curvature is not captured. As such, this model cannot predict a preferred orientation of stripes in *e.g.* a cylindrical geometry. Interestingly, Matsumoto *et al.*³² also allow the surface to adapt to the copolymer pattern by assuming a relaxational dynamics of the height field. Vega *et al.*³³ take a different approach, namely three-dimensional (3D) simulation of a Brazovskii-type model confined to a thin, curved patch of 3D space. They arrive at the prediction that the stripes tend to run around the cylinder. Their approach is simple in principle, requiring no covariantization or finite-thickness expansion of the free energy, but it has the disadvantage of not making the curvature-coupling explicit and of requiring simulation of a large number of degrees of freedom.

It is thus clear that attempts at modeling the effects of curvature on block copolymer stripe patterns have led to contradictory results. However, experiments^{32–34} may shed light on the features one should expect from a successful model of these phenomena. The simplest experimental paradigm in this regard is the cylinder, since it exhibits uniform extrinsic curvature while possessing no intrinsic curvature, thus allowing a separation of the effects owing to extrinsic geometry. In ref. 33,

polystyrene-*block*-poly(ethylene-*alt*-propylene) diblock copolymers were annealed on a substrate with trenches of vanishing Gaussian curvature, and it was clearly shown that the in-plane striped pattern tends to orient itself perpendicularly to the trenches. In ref. 32, the same type of block copolymer were deposited on more topographically diverse substrates, and the same tendency was seen.

In ref. 34 the authors perform experiments with polystyrene-*block*-poly(ethylene-*alt*-propylene) diblock copolymers on both a ridge-like geometry (with vanishing Gaussian curvature) and a bumpy geometry consisting of numerous Gaussian-like smooth bumps. For the cylindrical geometry, they find that the block copolymer cylinders tend to align along the direction of curvature. For the bumpy substrate, they find that both directions of principal curvature constitute preferred orientations for the block copolymer pattern. This is in accordance with what our model predicts, as will be shown in this paper, namely that stripes preferentially align with curvature.

The paper is organized as follows. In Section 2, we discuss the free energy functional which is the starting point of our description and the motivation for developing a finite-thickness model. In Section 3, we first describe the general strategy and then perform the expansion. Section 4 is devoted to studying pattern formation on different geometries by numerical simulations of the model. We go on to extract the general features of the ordering mechanism and discuss the implications for pattern formation. In Section 5 we make our closing remarks.

2 The free energy

The Brazovskii model³⁵ and closely related Phase Field Crystal models^{36,37} have been applied to a broad range of systems undergoing pattern formation and selection of a specific length scale. These Brazovskii-type models have previously been employed to describe block copolymers^{29,32,38–40} but the approach is very general and also nucleation and pattern formation processes,^{41–43} crystal defect dynamics,^{36,44–46} grain boundary melting^{47–49} and liquid crystals^{50,51} have been studied. A Brazovskii-type model was also famously shown by Swift and Hohenberg⁵² to describe Rayleigh–Bénard convection.

The Brazovskii mean field free energy $F(\psi)$ is a Ginzburg–Landau expansion in the order parameter $\psi(\mathbf{x})$. We will work with the corresponding free energy density $f = F/V$:

$$f(\psi) = \frac{1}{V} \int dV \left[2(\nabla^2 \psi)^2 - 2|\nabla \psi|^2 + \frac{\tau}{2} \psi^2 + \frac{1}{4} \psi^4 \right], \quad (1)$$

where V is volume, $\psi(\mathbf{x}) = \phi(\mathbf{x}) - \phi_0$ measures the local deviation from the average composition ϕ_0 at the critical temperature T_c . The model has one parameter, the reduced temperature $\tau = (T_c - T)/T_c$. We assume $\phi_0 = 0$ throughout, since we study the compositionally symmetric lamellar phase.

The negative sign of the gradient-squared in eqn (1) makes spatial modulations of the order parameter field ψ energetically favorable. In combination with the positive Laplacian-squared,

the gradient-squared favors a specific wavelength $\lambda = 2\pi\sqrt{2}$. To see this, consider the free energy density of a field $\psi = \psi_0 \sin(q_0 x)$:

$$f(\psi) = (q_0^4 - q_0^2)\psi_0^2 + \frac{\tau}{2}\psi_0^2 + \frac{3}{32}\psi_0^4. \quad (2)$$

The free energy density is minimized for $q_0 = 1/\sqrt{2}$ resulting in a characteristic wavelength $\lambda = 2\pi\sqrt{2}$. Any deviation from this spacing of the stripe pattern is energetically penalized. Since Brazovskii-type models favour a single wavelength, the simulated profiles most closely resemble block copolymers in the weak segregation limit where the composition profile (density of either component) is approximately sinusoidal,⁵³ but the patterns themselves are more general.

In the current work, we focus on thin films on curved surfaces. A simple way to describe the free energy of the thin film is to consider the two-dimensional surface version of eqn (1) where all derivatives have been replaced with their covariant surface equivalents:

$$f(\psi) = \frac{1}{A} \int d\tilde{A} \left[2(\tilde{\nabla}^2 \psi)^2 - 2|\tilde{\nabla} \psi|^2 + \frac{\tau}{2} \psi^2 + \frac{1}{4} \psi^4 \right]. \quad (3)$$

Here, $\tilde{\nabla}$ denotes a covariant surface derivative on the surface \tilde{S} and $d\tilde{A}$ is the area element of the curved surface. The strategy of replacing bulk derivatives with their surface equivalent has been applied to crystallization on curved surfaces using the related Phase Field Crystal model^{40,54} as well as in treatments of nematic crystals on curved surfaces using the Frank energy.^{31,55,56} Replacing the bulk derivatives with their surface equivalents preserves the optimal wavelength λ . This covariant formulation introduces a coupling between stripe orientation and intrinsic curvature which is geometrically clear: if the surface is intrinsically curved in some direction, the stripe pattern will effectively be stretched (or compressed) along this direction. Stretching the pattern along the stripe direction does not affect the spacing, while stretching orthogonal to the stripes does. Thus the effect is to align the stripes along the direction of maximal intrinsic curvature.

However, this approach does not take the third dimension into account and results, for example, in all stripe orientations on a cylinder being equally favorable, since the cylinder has vanishing intrinsic (Gaussian) curvature. This is not a proper description of the striped phase, which can be seen by considering Fig. 1. Whereas all layers in Fig. 1c have the same wavelength, this is not the case for the configuration in Fig. 1a, where the wavelength increases with the radial coordinate. Thus the configuration in Fig. 1c should have the lowest free energy.

To properly account for the third dimension, we will start with the three-dimensional free energy density in eqn (1) and expand it in the thickness of the film, to obtain a two-dimensional free energy density which takes both the intrinsic and extrinsic curvature of the surface into account.

3 Coupling between substrate curvature and texture orientation

3.1 Geometrical setup

We consider a thin three-dimensional region Ω of thickness h around a regular compact surface \tilde{S} – see Fig. 2. We define $\tilde{\mathbf{n}}$ to be the unit normal vector field to the surface \tilde{S} . The volume Ω is described by the three-dimensional position vector $\mathbf{p}(u, w, \xi)$ parametrized by the three internal parameters (u, w, ξ) :

$$\mathbf{p}(u, w, \xi) = \tilde{\mathbf{p}}(u, w) + \xi \tilde{\mathbf{n}}(u, w), \quad (4)$$

where $\tilde{\mathbf{p}}$ is the normal projection of the point \mathbf{p} onto \tilde{S} . The distance between \mathbf{p} and the surface \tilde{S} along the normal $\tilde{\mathbf{n}}$ at a point $\tilde{\mathbf{p}}$ is given by $|\xi|$. The surface is of thickness h and thus $\xi \in [-h/2; h/2]$.

The tangent vectors at the point $\tilde{\mathbf{p}}(u, w) \in \tilde{S}$ are

$$\tilde{\mathbf{a}}_i = \partial_i \tilde{\mathbf{p}}, \quad (5)$$

where the tilde indicates that the tangent vectors belong to the surface \tilde{S} and the index i runs over the reference coordinates u and w .

The induced metric (first fundamental form) on the surface \tilde{S} is

$$\tilde{g}_{ij} = \tilde{\mathbf{a}}_i \cdot \tilde{\mathbf{a}}_j, \quad (6)$$

where \cdot indicates the standard Euclidean inner product in \mathbb{R}^3 . The metric determinant will be denoted \tilde{g} . The metric inverse is \tilde{g}^{ij} and defined such that $\tilde{g}^{ik} \tilde{g}_{kj} = \delta_j^i$ where repeated indices indicate summation (Einstein convention). The metric and its inverse can be used to raise and lower indices. The curvature tensor† (second fundamental form) of the surface \tilde{S} is:

$$K_{ij} = \tilde{\mathbf{n}} \cdot \partial_i \tilde{\mathbf{a}}_j. \quad (7)$$

We denote the two principal curvatures at a point $\tilde{\mathbf{p}}$ as $\kappa_1(\tilde{\mathbf{p}})$ and $\kappa_2(\tilde{\mathbf{p}})$ respectively. If we define the local curvature length scale, $l(\tilde{\mathbf{p}})$ and the global curvature length scale ℓ as:

$$l(\tilde{\mathbf{p}}) = \min \left[\frac{1}{\kappa_1(\tilde{\mathbf{p}})}, \frac{1}{\kappa_2(\tilde{\mathbf{p}})} \right], \quad \ell = \min_{\tilde{\mathbf{p}} \in \tilde{S}} l(\tilde{\mathbf{p}}), \quad (8)$$

then the requirement of the volume V being a thin shell can be formulated as

$$\left(\frac{h}{\ell} \right)^2 \ll 1. \quad (9)$$

We consider a scalar order parameter ψ which is constant throughout the thickness of the shell – something which will be important once we derive the effective two-dimensional free energy:

$$\psi(\tilde{\mathbf{p}} + \xi \tilde{\mathbf{n}}) = \psi(\tilde{\mathbf{p}}) \quad \text{for all } \tilde{\mathbf{p}} \in \tilde{S}, \xi \in [-h/2; h/2].$$

Our model is thus valid for block copolymer configurations which are approximately homogeneous over the thickness of the (thin) film. This assumption could be violated by *e.g.* asymmetric

† Not to be confused with either the Riemann or Ricci curvature tensors which are purely intrinsic.

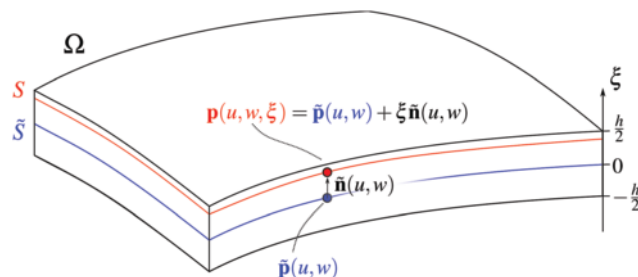


Fig. 2 The geometric setup for the expansion. Ω is the entire three-dimensional volume of the film while \tilde{S} defines the midsurface (blue) and S represents a surface (red) within Ω , separated from \tilde{S} by a distance ξ along the normal vector $\tilde{\mathbf{n}}$.

polymers or polymers which appreciably change conformation when placed under compression or dilation.

The goal is to express the three-dimensional free energy density described by eqn (1) in terms of the curvature tensor in eqn (7), invariants of the surface \tilde{S} such as the mean curvature H and the Gaussian curvature K and surface covariant derivatives of the fields. We write it as an expansion in the surface normal coordinate ξ . Once this has been done, the surface height $\xi \in [-h/2; h/2]$ can be integrated out to arrive at an effective two-dimensional free energy density to lowest order in the surface thickness to curvature ratio h/ℓ .

3.2 Expansion of the free energy density

Geometrically, the curvature tensor $K_j^i = g^{im} K_{mj}$ expresses the rate of change of the normal vector projected onto the surface:

$$K_{ij} = \tilde{\mathbf{n}} \cdot \partial_i \tilde{\mathbf{a}}_j = -\tilde{\mathbf{a}}_i \cdot \partial_j \tilde{\mathbf{n}} \quad (10)$$

In our analysis we will assume the curvature to vary slowly compared to the characteristic wavelength such that gradients of the curvature tensor can be neglected.

When combining eqn (4) with $\mathbf{a}_i = \partial_i \mathbf{p}$ it is clear that the tangent vectors of the surface S will involve the extrinsic curvature. The metric tensor $g_{ij} = \mathbf{a}_i \cdot \mathbf{a}_j$ then inherits this dependence on curvature:

$$g_{ij} = \tilde{g}_{ij} - 2\xi K_{ij} + \xi^2 K_i^k K_{kj}. \quad (11)$$

This expression is exact. To second order in the small quantity ξ/ℓ the inverse metric g^{ij} takes the following form:

$$g^{ij} = (1 - 3\xi^2 K) \tilde{g}^{ij} + 2(\xi + 3\xi^2 H) K^{ij} + \mathcal{O}(\xi/\ell)^3. \quad (12)$$

The curvature will then enter into the $|\nabla\psi|^2$ and $(\nabla^2\psi)^2$ terms of the free energy through the metric. However, the volume element itself is also affected. The invariant volume element in differential geometry is $\sqrt{g} d^d x$ where g is the determinant of the metric. We imagine the volume Ω to be foliated by a series of surfaces S , each being a level set of ξ described by (4) such that $\xi = 0$ corresponds to the midsurface \tilde{S} . The volume of Ω is denoted by V while the surface areas of S and \tilde{S} are denoted by A and \tilde{A} , respectively. The volume element $dV = d\xi dA$ depends on ξ through the area element dA . This dependence follows from

the expansion of the metric determinant in terms of ξ , which is given by:⁵⁷

$$\sqrt{g} = J_\xi \sqrt{\tilde{g}}, \quad J_\xi = 1 - 2H\xi + K\xi^2. \quad (13)$$

The area element of S is then $dA = J_\xi d\tilde{A}$. It follows that the total volume V which enters the free energy density is given by

$$V = \int_\Omega dV = \int_{-h/2}^{h/2} d\xi \int_S d\tilde{A} J_\xi = \tilde{A}h + \frac{h^3}{12}\chi,$$

where $\chi \equiv \int_S d\tilde{A} K$ is the integrated Gaussian curvature which, by the Gauss–Bonnet theorem, equals 2π times the Euler characteristic for a closed surface.

The gradient-squared term is expanded as

$$J_\xi |\nabla\psi|^2 = J_\xi g^{ij} \nabla_i \psi \nabla_j \psi = |\tilde{\nabla}\psi|^2 + c_1 \xi + c_2 \xi^2 + c_3 \xi^3 + \mathcal{O}(\xi/\ell)^4$$

where

$$c_2 = 2HKK^{ij} \tilde{\nabla}_i \psi \tilde{\nabla}_j \psi - 2K |\tilde{\nabla}\psi|^2.$$

The odd terms (c_1 and c_3) will not contribute to the effective 2D theory, since they integrate to zero over $\xi \in [-h/2, h/2]$.

The Laplacian-squared term can be expanded similarly to yield

$$J_\xi (\nabla^2 \psi)^2 = (\tilde{\nabla}^2 \psi)^2 + d_1 \xi + d_2 \xi^2 + d_3 \xi^3 + \mathcal{O}(\xi/\ell)^4$$

where the relevant curvature coupling term is given by

$$d_2 = 4(K^{ij} \tilde{\nabla}_i \psi \tilde{\nabla}_j \psi)^2 + 4H(K^{ij} \tilde{\nabla}_i \psi \tilde{\nabla}_j \psi) \tilde{\nabla}^2 \psi - 5K(\tilde{\nabla}^2 \psi)^2.$$

The last ingredient in the expansion is the local, polynomial part of the free energy. This depends on ξ only due to the metric determinant as it appears in the volume element.

The final effective two-dimensional energy up to and including order $(h/\ell)^3$ takes the form

$$f = \frac{h\tilde{A}}{V} \{f_{\tilde{S}} + k_{\tilde{S}}\} \quad (14)$$

Here, $f_{\tilde{S}}$ is the covariantized Brazovskii surface free energy as given by eqn (3), to which this effective energy reduces when $h \rightarrow 0$. The correction $k_{\tilde{S}}$ due to a finite film thickness is given by:

$$k_{\tilde{S}} = \frac{h^2}{12\tilde{A}} \int_{\tilde{S}} d\tilde{A} \left[8 \left((K^{ij} \tilde{\nabla}_i \tilde{\nabla}_j \psi)^2 + H(K^{ij} \tilde{\nabla}_i \tilde{\nabla}_j \psi) (\tilde{\nabla}^2 \psi) \right) - 10K(\tilde{\nabla}^2 \psi)^2 - 4 \left(HK^{ij} (\tilde{\nabla}_i \psi) (\tilde{\nabla}_j \psi) - K |\nabla\psi|^2 \right) + K \left(\frac{\tau}{2} \psi^2 + \frac{1}{4} \psi^4 \right) \right].$$

3.3 Relaxation towards equilibrium

We now turn to the dynamics of $\psi = \psi(\mathbf{p})$, the order parameter restricted to the midsurface. Assuming a conserved order parameter field, the relaxation in time t towards equilibrium can be described by the equation

$$\frac{\partial \psi}{\partial t} = M \tilde{\nabla}^2 \frac{\delta f}{\delta \psi}, \quad (15)$$

where M is a diffusion coefficient which sets the time scale of the dynamics, and $\delta f/\delta \psi$ denotes the functional derivative of the free energy in (14).

To study the effects of curvature on non-trivial surfaces, we solve the dynamic eqn (15) numerically using a finite element method in space and an implicit finite difference scheme in time. The numerical method is described in detail in Appendix A.

Note that the equation of motion (15) guarantees that the free energy f decreases in time, and thus the system will eventually reach at least a local free energy minimum. However, when the initial state $\psi(\mathbf{p}, t=0)$ is sufficiently disorganized, this local minimum state may be far from the global minimum in the free energy, which we typically seek. To get closer to this state, a cyclic annealing procedure, as outlined by Zhang *et al.*,⁴⁰ was implemented. This procedure is described in more detail in Section 4.1.

4 The effects of curvature as an ordering field

In order to understand how to design substrates in order to obtain specific textures, it is necessary to understand how curvature acts as an ordering field for the striped phase in specific geometries. In this section we will study the low-energy texture configurations on qualitatively different curved surfaces which exemplify the distinct configurations of Gaussian curvature K and mean curvature H . The surfaces considered are the cylinder ($K=0, H \neq 0$), the saddle geometry ($K \neq 0, H=0$) and the Gaussian bump ($H \neq 0, K \neq 0$).

Before delving into the details of the mechanism on model geometries, we have simulated the model on a random landscape of sinusoidal bumps and ridges, see Fig. 3. As film

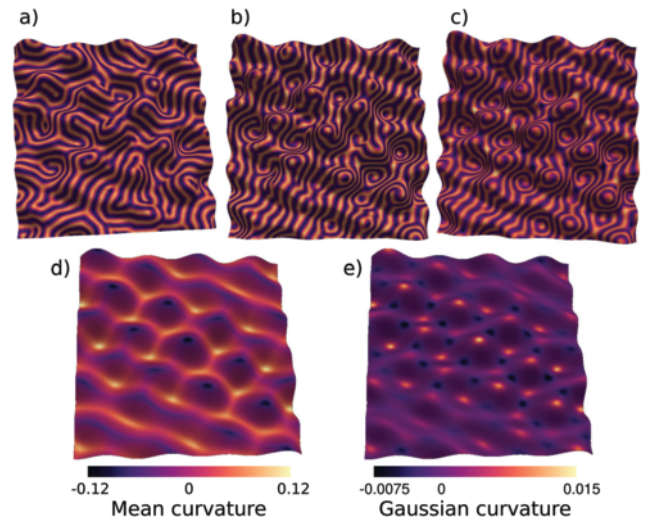


Fig. 3 Effect of thickness on pattern formation. Shown above are the results of quenching simulations on a random landscape for (a) $(h/\ell)^2 = 0$, (b) $(h/\ell)^2 = 0.49$ and (c) $(h/\ell)^2 = 1.25$ respectively. Note that the rightmost plot corresponds to $h/\ell > 1$ and thus lies outside the thin shell regime. (d) Mean curvature. (e) Gaussian curvature. Open boundary conditions were employed, see Appendix A for details.

thickness increases from left to right in the figure, it becomes clear that there is a tendency for the stripes to run perpendicularly to the ridge-like features and to encircle the bumps.

With this intuition in mind, let us turn to the study of pattern formation on simpler model surfaces. We take the surface to be locally parametrized in terms of coordinates $x^1 = u$ and $x^2 = w$ and that the coordinate curves are chosen as lines of curvature, rendering the metric as well as the curvature tensor diagonal. The curvature tensor may then be completely specified by the mean curvature H and Gaussian curvature K :

$$K_j^i = \begin{bmatrix} H \pm \sqrt{H^2 - K} & 0 \\ 0 & H \mp \sqrt{H^2 - K} \end{bmatrix}, \quad (16)$$

In order to see the role of intrinsic and extrinsic geometry separately in the finite-thickness energy contribution, it is instructive to study the model in the extremal cases of vanishing Gaussian curvature ($K = 0, H \neq 0$) and vanishing mean curvature ($K \neq 0, H = 0$), respectively. Below we study two simple examples of such extremal geometries and solve for the preferred pattern orientation, namely the cylinder and the saddle geometry.

We can parametrize a one-mode stripe pattern on a surface (such as those of Fig. 1) as

$$\psi = \psi_0 \cos[k_0(\cos(\alpha)s_u + \sin(\alpha)s_w)] \quad (17)$$

where s_u and s_w are the surface arc lengths along the coordinate curves of u and w , which are assumed orthogonal. In the next section we will use this expression to derive preferred orientations on different geometries.

4.1 Cylinder

The cylinder is an example of a geometry satisfying $K = 0, H \neq 0$ as described above. In this case the curvature effects are entirely extrinsic in nature, meaning that the finite-thickness contribution to the energy is crucial in breaking the symmetry between all the possible stripe orientations.

We parametrize a cylindrical surface \tilde{S} of radius R and length L by the cylindrical coordinates $\theta \in [0, 2\pi], z \in [0, L]$:

$$\tilde{\mathbf{p}}(\theta, z) = [R \cos(\theta) \quad R \sin(\theta) \quad z]^T \quad (18)$$

Relevant geometrical quantities associated for this specific parametrization are:

$$\tilde{g}_{ij} = \begin{bmatrix} \tilde{g}_{\theta\theta} & \tilde{g}_{\theta z} \\ \tilde{g}_{z\theta} & \tilde{g}_{zz} \end{bmatrix} = \begin{bmatrix} R^2 & 0 \\ 0 & 1 \end{bmatrix},$$

$$K_{ij} = \begin{bmatrix} R & 0 \\ 0 & 0 \end{bmatrix},$$

$$K = 0, \quad H = 1/(2R).$$

As in ref. 29, we consider a striped texture making an angle α with the axis of the cylinder, see Fig. 1. By applying eqn (17), we obtain the following expression:

$$\psi(\theta, z) = \psi_0 \cos[q_0(R\theta \cos(\alpha) + z \sin(\alpha))].$$

Since the Gaussian curvature vanishes everywhere, the curvature contribution to the free energy in eqn (14) reduces to:

$$k_{\tilde{S}} = \frac{1}{12} \left(\frac{h}{R} \right)^2 \psi_0^2 \cos^4(\alpha).$$

where we have inserted the preferred wavenumber $q_0 = 1/\sqrt{2}$ in the last step. The curvature contribution to the energy is minimized when $\alpha = \pi/2$ and the stripes on every parallel surface are able to maintain the preferred lattice spacing q_0 , as shown in Fig. 1c. This specific result for the cylinder has previously been derived²⁹ and is in agreement with the observation of stripe textures running perpendicular to substrate ridges.^{32,33}

We have simulated the model on a cylinder and measured the angle of the gradient $\tilde{\nabla}\psi$, which is of course perpendicular to the stripes. The angle histograms for the $h = 0$ (vanishing thickness) and $h > 0$ cases are shown in Fig. 4. The tendency for stripes to run around the cylinder is very clear.

To obtain such a clear result, it was necessary to perform a cyclical heat treatment – annealing – in order to decrease the number of dislocations and reach a low-energy state. Our annealing protocol consists of cycling sinusoidally between a low temperature ($\tau = 0.1$) and a high temperature ($\tau = 0.99$) which lies very close to the order–disorder transition point at $\tau = 1$.

4.2 Saddle geometry

An example of the $H = 0, K \neq 0$ situation can be realized in a simple saddle geometry. This geometry can be parametrized as

$\tilde{\mathbf{p}}(x, y) = [x, y, \frac{a}{2}(y^2 - x^2)]^T$ in Monge gauge. In this case the

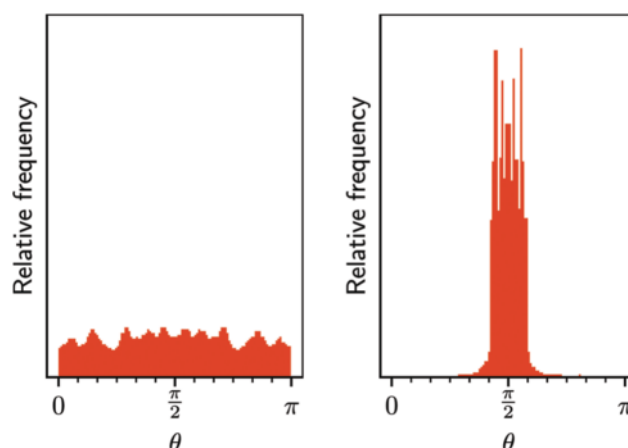


Fig. 4 Average angle distribution on the cylinder after annealing. Left: $(h/l)^2 = 0$. Right: $(h/l)^2 = 0.5$. Here $\theta = \angle(\hat{\phi}, \tilde{\nabla}\psi)$. For $h > 0$ there is a clear tendency for $\tilde{\nabla}\psi$ to be oriented in the axial direction (along $\hat{\mathbf{z}}$), meaning that the stripes tend to run around the cylinder (along $\hat{\phi}$). The histograms are based on 5 (left) and 17 (right) simulations. Periodic boundary conditions were employed.

metric and curvature tensor is

$$\tilde{g}_{ij} = \begin{bmatrix} 1 + a^2x^2 & -a^2xy \\ -a^2xy & 1 + a^2y^2 \end{bmatrix},$$

$$K_{ij} = \frac{1}{\sqrt{1 + a^2(x^2 + y^2)}} \begin{bmatrix} -a & 0 \\ 0 & a \end{bmatrix}$$

We will focus on the saddle point $x = y = 0$, where these tensors reduce to $g_{ij} = \text{diag}(1,1)$ and $K_{ij} = \text{diag}(-a,a)$. Applying (17), we get an expression for a stripe pattern:

$$\psi(x,y) = \psi_0 \cos[q_0(s(x)\cos(\alpha) + s(y)\sin(\alpha))]$$

where

$$s(x) = \int_0^x \sqrt{g_{xx}} dx' = \frac{1}{2}x\sqrt{1 + a^2x^2} + \frac{1}{2a} \sinh^{-1}(ax)$$

In this case the finite-thickness energy density $k_{\tilde{S}}$ as a function of azimuthal angle α reduces to

$$k_{\tilde{S}} = \frac{h^2}{3A} \psi_0^2 \cos(4\alpha) + \text{const.}$$

We see that the minimum energy configuration occurs for $\alpha = \pm\pi/4$. This fits well with what we find in simulations of the striped phase on a saddle geometry, see Fig. 5.

4.3 Gaussian bump

A Gaussian bump can be parametrized in the following way in Monge gauge:

$$\tilde{\mathbf{p}}(r, \phi) = [r \cos(\phi) \quad r \sin(\phi) \quad h_0 \exp[-r^2/(2\sigma^2)]]^T \quad (19)$$

with $r \geq 0$, $\phi \in [0, 2\pi]$.

In general the Gaussian bump has mean curvature as well as Gaussian curvature. However, at the ring given by $r = \sigma$, the Gaussian curvature vanishes and the surface is locally equivalent to a cylinder, with r corresponding to the axis direction and ϕ to the azimuthal direction. The model proposed in this paper therefore predicts that there should be a local tendency for the stripes to run around the circle as if it was a cylinder.

For a more global view, we must investigate the curvatures of the Gaussian bump in its entirety (not just the $K = 0$ circle). The two principal curvature directions of the bump are given by $\hat{\mathbf{r}}$ and $\hat{\phi}$. Whenever two non-zero principal curvatures are present, there will correspondingly exist two stripe orientations which corresponds to local minima of the curvature energy. If properly annealed, the stripes should align along the direction of greatest curvature – however this tendency is of course more pronounced when the two curvatures are markedly different.

To study the relative strength of the two minimal-energy orientations, we form the ratio of the two principal curvatures $F(\tilde{r}, \tilde{h}) = \frac{\kappa_r}{\kappa_\phi}$ where we have defined the two dimensionless quantities $\tilde{r} = r/\sigma$ and $\tilde{h} = h_0/\sigma$. One finds:

$$F(\tilde{r}, \tilde{h}) = \frac{(1 + \tilde{r})(1 - \tilde{r})}{1 + (\tilde{h}\tilde{r})^2 e^{-\tilde{r}^2}} \quad (20)$$

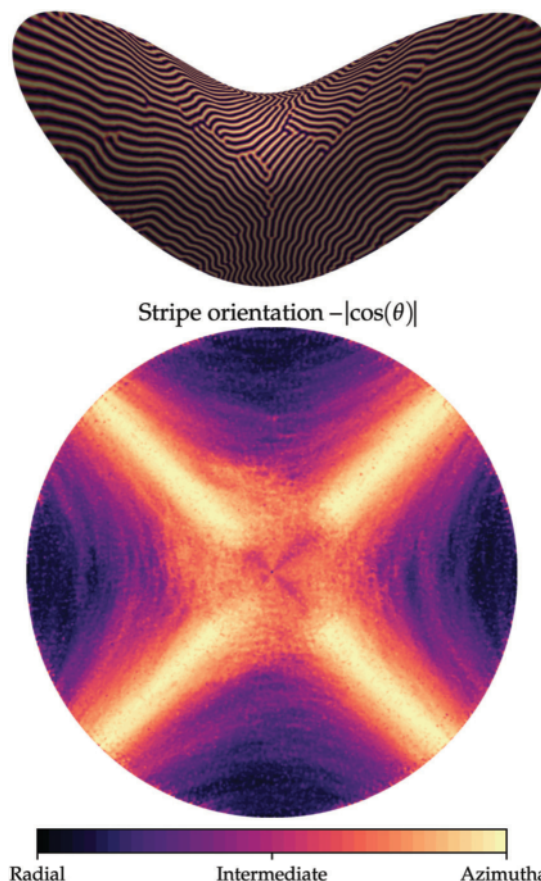


Fig. 5 Stripe textures in the vicinity of a saddle point. Top: A tilted view of a representative stripe pattern on the saddle geometry, reached by annealing. Bottom: Spatial angle distribution, averaged over 12 runs such as the one shown in the top figure. The color denotes $|\cos(\theta)|$ with $\theta = \angle(\tilde{\mathbf{r}}, \nabla\psi)$ being the angle between the radial vector and the gradient of ψ along the curved surface. These simulations were run with $(h/l)^2 = 0.23$. Open boundary conditions were employed, see Appendix A for details.

This ratio is plotted in Fig. 6 as a function of \tilde{r} for several values of the height-to-width ratio \tilde{h} . We see that F drops quickly as a function of \tilde{r} when \tilde{h} is large, corresponding to a tall and narrow Gaussian bump. Thus we should see the strongest tendency to orient the stripes azimuthally for narrow and tall Gaussian bumps.

In Fig. 7, the average orientation on the Gaussian bump is shown, with the stripe pattern quite clearly displaying a tendency to run azimuthally (*i.e.* around the bump). As with the cylinder, these were obtained by annealing.

5 Discussion and conclusions

Stripe textures of copolymers have frequently been modeled as two-dimensional nematic crystals with a one constant Frank free energy.^{31,55,56} However, due to the use of surface derivatives, this approach does not take the extrinsic geometry into account, and results in all orientations on the cylinder being equivalent. Napoli and Vergori⁵⁸ considered the influence of

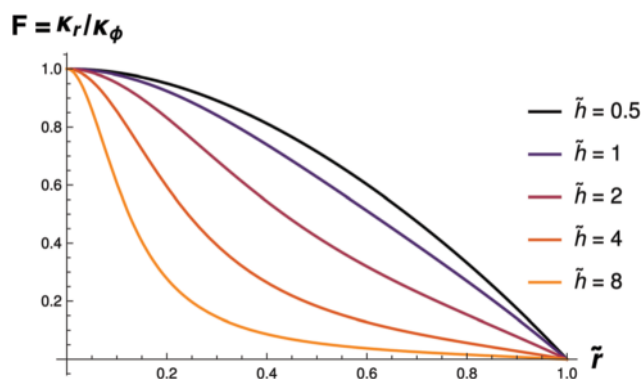


Fig. 6 Ratio of the principal curvatures on the Gaussian bump. The azimuthal curvature κ_ϕ is always strongest, regardless of $\tilde{r} = r/\sigma$ and $\tilde{h} = h/\sigma$.

extrinsic geometry on a nematic phase on a curved surface, by expanding the three-dimensional Frank elastic energy of nematic crystals to zeroth order in the small parameter ξ/l following the technique described in ref. 59. They also considered a cylindrical surface and found that the absolute free energy minimum occurs when the director field is aligned with the axis ($\alpha = 0$). Thus a similarity between the nematic description and ours arises if the director field is identified with the direction of $\vec{\nabla}\psi$. A nematic description was also employed by Mbangwa *et al.*⁶⁰ who studied defects in a nematic phase on a catenoid, and by Segatti *et al.*⁶¹ who investigated the behavior of the model by Napoli and Vergori^{58,59} on a torus. The block copolymer textures considered in this paper are smectic, rather than nematic, but preferred orientations will in many cases turn out to be similar to those obtained in Napoli and Vergori's model,⁵⁸ due to their model penalizing normal curvature of the director field. However their approach also

penalizes any geodesic torsion of the director, something which does not arise in our model.

Hexemer³⁰ experimentally studied triblock co-polymer films on an approximately Gaussian bump. In their study, they focused on the orientation at the ring of vanishing Gaussian curvature and found that the stripes tend to be perpendicular to the circle at $r = \sigma$. Our model describes only diblock copolymers, but their result points to the possibility to extend this type of substrate curvature analysis to triblock systems, which have quite different mechanical properties from diblock copolymers.⁶²

Gómez and Vega⁶³ and Garca *et al.*⁶⁴ found that the strain induced by positive (negative) defects could be reduced by positive (negative) Gaussian curvature, albeit in the $h = 0$ case (*i.e.* considering only intrinsic geometry). This echoes the defect formation seen on and between the bumps in Fig. 3. Thus we see that curvature may stabilize defects, which in turn have long-range effects on the pattern formed. In simulations, the local influence of defects can sometimes overpower the organizing effect of curvature, and thus annealing is usually necessary. Even with annealing, the ordering effect of extrinsic curvature often only becomes significant for relatively large ratios of film thickness to substrate radius of curvature, on the order of $(h/l)^2 \sim 0.1$ –1. This should be contrasted with the fact that our perturbative approach is strictly speaking limited to thin films and moderate curvatures – *i.e.* systems for which the film thickness does not exceed the local radius of curvature. Experiments are often conducted outside this regime – consider *e.g.* the experiments of Hexemer³⁰ in which the film is several layers thick.

Furthermore, annealing on highly curved substrates has been shown in some cases to lead to the formation of dewetted regions.³³ In Vega *et al.*,³³ orientation was observed to be random close to these dewetted regions. While this phenomenon can thus clearly affect ordering, it is not captured by our model.

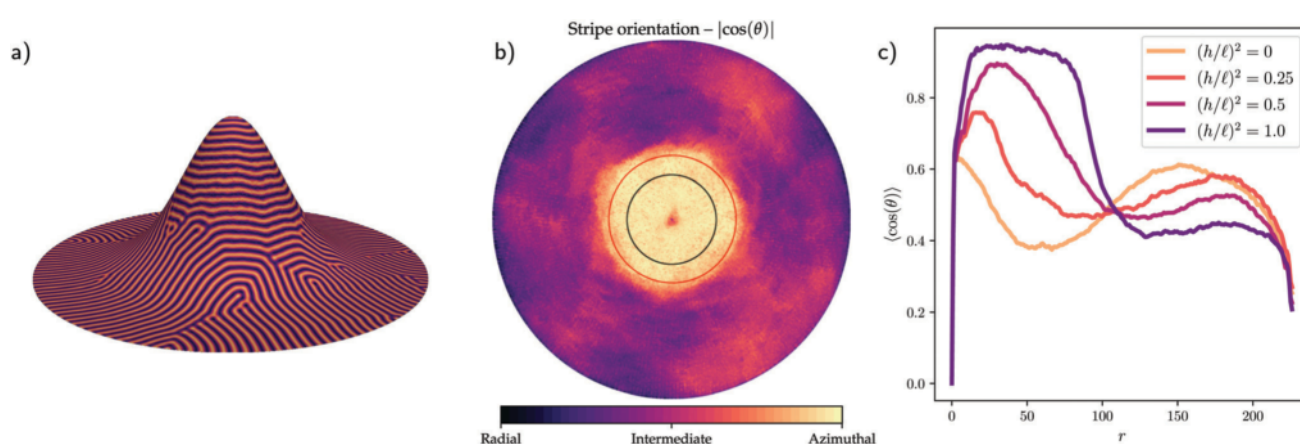


Fig. 7 Pattern formation on the Gaussian bump. (a) Stripe pattern formed on Gaussian bump after annealing at $(h/l)^2 = 1.0$. (b) Average orientation on the Gaussian bump. The color denotes $|\cos(\theta)|$ with $\theta = \angle(\hat{r}, \vec{\nabla}\psi)$ being the angle between the radial vector and the gradient of ψ along the curved surface. There is a clear tendency for stripes to run azimuthally, *i.e.* “around the bump”. The black circle represents the curve $r = \sigma$ where the Gaussian curvature vanishes. The red circle represents the curve where the two principal curvatures have equal strength, *i.e.* $|\kappa_r/\kappa_\phi| = 1$. For this particular bump, $\tilde{h} = h_0/\sigma = 3$. This plot is an average over 12 simulations run with $(h/l)^2 = 1.0$. (c) Average orientation of stripes as a function of radial coordinate. The strength of the orientation effect is clearly controlled by the ratio $(h/l)^2$. In these simulations, open boundary conditions were employed. See Appendix A for details.

To conclude, our study finds that curvature has an important effect on pattern formation in thin film block copolymers. The model that we have developed shows that it is necessary to take the effects of extrinsic curvature into consideration – effects which become apparent due to the finite thickness of the film. Through analysis of three geometries exhibiting distinct signatures of mean *vs.* Gaussian curvature, we conclude that the general tendency is for stripes to align with the direction of maximal curvature. This simple principle provides a straightforward recipe for optimizing the substrate to favour a desired pattern.

Conflicts of interest

The authors have no conflicts to declare.

Appendix A: numerical method

To solve the relaxation dynamics (15) with the free energy functional (14), we have used a numerical framework⁶⁵ developed by the authors for solving partial differential equations on arbitrary parametrized surfaces. The numerical framework is built as a layer on top of the FEniCS/Dolfin framework^{66,67} for solving PDEs using the finite element method. FEniCS interfaces to *e.g.* PETSc⁶⁸ for solving large sparse linear systems arising in the finite element method. Our numerical framework is not constrained to the equations of motion described here, but the full functionality will be documented and published elsewhere. In particular, the framework is accessed using Python and supports arbitrary analytical surface parametrizations. Surface derivatives, which enter in the metric curvature tensor fields, are computed symbolically using SymPy,⁶⁹ thereby achieving accuracy only limited by the interpolation onto the unstructured spatial mesh.

In order for the results to be directly reproducible by the reader, the numerical cases presented here are found as example scripts at the GitHub repository https://github.com/gautelinga/surface_pfc/SoftMatter2019.

A.1 Functional derivative of the free energy

The functional derivative of the free energy f which enters in the equation of motion (15), is given by

$$\frac{\delta f}{\delta \psi} = \frac{h}{V} \mu, \quad (21)$$

where the chemical potential $\mu = \mu_0 + h^2 \mu_2$ can be decomposed into

$$\mu_0 = W'(\psi) + 4\tilde{\Delta}\psi + 4\tilde{\Delta}^2\psi, \quad (22a)$$

$$\mu_2 = \frac{1}{12} K W''(\psi) + Q_1[\psi] + Q_2[\psi]. \quad (22b)$$

Here, we have introduced the operators

$$\tilde{\Delta}f = \tilde{\nabla}_i \tilde{\nabla}^i f \quad (\text{surface Laplacian}) \quad (23)$$

$$\tilde{\Omega}f = \tilde{\nabla}_i (K^{ij} \tilde{\nabla}_j f) \quad (\text{curvature Laplacian}) \quad (24)$$

and further

$$Q_1[f] = \frac{2}{3} [H\tilde{\Omega}f - K\tilde{\Delta}f], \quad (25)$$

$$Q_2[f] = \frac{1}{3} [4\tilde{\Omega}^2 f - 5K\tilde{\Delta}^2 f + 2H(\tilde{\Delta}\tilde{\Omega}f + \tilde{\Omega}\tilde{\Delta}f)], \quad (26)$$

for an arbitrary scalar function f . Finally, $W'(\psi)$ denotes the derivative of the double well potential $W(\psi) = \tau\psi^2/2 + \psi^4/4$.

A.2 Time-stepping scheme

We discretized the equations of motion (15) using an implicit approach:

$$\frac{\psi^k - \psi^{k-1}}{\Delta t^k} = \tilde{M} \tilde{\nabla}^2 \mu^k, \quad (27)$$

where ψ^k approximates $\psi(\tilde{\mathbf{p}}, t^k)$, μ^k approximates μ at time t^k , and the constant mobility \tilde{M} has absorbed the prefactor in (21), *i.e.* $\tilde{M} = (h/V)M$ (compare (15)). The time step is given by $\Delta t^k = t^k - t^{k-1}$ and selected adaptively; an initial estimate is based on

$$\Delta t_*^k = \frac{c}{\max\{|\nabla \mu^{k-1}|\}} \quad (28)$$

where c is a heuristically chosen constant. An estimate like (28) is fairly standard for phase-field models, *cf.* ref. 70, and it leads to large (small) time steps when the driving forces are small (large). If the time step is still too large to achieve convergence within a few iterations, a new time step is chosen as half of the previous estimate, *i.e.* $\Delta t_*^k \rightarrow \Delta t_*^k/2$, which is repeated until convergence.

The chemical potential is given by

$$\mu^k = \mu_0^k + h^2 \mu_2^k, \quad (29)$$

where

$$\mu_0^k = \overline{W'}(\psi^k, \psi^{k-1}) + 4\tilde{\Delta}\psi^k + 4\tilde{\Delta}^2\psi^k, \quad (30a)$$

$$\mu_2^k = \frac{1}{12} K \overline{W''}(\psi^k, \psi^{k-1}) + Q_1[\psi^k] + Q_2[\psi^k], \quad (30b)$$

In (30a) and (30b), the function $\overline{W'}(\psi^k, \psi^{k-1})$ approximates the derivative of the double well potential $W'(\psi)$. Herein, we choose the fully implicit nonlinear discretization

$$\overline{W'}(\psi^k, \psi^{k-1}) = W'(\psi^k). \quad (31)$$

Apart from the terms involving W' , the model is linear. Further, with the choice (31), it can be shown that the numerical scheme satisfies a second law of thermodynamics on the discrete level; *i.e.* the discrete free energy replacing $\psi \rightarrow \psi^k$ in (14) decays in time:

$$f[\psi^k] \leq f[\psi^{k-1}]. \quad (32)$$

A.3 Boundary conditions

We choose trivial boundary conditions, *i.e.* the boundary conditions which allow us to pass from eqn (27), (29) and (30), to

the weak form (37) below by integration by parts, such that all boundary integral terms vanish. For open (non-periodic) boundaries, this implies, most importantly, that $n_i \tilde{\nabla}^i \psi^k = 0$ (which encodes a preferred 90-degree ‘‘contact angle’’ between the boundary and stripe pattern) and $n_i \tilde{\nabla}^i \mu^k = 0$ (no energy flux across the boundary). Here n_i is the boundary normal on the reference domain. It should be noted that the boundary conditions were chosen mainly for numerical convenience and boundary effects are not a main focus in this study. A more systematic study of the implications of boundary effects should be undertaken in the future.

A.4 Variational form

The problem is solved using mixed finite elements, all of which belong to the space of piecewise continuous functions. In particular, we introduce the auxiliary fields

$$\nu^k = \tilde{\Delta} \psi^k, \quad \text{and} \quad \hat{\nu}^k = \tilde{\nabla} \psi^k, \quad (33)$$

such that our trial functions are given by

$$[\psi^k, \mu^k, \nu^k, \hat{\nu}^k] \in W = (H^1(\Omega))^4. \quad (34)$$

Further, to save some notation, we define the ‘‘gradient products’’

$$\mathfrak{S}[a, b] = g^{ij} a_i b_j, \quad (35)$$

$$\mathfrak{R}[a, b] = K^{ij} a_i b_j, \quad (36)$$

where a, b are scalar fields.

The variational problem to be solved can now be posed as the following: given ψ^{k-1} , find $[\psi^k, \mu^k, \nu^k, \hat{\nu}^k] \in W$ such that for all test functions $[\chi, \eta, \zeta, \hat{\zeta}] \in W$, we have

$$0 = \int_{\tilde{\Omega}} \left[\frac{\psi^k - \psi^{k-1}}{\Delta t^k} \chi + M \mathfrak{S}[\mu^k, \chi] \right] d\tilde{S}, \quad (37a)$$

$$0 = \int_{\tilde{\Omega}} \mu^k \zeta d\tilde{S} - m, \quad (37b)$$

$$0 = \int_{\tilde{\Omega}} [\nu^k \zeta + \mathfrak{S}[\psi^k, \zeta]] d\tilde{S}, \quad (37c)$$

$$0 = \int_{\tilde{\Omega}} [\hat{\nu}^k \hat{\zeta} + \mathfrak{R}[\psi^k, \hat{\zeta}]] d\tilde{S}, \quad (37d)$$

where, in (37b):

$$m = m_{\text{NL}} + m_0 + h^2 m_2. \quad (38a)$$

The nonlinear contribution is given by

$$m_{\text{NL}}[\psi^k, \psi^{k-1}, \xi] = \int_{\tilde{\Omega}} \left(1 + \frac{h^2}{12} K \right) \overline{W'}(\psi^k, \psi^{k-1}) \xi d\tilde{S}, \quad (38b)$$

and the (zeroth and second order in h) linear contributions are given by

$$m_0[\nu^k, \xi] = 4 \int_{\tilde{\Omega}} [\nu^k \xi - \mathfrak{S}[\nu^k, \xi]] d\tilde{S}, \quad (38c)$$

$$m_2[\nu^k, \hat{\nu}^k, \xi] = \frac{1}{3} \int_{\tilde{\Omega}} [2[H\hat{\nu}^k - K\nu^k] \xi - 4\mathfrak{R}[\hat{\nu}^k, \xi] + 5K\mathfrak{S}[\nu^k, \xi] - 2H(\mathfrak{S}[\hat{\nu}^k, \xi] + \mathfrak{R}[\nu^k, \xi])] d\tilde{S}. \quad (38d)$$

At each time step k , this gives the solution vector $[\psi^k, \mu^k, \nu^k, \hat{\nu}^k]$ wherein ψ^k is used for the next time step $k + 1$.

Since the variational problem has a nonlinear contribution from m_{NL} , the problem must be linearised and solved in an inner iteration cycle at each time step. In practice, this is automatically handled in FEniCS, which automatically generates the Jacobian of the system based on the symbolic expression for $W'(\psi^k)$ to be used in a Newton method with ψ^{k-1} as an initial guess for ψ^k .

The full set of eqn (37) with eqn (38) was discretized on the reference domain, *i.e.* a linear system was found *i.e.* using

$$\int_{\tilde{\Omega}} (\bullet) d\tilde{S} = \int_{\Omega} (\bullet) \sqrt{|g|} dS. \quad (39)$$

For stability purposes (particularly when h/ℓ was large), the best convergence rate was achieved using a direct linear solver.

Acknowledgements

BFN was supported by the VILLUM Foundation through the research grant (13168). GL was partly supported by the Research Council of Norway through its Centres of Excellence funding scheme, Project No. 262644.

Notes and references

- 1 M. P. Stoykovich and P. F. Nealey, *Mater. Today*, 2006, **9**, 20–29.
- 2 X. Gu, I. Gunkel and T. P. Russell, *Philos. Trans. R. Soc., A*, 2013, **371**, 20120306.
- 3 S. Y. Yang, J. Park, J. Yoon, M. Ree, S. K. Jang and J. K. Kim, *Adv. Funct. Mater.*, 2008, **18**, 1371–1377.
- 4 Y. Gu and U. Wiesner, *Macromolecules*, 2015, **48**, 6153–6159.
- 5 H. Ahn, S. Park, S.-W. Kim, P. J. Yoo, D. Y. Ryu and T. P. Russell, *ACS Nano*, 2014, **8**, 11745–11752.
- 6 S. Y. Yang, I. Ryu, H. Y. Kim, J. K. Kim, S. K. Jang and T. P. Russell, *Adv. Mater.*, 2006, **18**, 709–712.
- 7 T. Xu, J. Stevens, J. Villa, J. T. Goldbach, K. W. Guarini, C. T. Black, C. J. Hawker and T. P. Russell, *Adv. Funct. Mater.*, 2003, **13**, 698–702.
- 8 A. Checco, A. Rahman and C. T. Black, *Adv. Mater.*, 2014, **26**, 886–891.
- 9 P. Mansky, C. Harrison, P. Chaikin, R. A. Register and N. Yao, *Appl. Phys. Lett.*, 1996, **68**, 2586–2588.
- 10 M. Park, C. Harrison, P. M. Chaikin, R. A. Register and D. H. Adamson, *Science*, 1997, **276**, 1401–1404.
- 11 V. Pelletier, K. Asakawa, M. Wu, D. H. Adamson, R. A. Register and P. M. Chaikin, *Appl. Phys. Lett.*, 2006, **88**, 211114.

- 12 Y.-R. Hong, K. Asakawa, D. H. Adamson, P. M. Chaikin and R. A. Register, *Opt. Lett.*, 2007, **32**, 3125–3127.
- 13 T. Thurn-Albrecht, J. Schotter, G. Kästle, N. Emley, T. Shibauchi, L. Krusin-Elbaum, K. Guarini, C. Black, M. Tuominen and T. Russell, *Science*, 2000, **290**, 2126–2129.
- 14 E. W. Edwards, M. F. Montague, H. H. Solak, C. J. Hawker and P. F. Nealey, *Adv. Mater.*, 2004, **16**, 1315–1319.
- 15 L. Rockford, Y. Liu, P. Mansky, T. Russell, M. Yoon and S. Mochrie, *Phys. Rev. Lett.*, 1999, **82**, 2602.
- 16 J. Y. Cheng, C. T. Rettner, D. P. Sanders, H.-C. Kim and W. D. Hinsberg, *Adv. Mater.*, 2008, **20**, 3155–3158.
- 17 D. E. Angelescu, J. H. Waller, D. H. Adamson, P. Deshpande, S. Y. Chou, R. A. Register and P. M. Chaikin, *Adv. Mater.*, 2004, **16**, 1736–1740.
- 18 T. Morkved, M. Lu, A. Urbas, E. Ehrichs, H. Jaeger, P. Mansky and T. Russell, *Science*, 1996, **273**, 931–933.
- 19 K. G. Yager, N. J. Fredin, X. Zhang, B. C. Berry, A. Karim and R. L. Jones, *Soft Matter*, 2010, **6**, 92–99.
- 20 K. Mita, H. Tanaka, K. Saijo, M. Takenaka and T. Hashimoto, *Macromolecules*, 2007, **40**, 5923–5933.
- 21 S. Park, D. H. Lee, J. Xu, B. Kim, S. W. Hong, U. Jeong, T. Xu and T. P. Russell, *Science*, 2009, **323**, 1030–1033.
- 22 R. A. Segalman, H. Yokoyama and E. J. Kramer, *Adv. Mater.*, 2001, **13**, 1152–1155.
- 23 S.-J. Jeong, J. E. Kim, H.-S. Moon, B. H. Kim, S. M. Kim, J. B. Kim and S. O. Kim, *Nano Lett.*, 2009, **9**, 2300–2305.
- 24 H. Xiang, K. Shin, T. Kim, S. Moon, T. McCarthy and T. Russell, *J. Polym. Sci., Part B: Polym. Phys.*, 2005, **43**, 3377–3383.
- 25 J. Y. Cheng, C. Ross, E. Thomas, H. I. Smith and G. J. Vancso, *Appl. Phys. Lett.*, 2002, **81**, 3657–3659.
- 26 B. H. Kim, Y. Choi, J. Y. Kim, H. Shin, S. Kim, S.-W. Son, S. O. Kim and P. Kim, *Adv. Mater.*, 2014, **26**, 4665–4670.
- 27 B. H. Kim, H. M. Lee, J.-H. Lee, S.-W. Son, S.-J. Jeong, S. Lee, D. I. Lee, S. U. Kwak, H. Jeong and H. Shin, *et al.*, *Adv. Funct. Mater.*, 2009, **19**, 2584–2591.
- 28 B. H. Kim, D. O. Shin, S.-J. Jeong, C. M. Koo, S. C. Jeon, W. J. Hwang, S. Lee, M. G. Lee and S. O. Kim, *Adv. Mater.*, 2008, **20**, 2303–2307.
- 29 A. D. Pezzutti, L. R. Gomez and D. A. Vega, *Soft Matter*, 2015, **11**, 2866–2873.
- 30 C. D. Santangelo, V. Vitelli, R. D. Kamien and D. R. Nelson, *Phys. Rev. Lett.*, 2007, **99**, 017801.
- 31 R. D. Kamien, D. R. Nelson, C. D. Santangelo and V. Vitelli, *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.*, 2009, **80**, 051703.
- 32 E. A. Matsumoto, D. A. Vega, A. D. Pezzutti, N. A. Garcia, P. M. Chaikin and R. A. Register, *Proc. Natl. Acad. Sci. U. S. A.*, 2015, **112**, 12639–12644.
- 33 D. A. Vega, L. R. Gómez, A. D. Pezzutti, F. Pardo, P. M. Chaikin and R. A. Register, *Soft Matter*, 2013, **9**, 9385.
- 34 G. T. Vu, A. A. Abate, L. R. Gómez, A. D. Pezzutti, R. A. Register, D. A. Vega and F. Schmid, *Phys. Rev. Lett.*, 2018, **121**, 087801.
- 35 S. A. Brazovskii, *Soviet Phys. JETP*, 1975, **41**, 85–89.
- 36 K. R. Elder, M. Katakowski, M. Haataja and M. Grant, *Phys. Rev. Lett.*, 2002, **88**, 245701.
- 37 K. R. Elder and M. Grant, *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.*, 2004, **70**, 051605.
- 38 K. Yamada and S. Komura, *J. Phys.: Condens. Matter*, 2008, **20**, 155107.
- 39 S. Villain-Guillot and D. Andelman, *Eur. Phys. J. B*, 1998, **4**, 95–101.
- 40 L. Zhang, L. Wang and J. Lin, *Soft Matter*, 2014, **10**, 6713.
- 41 T. Pusztai, G. Tegze, G. I. Tóth, L. Környei, G. Bansel, Z. Fan and L. Gránágy, *J. Phys.: Condens. Matter*, 2008, **20**, 404205.
- 42 K. Elder, G. Rossi, P. Kanerva, F. Sanches, S. Ying, E. Granato, C. Achim and T. Ala-Nissila, *Phys. Rev. Lett.*, 2012, **108**, 226102.
- 43 K. Elder and Z. Huang, *J. Phys.: Condens. Matter*, 2010, **22**, 364103.
- 44 A. D. Pezzutti and D. A. Vega, *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.*, 2011, **84**, 011123.
- 45 J. M. Tarp, L. Angheluta, J. Mathiesen and N. Goldenfeld, *Phys. Rev. Lett.*, 2014, **113**, 265503.
- 46 A. Skaugen, L. Angheluta and J. Viñals, *Phys. Rev. B*, 2018, **97**, 054113.
- 47 M. Bjerre, J. M. Tarp, L. Angheluta and J. Mathiesen, *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.*, 2013, **88**, 020401.
- 48 J. M. Tarp and J. Mathiesen, *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.*, 2015, **92**, 012409.
- 49 J. Mellenthin, A. Karma and M. Plapp, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2008, **78**, 184110.
- 50 H. Löwen, *J. Phys.: Condens. Matter*, 2010, **22**, 364105.
- 51 R. Wittkowski, H. Löwen and H. R. Brand, *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.*, 2010, **82**, 031708.
- 52 J. Swift and P. C. Hohenberg, *Phys. Rev. A: At., Mol., Opt. Phys.*, 1977, **15**, 319–328.
- 53 V. Castelletto and I. W. Hamley, *Curr. Opin. Solid State Mater. Sci.*, 2004, **8**, 426–438.
- 54 R. Backofen, A. Voigt and T. Witkowski, *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.*, 2010, **81**, 025701.
- 55 D. R. Nelson, *Nano Lett.*, 2002, **2**, 1125–1129.
- 56 T. Lopez-Leon, A. Fernandez-Nieves, M. Nobili and C. Blanc, *Phys. Rev. Lett.*, 2011, **106**, 247802.
- 57 M. Deserno, *Notes on Differential Geometry – with special emphasis on surfaces in \mathbb{R}^3* , 2004, https://www.cmu.edu/biolphys/deserno/pdf/diff_geom.pdf.
- 58 G. Napoli and L. Vergori, *Phys. Rev. Lett.*, 2012, **108**, 207803.
- 59 G. Napoli and L. Vergori, *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.*, 2012, **85**, 061701.
- 60 B. L. Mbanga, G. M. Grason and C. D. Santangelo, *Phys. Rev. Lett.*, 2012, **108**, 017801.
- 61 A. Segatti, M. Snarski and M. Veneroni, *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.*, 2014, **90**, 012501.
- 62 M. W. Matsen and R. Thompson, *J. Chem. Phys.*, 1999, **111**, 7139–7146.
- 63 L. R. Gómez and D. A. Vega, *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.*, 2009, **79**, 031701.
- 64 N. A. Garcia, A. D. Pezzutti, R. A. Register, D. A. Vega and L. R. Gómez, *Soft Matter*, 2015, **11**, 898–907.
- 65 G. Linga and B. F. Nielsen, *Surfaise: GitHub repository*, 2019, <https://github.com/gautelinga/surfaise/>.

- 66 A. Logg, K.-A. Mardal and G. Wells, *Automated solution of differential equations by the finite element method: The FEniCS book*, Springer Science & Business Media, 2012, vol. 84.
- 67 A. Logg, G. N. Wells and J. Hake, *Automated Solution of Differential Equations by the Finite Element Method*, Springer, 2012, pp. 173–225.
- 68 S. Balay, S. Abhyankar, M. F. Adams, J. Brown, P. Brune, K. Buschelman, L. Dalcin, V. Eijkhout, W. D. Gropp, D. Kaushik, M. G. Knepley, D. A. May, L. C. McInnes, K. Rupp, B. F. Smith, S. Zampini, H. Zhang and H. Zhang, PETSc Web page, 2017, <http://www.mcs.anl.gov/petsc>.
- 69 A. Meurer, C. P. Smith, M. Paprocki, O. Čertík, S. B. Kirpichev, M. Rocklin, A. Kumar, S. Ivanov, J. K. Moore, S. Singh, T. Rathnayake, S. Vig, B. E. Granger, R. P. Muller, F. Bonazzi, H. Gupta, S. Vats, F. Johansson, F. Pedregosa, M. J. Curry, A. R. Terrel, V. Roučka, A. Saboo, I. Fernando, S. Kulal, R. Cimrman and A. Scopatz, *PeerJ Comput. Sci.*, 2017, **3**, e103.
- 70 E. Campillo-Funollet, G. Grün and F. Klingbeil, *SIAM J. Appl. Math.*, 2012, **72**, 1899–1925.

CHAPTER 3

SPREADING AND HETEROGENEITY: COVID-19

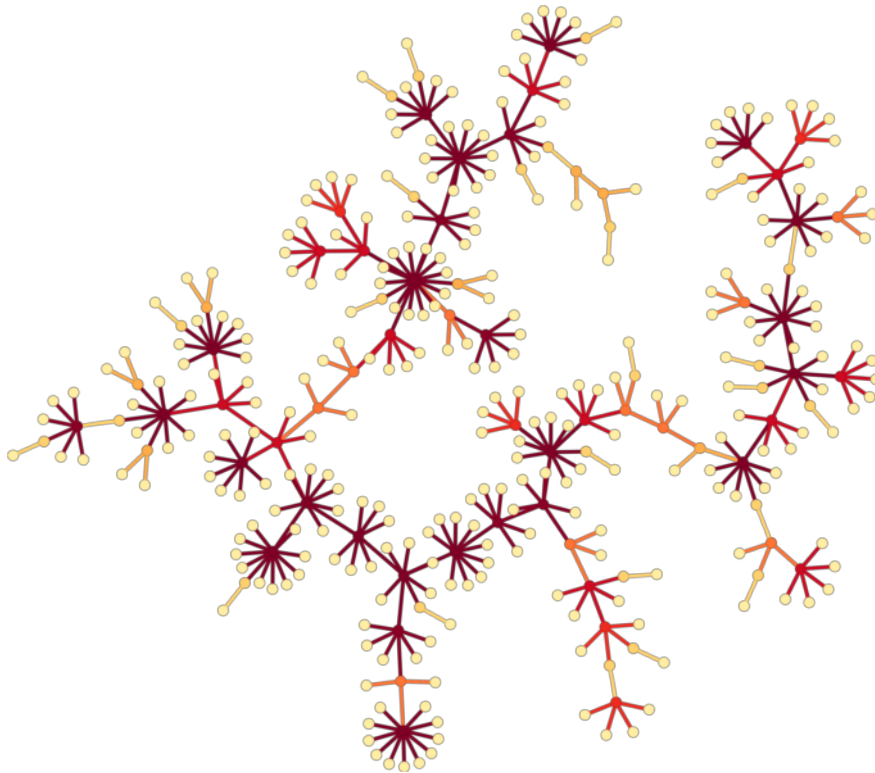


Figure 23: Simulated infection network due to a superspreading disease propagating in a restricted social network.

This is the last of the three main chapters of the thesis. It deals with a topic that is very different from the preceding two, namely dynamical modeling of disease spread, particularly as it pertains to COVID-19. The main focus of this chapter is on heterogeneity and how it impacts spreading and interacts with mitigation strategies. *Heterogeneity* is of course a very broad term, and we are being intentionally general here, since we will consider some social as well as biological sources of heterogeneity.

The bulk of the simulations in this chapter are of the agent-based – or individual-based – type. These are implemented computationally by having a number of separate agents interact, each possessing their own properties and state information. This is to be contrasted with e.g. aggregate, compartmental models such as the SIR (**S**usceptible-**I**nfectious-**R**ecovered) model, which is not formulated in terms of individual agents, but rather by stratifying the population into a few compartments, the state of each being described by a single real number – namely its occupancy.

Our reasons for choosing to work with agent-based models are several, but let me describe just a few

of them here. The first has to do with contact structures. In our work, contact structure is of great importance – the fact that people generally have a mixture of repeated and one-off (or just very seldom repeated) contacts. It is not possible to include this in aggregate compartmental models for the simple reason that the identities of individuals are not tracked, so repeat contacts cannot be ensured. In aggregate models, homogeneous mixing is generally assumed within each compartment. Formulated in agent-based terms, this would mean that any individual belonging to certain compartment has the same chance of meeting any member of some other, given compartment. Even with age stratified populations, you get statements such as “any 25-30 year-old infectious person has the same chance of infecting any 30-35 year old susceptible person”. This is an approximation which can be entirely valid in certain situations and inadequate in others. Examples of this will become clear throughout this chapter.

The first part of the chapter concerns a special kind of heterogeneity, namely *overdispersed transmission* or *superspreading*. This term covers a phenomenon where the distribution of individual reproductive numbers is very wide. In other words, some infected persons end up causing a large number of new infections, while the majority hardly infect at all. Investigations into this phenomenon have been rendered especially relevant by the emergence of SARS-CoV-2, the causative agent of the COVID-19 pandemic. This pandemic has been characterized by overdispersion, manifesting itself as superspreading. Evidence suggests that around 10% of infected individuals with COVID-19 are responsible for 80% of new cases [31, 32, 33, 34].

The second part of the chapter concerns contact tracing as a mitigation strategy, and how it is impacted by heterogeneities of real-world contact networks and social activity patterns. We have analyzed a social contact network between a cohort of university students with high temporal resolution and simulated a disease outbreak in this setting. On top of this simulation, a contact tracing scheme was implemented, allowing us to study how e.g. differences in social activity, network structure and degree distributions affects the effectiveness of contact tracing. Our method for doing this largely relies on comparing the performance of contact tracing on the real-world contact network with several artificially homogenized analogues. The study was motivated by the importance of contact tracing in COVID-19 in general, as well as by the introduction of multiple app-based contact tracing solutions [35, 36].

3.1 SUPERSPREADING

Our research into the superspreading phenomenon has focused not on the mechanism behind it, but rather on the consequences. Specifically, we have focused on three main questions:

How does overdispersion ...

- ... affect mitigation strategies?
- ... affect the sensitivity of an epidemic to contact network structure?
- ... affect the evolution of a pathogen?

We have addressed aspects of each of these questions in Refs. [37], [38] and [39], respectively. We also included a section on the impact of overdispersion in the broad COVID-19 review article [40]. We will begin by looking at the first of these questions, namely how overdispersion impacts the effectiveness of lockdowns. This particular non-pharmaceutical intervention has been used very widely during the COVID-19 pandemic. By April of 2020, more than half of the world population were under some kind of lockdown [41] and their use has continued in response to periods of rising case numbers and/or fatalities.

Originally, we were motivated by the observation that lockdowns were remarkably effective for COVID-19 – much more so than typical aggregated models predicted when accounting for the observed reduction in contact rates. Meanwhile, it was becoming increasingly clear that superspreading

was a prominent feature of the pandemic – to a much higher extent than what has been seen in several other respiratory infectious diseases such as pandemic influenza [42, 43, 44]. Now, it is not a priori clear how to include superspreading in a model of disease transmission, as superspreading can have diverse origins, ranging from behavioural to biological [45]. One might assume that biological infectiousness does not vary very much between individuals and that the observed overdispersion is primarily a result of some people being hypersocial. There are, however, a few problems with this assumption: 1) observed low household transmission in COVID-19, 2) overdispersion patterns vary widely between diseases with similar routes of transmission and 3) observed correlations between respiratory viral load variations and overdispersed transmission. We will comment on each of these in more detail below.

Low household transmission. Several studies of household transmission of COVID-19 from around the globe have found that household transmission is quite low in this disease. Two studies from China reported household attack rates of 15% and 12% [46, 47], a smaller South Korean study reported 16% and a nationwide Danish study found an average household attack rate of 17%. Why is this relevant for superspreading? Well, if behaviour and highly variable numbers of contacts are to blame, it is difficult to explain why so many infected individuals do not infect even their closest contacts. If, on the other hand, there is a large biological inter-individual heterogeneity component, so that the majority of individuals simply do not become very infectious, while a few become highly infectious, this pattern of low household transmission is *exactly* what one would expect to see.

Overdispersion varies between comparable diseases. As mentioned above, the level of overdispersion for e.g. pandemic influenza is very different from that of COVID-19. In fact, overdispersion is a widely variable trait between different diseases, even those sharing common modes of transmission [48, 49, 42]. This is difficult to explain from a purely behavioural superspreading perspective.

Respiratory viral load variability and overdispersion are correlated. A recent study showed that just 2% of COVID-positive individuals carry 90% of the virus [49], consistent with a biological mechanism. Another recent study compared the overdispersion levels of influenza A (H1N1), SARS-CoV-1 and SARS-CoV-2 and found that increased variability in respiratory viral load was associated with a higher transmission heterogeneity [50].

In this context, it is worth discussing a fundamental concept from statistical physics, which turns out to be directly relevant for the problem of superspreading – and for our studies of contact tracing as well – namely that of **annealed versus quenched disorder**. Quenched disorder occurs when some variables describing the behaviour of a system are *random* but do not evolve in time. An example could be a material which has been rapidly cooled into a frozen yet disordered state (such as in the glass phase). Annealed disorder, on the other hand, occurs when some variables fluctuate randomly in time, such as the speeds of individual particles in a gas. In a superspreading context, these two types of disorder can come into play in several ways. For instance:

- **Model 1 (annealed disorder):** Each infected individual has a chance of causing a superspreading event at any given time. This can be modeled by a temporally varying contact rate, such that each individual has the same time-averaged contact rate but large temporal fluctuations.
- **Model 2 (quenched disorder):** Some individuals develop high infectiousness while others never do. This can be modeled by a person-specific (fixed) infectiousness, drawn from a statistical distribution once the individual becomes infected.

These are of course far from being the only possible superspreading models, but a simple choice between quenched and annealed disorder makes an enormous difference – for instance, one neatly

explains the household transmission observation while the other does not. In our research, we have mainly assumed a quenched-type model, but other schemes have been explored by collaborators [51].

Chronologically, we first studied how overdispersion affects the effectiveness of lockdowns in Ref. [37]. While the model used in that study is of course very much an idealization, it is more complex than most statistical physics models, having a three-tiered social structure, age structure, and rules for interactions between age groups in each of the social tiers. However, the model suggested that some network theoretical aspects of overdispersed transmission dynamics had previously been overlooked, and we decided that it would be worthwhile to study them in a simpler, dedicated network model. To this end, we developed a simple model of superspreading dynamics in a static network and implemented it in a modular fashion (in C++, as opposed to the PNAS model's FORTAN code). This allowed us to straightforwardly probe the sensitivity of a superspreading epidemic to network structure, connectivity etc. This led to the study published in Physical Review Letters as Ref. [38].

After finishing that study, data emerged which suggested that the novel Alpha (B.1.1.7) variant of SARS-CoV-2 had a lower variability in respiratory viral loads, compared with the ancestral type [52, 53]. This, in turn, suggested that overdispersion could be an unstable trait, and that it may even be the object of evolutionary pressures. Having, by now, a reasonably firm grasp on the network aspects of superspreading, we turned to the question of whether mitigation strategies and overdispersion affected pathogen evolution. This led to the study in Ref. [39] which has yet to undergo peer review at the time of writing.

This remainder of the chapter is organized as follows. Our three main studies within this topic ([37, 38, 39]) are each given a subsection in which we discuss the Methods and Results. Perspectives of these findings are then discussed at the end of the chapter, together with the contact tracing research of section 3.2. The reason for opting for this structure is that the models and results of the underlying papers are sufficiently disjoint that a combined description of all three would only create confusion. However, taken together the results are part of a bigger story which deserves to be discussed together and placed in a coherent perspective. The three subsections which follow are not faithful to the chronology of the projects, but rather begins with the simplest of the models.

3.1.1 SUPERSPREADERS ON NETWORKS

This subsection is based on the study [38]. As with most of our treatments of superspreading, this is mainly based on an agent-based model, the details of which are explained below.

METHODS

The model used in Ref. [38] is the simplest of our superspreading models and can serve as a basis for describing the other models we have developed.

Infectiousness is treated as a stochastic variable representing a quenched disorder in the sense described above. The infectiousness of an individual is drawn from a statistical distribution at the outset of the simulation (or once an individual becomes infected – this is immaterial as long as the superspreading tendency/overdispersion doesn't change over time). A Gamma distribution is commonly used to parametrize infectiousness [48], and this is also the case in our model. This distribution has two parameters, the mean μ and the dispersion parameter k . This dispersion parameter is related to the coefficient of variation (CV) as $CV = 1/\sqrt{k}$, such that a higher k value corresponds to a more homogeneous pattern of infectiousness while a low k value is the signature of a heterogeneously spreading disease. For COVID-19, the k value has been estimated by multiple methods to be around 0.1 [31, 32, 33, 34]. For low values of the dispersion parameter ($k \ll 1$), k itself approximately corresponds to the fraction of infectious individuals responsible for 80% of new infections. For COVID-19, it thus holds that approximately 10% of infected persons lead to 80%

of the new infections. Note that the Gamma distribution becomes a Dirac δ distribution in the limit of $k \rightarrow \infty$, corresponding to completely deterministic infectiousness. In Figure 24b we show the relation between the k value and the fraction of infectious individuals responsible for 80% of new infections. Along with this, we show estimates of k for SARS, MERS [48, 54] and pandemic influenza [42].

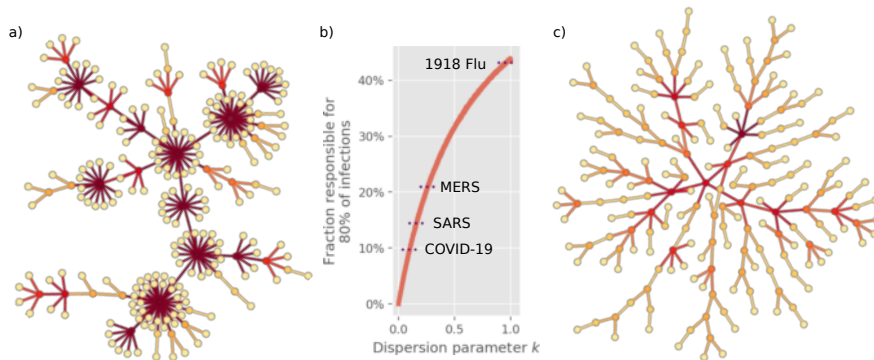


Figure 24: **Characteristics of superspreading infectious networks.** a) Infection network characterized by superspreading with $k = 0.1$. b) The relation between the k value and the fraction of infectious individuals responsible for 80% of new infections. Along with this, we show estimates of k for SARS, MERS [48, 54] and pandemic influenza [42]. c) Infection network characterized by homogeneous transmission with $k \rightarrow \infty$, i.e. a Dirac δ distributed infectiousness. Figure from Ref. [38].

The basic disease progression model that we employ is a compartmental SE(P)IR type model, i.e. **S**usceptible-**E**xposed-(**P**resymptomatic-)**I**nfectious-**R**ecovered. The presymptomatic stage is only important in models of contact tracing (which we will cover later) – otherwise it is treated identically to the Infected state. The individual states along with their average durations are shown in Figure 25. Each stage is modeled as having a constant rate for leaving, leading to exponentially distributed occupation times. The durations indicated in the figure are thus the reciprocals of the rates for leaving each state.

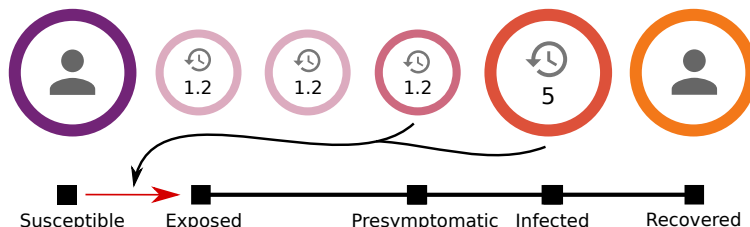


Figure 25: The basic disease progression models used in our study. Figure from Ref. [38], Supplemental Material.

The exposed state is divided into two exponentially distributed 1.2-day stages in order to produce a Gamma-distributed state with $k = 2$. This splitting is done to prevent an unrealistically short exposed period, which would occur with a high probability if the state had been modeled as a single constant-rate process, as is the case in traditional S(E)IR models. The reason for this is of course that an exponential distribution ($p(\tau) \sim e^{-c\tau}, \tau \geq 0$) attains its maximal value at the origin, i.e. for $\tau = 0$, corresponding to skipping the exposed period entirely. The infectious state is also split into two stages, the Presymptomatic and the Infected stage. This has a dual purpose of, again, preventing unrealistically short infectious periods, and furthermore the inclusion of a presymptomatic stage makes extension to contact tracing situations more straightforward (more on this in the next section).

The contact dynamics of our base model is based on static random networks. We will describe the nature of these random networks in more detail below. Once a network has been established, the contact dynamics consists of each agent choosing a number of its contacts to interact with in

each time step. In the base model, with uniformly distributed social activity, this number has been normalized to 1. If an infected and a susceptible person come into contact, the disease is transmitted with a probability rate given by the personal infectiousness parameter of the infected individual.

The basic network simulation scheme can be summarized as follows:

- Initialization:
 - Generate the contact network (e.g. Erdős–Rényi, clustered network or scale-free)
 - For each agent i , draw a personal infectiousness parameter r_i from a Gamma distribution with dispersion parameter k .
 - Initiate all agents in the **Susceptible** state (except for a few, see next step).
 - Designate a certain fraction (e.g. 0.1%) as initially **Infected**.
- Time evolution (re-iterate for each timestep):
 - For each agent in the network, activate one associated edge, simulating a contact event.
 - * (Note: This is in the simplest case of uniform social activity levels.)
 - When a network edge between an infectious individual i and a susceptible individual j is activated, the susceptible individual becomes exposed with a probability rate r_i .
 - * (Note: This assumes uniform susceptibility. Stochastic susceptibility is a straightforward extension.)
 - Exposed and infectious individuals leave their current stage at the probability rate given by the inverse of the mean duration of the stage given in Figure 25.
 - Increment time t by $\Delta t = 30\text{min}$.

Analytic description. Although a full agent-based model, as described above, is necessary for some of our simulations, some particularly simple results can be obtained directly from an analytic description – as long as global saturation effects can be ignored.

We consider an infected person with c contacts (the *degree* of the *node*, in network parlance). All contacts are assumed susceptible. The infectiousness of this individual is drawn from a distribution $P_I(r)$, which we assume to be a Gamma distribution with dispersion parameter k and mean μ . For simplicity, we notationally suppress these parameters when referring to the distribution $P_I(r)$. The distribution of the reproductive number R of an individual with a *known* infectious parameter r and degree (number of contacts) c is then:

$$P(R; r, c) = \binom{c}{R} \left(1 - e^{-r/c}\right)^R \left(e^{-r/c}\right)^{c-R} \quad (3.1)$$

In the limit of $c \rightarrow \infty$ this becomes a Poisson distribution with mean r , as would be expected when there are no limitations on available susceptible contacts.

For an individual with an a priori unknown infectiousness, the number r must be drawn from the Gamma distribution P_I . The distribution of the reproductive number thus becomes

$$P(R; c) = \int_{r=0}^{r=\infty} dr P_I(r) P(R; r, c) \quad (3.2)$$

In the aforementioned limit of infinite connectivity, $c \rightarrow \infty$, this becomes a negative binomial distribution, which is thus the distribution of personal reproductive number of an overdispersed disease in a homogeneous-mixing contact structure [48]. If the limit $k \rightarrow \infty$ is taken as well, the distribution simply becomes Poissonian.

The above computations were for an individual with a known number of contacts c , but if a degree distribution $P_C(c)$ is given, the expected reproductive number for an individual with an a priori

unknown number of contacts can be computed as $\sum_c P_C(c)P(R; c)$ (since P_C is necessarily discrete). As can be gathered from these equations, the actual reproductive number does not only depend on infectiousness, but on connectivity and contact structure as well.

RESULTS

In panels a and b of Figure 26, we compare how the distribution of reproductive numbers, as well as the mean reproductive number is affected by reductions in network size. The conclusion is that even moderate reductions in network size have a drastic effect in a disease characterized by superspreading (panel b) while a homogeneously spreading disease is much less affected (panel a). In panel c we show the basic reproductive number as a function of the average connectivity and dispersion factor, thus giving an overview of the connection between superspreading and contact network size. These computations are all based on the analytic description described towards the end of the Methods section above. Details can be found in Ref. [38] and Supplemental Material.

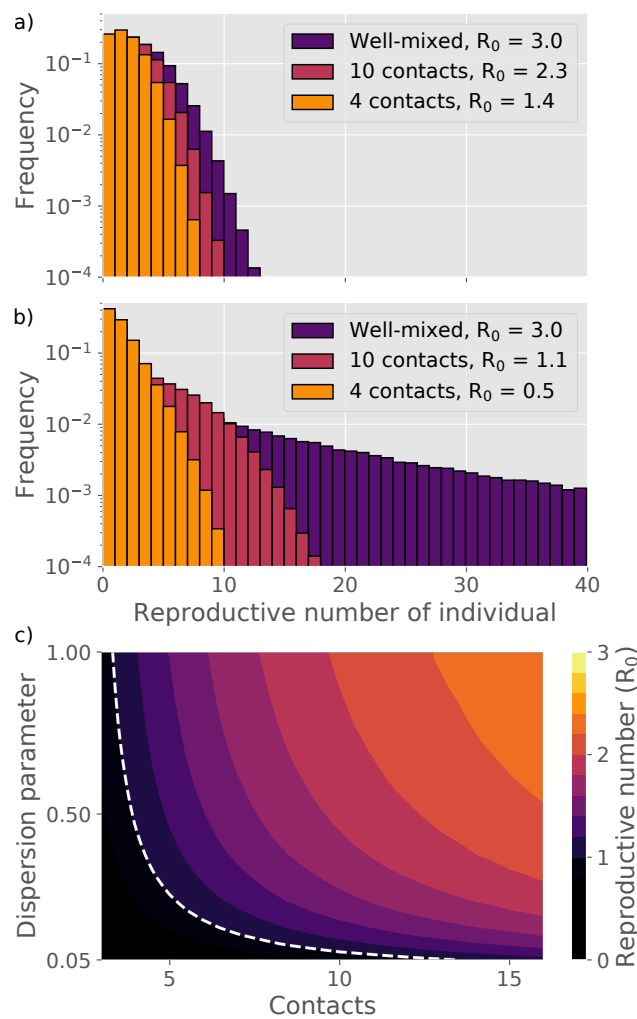


Figure 26: **The effect of network reduction depends strongly on overdispersion.** **a)** Distribution of personal (basic) reproductive number in a homogeneously spreading disease ($k \rightarrow \infty$) for three different connectivities. **b)** Distribution of personal (basic) reproductive number in a superspreading disease ($k = 0.1$, similar to COVID-19) for three different connectivities. **c)** Basic reproductive number as a function of connectivity (average number of contacts) and overdispersion (the k value – higher k corresponding to more homogeneous spreading). Figure from [38].

While these computations are tractable in the analytic description, it does not suffice for simulating an entire epidemic trajectory or computing e.g. attack rates. The reason for this is that it fundamentally neglects global saturation (as well as some aspects of local saturation). For the full simulations, we turn to the agent based model described in the Methods section. The disease in question has been parametrized to have an initial growth rate of 23% per day in a homogeneous

mixing contact structure, giving an R_0 of 3, both figures representative of the early COVID-19 pandemic [55, 56, 38, 57]. In Figure 27, the trajectory of a homogeneously spreading (panel a) and a superspreading disease ($k = 0.1$, panel b) are shown in each of the three connectivity cases ($\langle c \rangle = 10$, $\langle c \rangle = 15$ and well-mixed, which formally corresponds to $c \rightarrow \infty$).

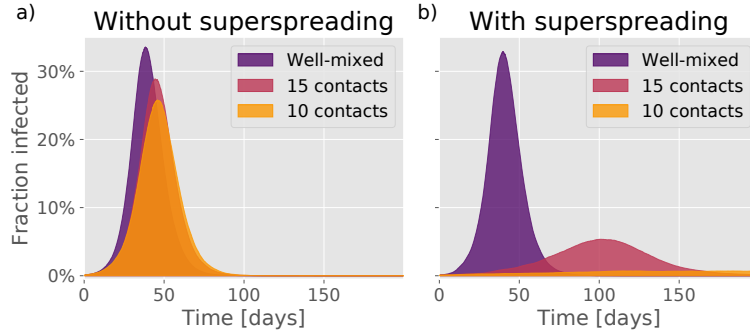


Figure 27: The spread of an infectious disease in contact networks with different mean connectivities. **a)** A disease with a homogeneous infectiousness distribution ($k \rightarrow \infty$). **b)** An overdispersed disease prone to superspreading ($k = 0.1$). Figure from [38].

This shows very clearly how sensitive a superspreading (i.e. highly overdispersed) epidemic is to contact network size – something which is perhaps even clearer here than when just comparing reproductive numbers. These simulations are run on Erdős-Renyi networks, which are characterized by being essentially devoid of clusters. However, by introducing a network with a high degree of clustering, we can assess the sensitivity of a superspreading disease to this aspect of network structure. The exact algorithm for generating this type of network can be found in the Supplemental Material of Ref. [38], but the concept is as follows. Each person is a member of two ”groups”, each of which is fully internally connected. As an example, consider a network with mean connectivity 10. A representative person will then be a member of two groups of size five⁵. This person will typically be the only link between those two groups, but everyone who is a member of either group will also have a connection with every other member of the same group.

In Figure 28, we show the total attack rate (also known as the final epidemic size) as a function of connectivity and dispersion parameter k , similarly to how we plotted the reproductive number in Figure 26. The new aspect here is the clustered network in panel b. Clearly, clustering has a large effect on the attack rate, and even more so for a very heterogeneously spreading disease.

The last result of this section is an analytical one. We present an analytic derivation of an equation for the reproductive number as a function of connectivity and dispersion which qualitatively reproduces the plot of Figure 26c.

The derivation starts from equation (3.1). The basic reproductive number, given an infectiousness (Gamma) distribution P_I with dispersion parameter k and average infectiousness μ , can then be written as:

$$\begin{aligned}
 R_0(c, k, \mu) &= \sum_{R=0}^{\infty} \int_{r=0}^{r=\infty} dr R P_I(r, k, \mu) P(R; r, c) \\
 &= c - c \left(\frac{1}{1 + \mu/(ck)} \right)^k
 \end{aligned} \tag{3.3}$$

In the limit $c \rightarrow \infty$, this becomes $R_0 = \mu$, a reflection of the fact that the average infectiousness μ precisely equals the basic reproductive number in the homogeneous mixing limit.

Recall that all c contacts were assumed to be susceptible. More realistically, we can take into account that one contact in the personal network of each infected person will in general be insusceptible, even where computations of the *basic* reproductive number R_0 are concerned. This insusceptible individual is the one who transmitted the disease to the infected person we are considering. This is

⁵Note that the group size is stochastic, but we’re just considering a representative node in the network for simplicity.

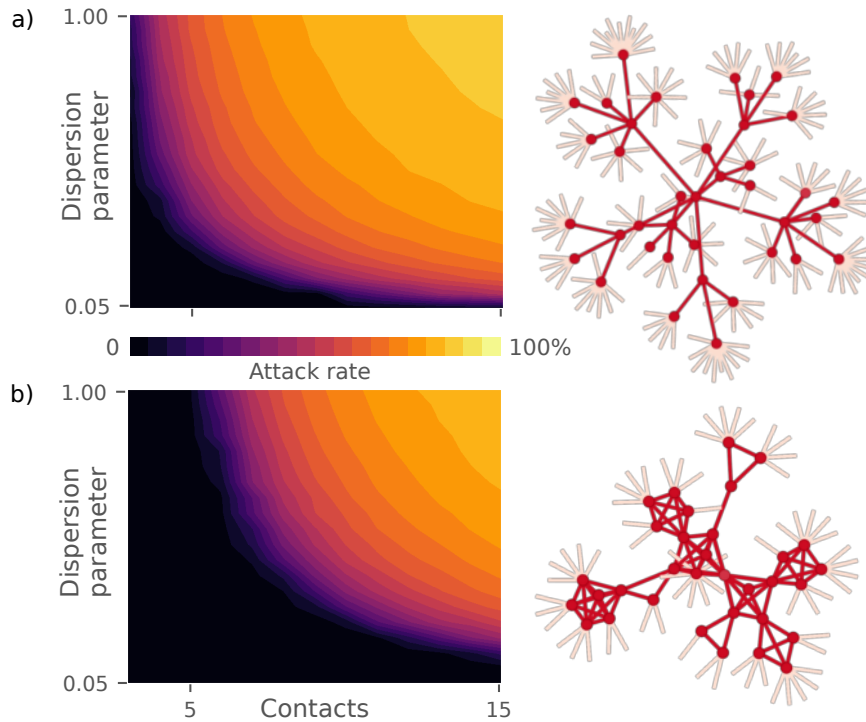


Figure 28: The dependence of the finale epidemic size (attack rate) on network connectivity, superspreading tendency and network clustering. **a)** Attack rates in an Erdős-Renyi network which is free from clustering. **b)** Attack rates in a clustered network, where the contact network of each person is divided into two internally connected groups. Figure from [38].

however easy to implement, by considering instead the function $R_0[c - 1, k, (1 - \frac{1}{c+1})\mu]$, which we have plotted in Figure 29. Note that it closely resembles Figure 26c.

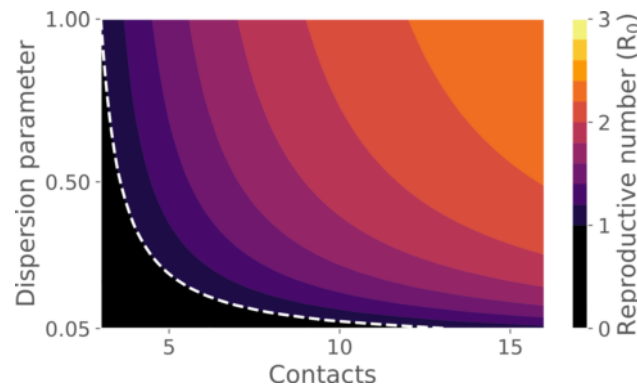


Figure 29: The basic reproductive number as a function of connectivity and dispersion factor. This plot is obtained directly from Equation (3.1) and reproduces Figure 26 with two minor differences: 1) a fixed (δ distributed) connectivity c is assumed, while in Figure 26 it followed a Poisson distribution, and 2) here, all integrals could be performed analytically, while the results in Figure 26 were evaluated numerically, out of necessity.

3.1.2 SUPERSPREADING, LOCKDOWNS AND NON-REPEATED CONTACTS

This section is based on the study published in PNAS as Ref. [37]. Here, we model the impact of superspreading on the effectiveness of lockdowns as non-pharmaceutical interventions. This turns out to reveal a plausible explanation for the relatively large observed effect of lockdowns in mitigating the COVID-19 pandemic.

METHODS

The model employed here bears many similarities to the simple network model described in the previous section, but is substantially more complex. Like the previous model, this is an agent-based

model. However, this study aims to capture some more subtle aspects of mitigation strategies and includes a three-tiered contact structure as well as age-dependent social activity levels.

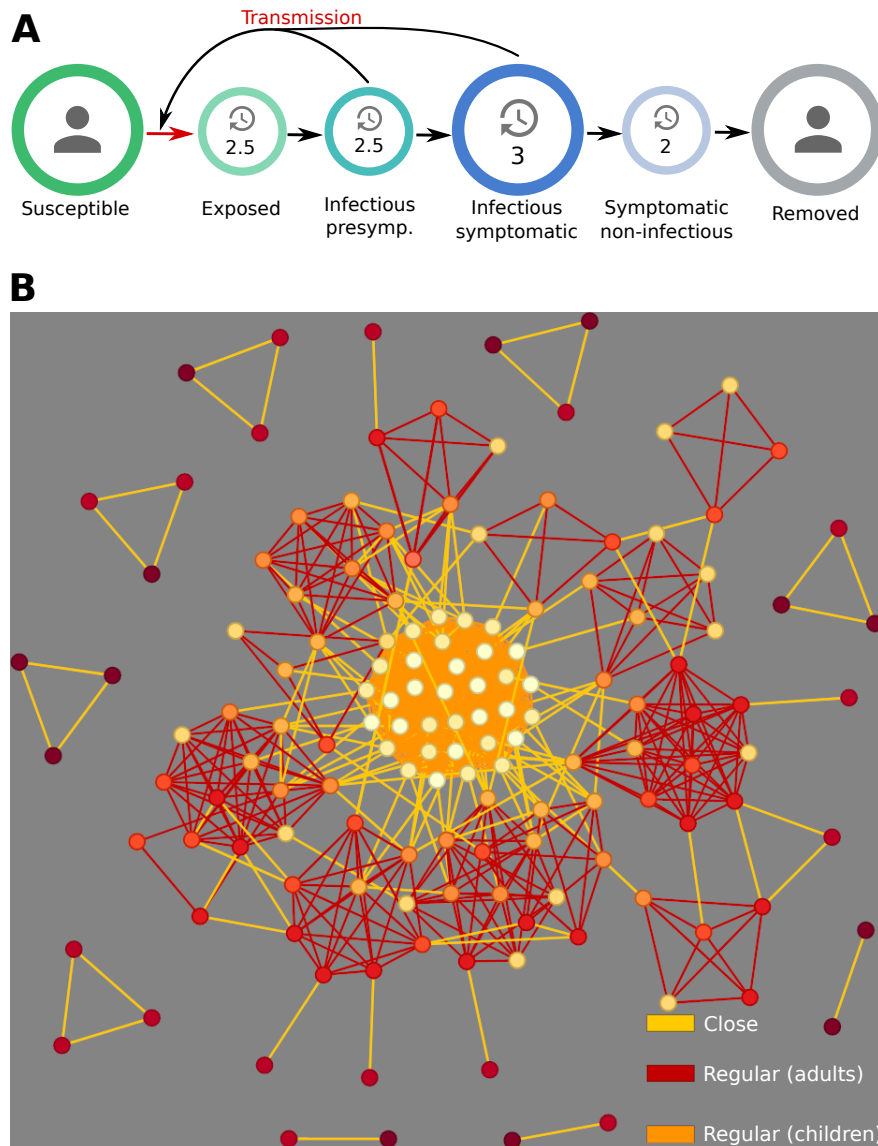


Figure 30: **A)** The disease progression model employed in Ref. [37]. **B)** A scaled-down version of the social networks employed in Ref. [37]. The link colours denote the sectors to which the contact relations belong, as indicated in the legend. The node colours indicate the age of the individual in question, with darker reds indicating older individuals and pale yellow indicating a young individual. Figure from Ref. [37].

In these simulations, each individual is assigned to one *close* and one *regular* unit. These units are part of a static network, so contacts are repeated over the course of the epidemic. Aside from these fixed units, each individual also participates in random interactions with contacts drawn from the entire population. These latter contacts are non-repeated, i.e. devoid of temporal correlation. The population is structured by age into 10-year intervals, each with a corresponding social activity level a_i . The age structure of the population and the associated relative contact rates are given in Table 1. The activity levels are fitted such that the observed contact rates in an unmitigated simulation scenario match those given in the table. The age structure is based on Ref. [58] while the activity levels are based on Ref. [59].

Each *close* unit is reminiscent of a household. It has an average size of 2.3 and a coefficient of (size) variation of 0.59. These figures closely match those reported in The European Union Statistics on Income and Living Conditions Survey, which gives an average size of 2.3 with $CV = 0.57$ [60]. In *close* units, adults are in the same or adjacent age bands, while children are taken to be 20 to 40 years younger than adults of the same unit.

Table 1: Population age structure and age-structured social activity data used in [37]. The age structure data is from [58] while age-structured activity data is from [59].

Age (y)	Share of population (%)	Relative social time per person
0-9	10.9	1.21
10-19	11.9	1.70
20-29	13.3	1.45
30-39	11.7	1.45
40-49	13.6	1.38
50-59	13.6	1.31
60-69	11.7	1.06
70-79	8.9	0.81
80+	4.3	0.81

The *regular* units have more structure. For agents 20-70 years of age, the regular unit is Poisson-distributed with an average size of 8 and approximately represents a workplace. Individuals under 20 years of age are assigned to *regular* units of 18 members, which are also assigned two adults aged 20 to 70 each. Agents above the age of 70 are not assigned to a *regular* unit.

The disease progression model is an agent-based SEIR framework similar to the one described in section 3.1.1. See Ref. [37] for the exact details. As before, infectiousness is taken to be Gamma distributed and is treated as a quenched variable.

The time evolution scheme is also similar to that described in section 3.1.1, with the addition that each individual now has an activity-dependent (and thus age-dependent) probability of making a contact in each time step. These possible contacts are chosen from the three social sectors, and the relative frequency of each of these sectors is based on a population-based survey of mixing patterns in eight European countries by Mossong et al. [59]. They found that the “home” sector made up 19 to 50% of all contacts, while the “work/school” sector accounted for 23 to 37%, and the remaining sectors amounted to 27 to 44%. These intervals are consistent with equipartitioning, so we chose to simplify by letting one third of social time fall into each of the three sectors of our model. This holds for the unmitigated base scenario – mitigation strategies will generally affect this partitioning. The purpose of this additional social structure is mainly to simulate mitigation strategies which rely on reducing contact in either the *regular* or the *random* sectors.

RESULTS

Our main result concerns the effect of limiting contacts in the *regular* and *random* sectors in an overdispersed or homogeneously spreading disease. We do not limit contacts in the *close* sector since this does not correspond to a credible mitigation strategy. In Figure 31, we show the effects of each of these two mitigation strategies. In panel A, the disease in question is assumed to have a δ -distributed infectiousness profile – i.e. every infected person develops the same level of infectiousness – corresponding to the limit $k \rightarrow \infty$. Here, the conclusion is that restricting contacts in either of the two sectors has a similar effect. Contrast this with panel B. Here a dispersion factor of $k = 0.1$ is assumed, and the effect of limiting *random* contacts is drastically enhanced despite the biological mean infectiousness of the disease being unchanged. The interpretation of this is quite clear: For a superspreading disease, the number of *different* contacts an individual comes into contact with is more important than the total contact *time*. For a homogeneously spreading disease, the opposite is true: contact time is the important variable.

In Figure 32, we systematically vary the dispersion parameter and show the trajectory of the

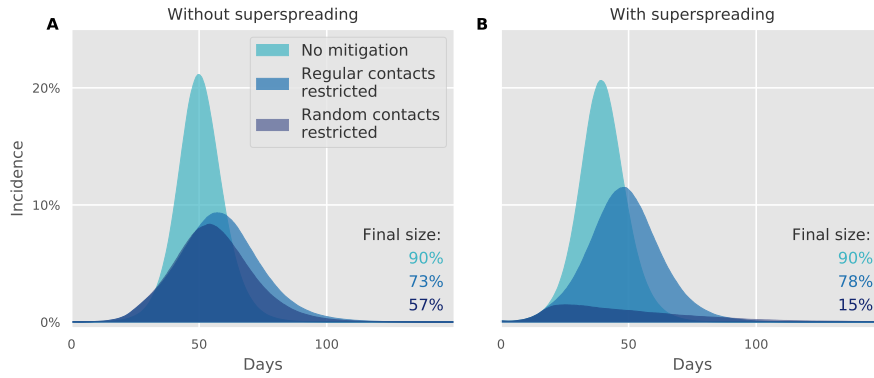


Figure 31: The effect of mitigation strategies based on restricting either *regular* or *random* contacts. **A)** A homogeneously spreading disease, meaning that infectiousness is δ -distributed, corresponding to taking the limit $k \rightarrow \infty$. **B)** A superspreading disease with an overdispersion of $k = 0.1$, similar to COVID-19.

epidemic under the lockdown-like random sector mitigation for each value of k . The dependence on the dispersion parameter is quite dramatic.

Further details and sensitivity analyses can be found in Ref. [37]. We analyzed the sensitivity of these results to variations in initial growth rate, relative sizes of compartments (in terms of social time in all three as well as the number of distinct contacts in *close* and *regular*), inter-individual heterogeneity in number of random contacts and total social time. The overarching finding was that the result is robust to these perturbations.

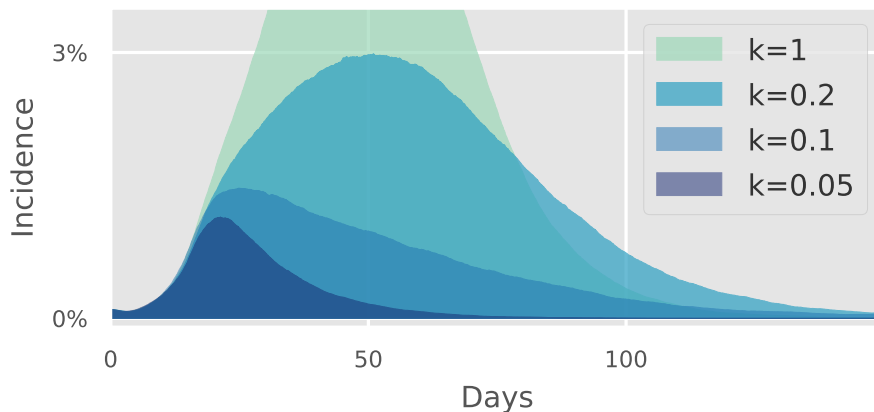


Figure 32: The trajectory of an epidemic under a mitigation strategy based on the reduction of random sector contacts, at different levels of transmission heterogeneity. Lower values of k corresponds to more heterogeneous transmission. The overdispersion level for COVID-19 is likely around $k = 0.1$. The effectiveness of this type of mitigation strategy is seen to depend strongly on overdispersion.

3.1.3 OVERDISPERSION AS AN EVOLVING CHARACTERISTIC

In the beginning of 2021, preliminary data on the distributions of respiratory viral loads (or Ct values) in persons infected with the then-emerging Alpha-variant (B.1.1.7) of SARS-CoV-2 became available [52, 53]. This variant has been reported to be $\sim 50\%$ more transmissible than the ancestral SARS-CoV-2 virus under varying degrees of lockdown [61, 62, 63]. In the aforementioned viral load studies themselves, focus was mostly on changes in the mean or median viral load, since these are most obviously correlated with infectiousness. However, we noticed that it was not only the average viral load that had changed, but the variability as well. Using data from Ref. [52], we calculated that the viral loads in samples of the Alpha variant were associated with a lower coefficient of variation (≈ 2), compared to the ancestral strain (≈ 4). In Ref. [53], the authors presented viral load distributions for samples on the basis of the presence or absence of some of the defining mutations of the Alpha variant, namely the N501Y substitution and the Δ H69/V70 deletion. Again, these distributions suggested a reduced variability in the Alpha variant.

In [37] we had shown that the highly overdispersed variants were at a significant disadvantage when it came to spreading under (partial) lockdowns - i.e. mitigation strategies which primarily rely on reducing the number of contacts. Together, these findings led us to speculate that a) overdispersion could be an evolving characteristic and b) non-pharmaceutical interventions may exert an evolutionary pressure in this direction.

We emphasize that the phrase “evolutionary advantage” has more than one meaning. In terms of viral strains, it is natural to distinguish between the “survival ability” and the “competitive advantage” of a variant. Our goal is to explore these two aspects separately. The first aspect we explore by performing a stochastic extinction analysis under a lockdown-like scenario as well as in an unmitigated scenario. This involves simulating the evolution of an outbreak starting from a single case of a given variant and recording the probability that the outbreak is sustained beyond a few generations or goes extinct. This can be done in a fairly simple branching process type model as well as using generating function methods – see the Methods section below and the Supporting Information of [39] for details.

Studying the “competitive advantage” aspect, on the other hand, requires a full-fledged epidemic simulation. We created an agent-based model which allowed for coexistence of multiple variants with different dispersion and (mean) infectiousness as well as evolution and inheritance of those characteristics.

METHODS

The main model of this study is an extension of the agent-based model described in section 3.1.1. The disease progression and social network aspects are the same. We have however implemented a few extensions, mainly to allow us to study the evolution of overdispersion (modeled as mutations affecting the k parameter of the infectiousness distribution).

Stochastic extinction model We model stochastic extinction using a branching process algorithm based on sampling of probability distributions with an analytic description. In practice, we have performed the computation by numerical sampling. Initially, a single infected individual is introduced. This individual is infected with a disease characterized by a certain mean infectiousness and dispersion parameter.

We present the algorithm below, which is reiterated for each new generation of the outbreak. Without loss of generality, we therefore here describe just a single generation which initially consists of I infected individuals. Note that for the initial generation, $I = 1$.

- For $i \in \{1, \dots, I\}$:
 - Draw individual infectiousness r_i from Gamma distribution $P_r(r; k, \mu)$
 - Draw number of contacts c from a Poisson distribution with a given mean connectivity.
 - Given number of contacts c , draw personal reproductive number R_i from the distribution (3.4)

$$P_R(R; r, c) = \binom{c}{R} \left(1 - e^{-r/c}\right)^R \left(e^{-r/c}\right)^{(c-R)}. \quad (3.4)$$

- Let the number of newly infected be $I = \sum_i R_i$ and repeat the algorithm with this updated value of I .

If the number of infected I drops to zero at any point, the outbreak is said to have gone extinct in that generation. By performing multiple such branching process simulations for each value of the parameters μ (mean infectiousness) and k (dispersion factor) we build up a statistic of the survival chance of each specific variant.

Agent-based evolution model The basic model, including the disease progression and contact structure, is similar to that described in section 3.1.1. Here we will describe only the extensions which we have introduced.

Two-strain simulations: In the two-strain simulations, we study only the fitness of the two variants, and mutations affecting overdispersion do not occur during the simulation. Initially we designate some fraction x_1 of the infected individuals as having been infected with the heterogeneous ancestral variant ($k = 0.1$) and the remaining fraction $x_2 = 1 - x_1$ as having been infected with a more homogeneous emerging variant ($k = 0.2$). When an individual infected with a given variant infects a susceptible person, the characteristics of the disease are passed on to the newly infected individual. The infectiousness of this individual is then drawn from a Gamma distribution with dispersion parameter k determined by the variant in question.

Evolution simulations: In these simulations, mutations affecting overdispersion are introduced. They occur randomly and without any intrinsic bias, allowing us to study any external evolutionary pressures exerted by e.g. mitigations. We allow the pathogen to stochastically mutate its overdispersion upon transmission with a certain probability p . When such a mutation occurs, the dispersion parameter is either increased ($k \rightarrow rk$, $r > 1$) or decreased ($k \rightarrow k/r$) with equal probability. On the microscopic level, the dispersion parameter thus performs an unbiased geometric Brownian motion. In our simulations, we have chosen the mutation probability $p = 1/3$ and magnitude $r = 3/2$. The figure 1/3 was chosen since the SARS-CoV-2 pathogen has been estimated to mutate at a rate of approximately 2 substitutions per genome per month [64], translating to about one mutation per three transmissions. However, these simulations should be seen as entirely conceptual, since the frequency of mutations affecting dispersion are likely to be much lower, and the expected magnitude of such mutations is currently not known.

RESULTS

The first aspect of evolutionary advantage that we study is the ability of a pathogen to avoid stochastic extinction in the initial stage of a potential outbreak. In Figure 33, we plot the survival chance after 10 generations as a function of dispersion and mean infectiousness. In panel A, we consider an unmitigated scenario, modeled by a homogeneous mixing contact structure (formally infinite connectivity). The central finding here is that the initial survival chance depends very strongly on overdispersion. In panel B, contact restrictions are in place, limiting each person to 10 distinct contacts in order to simulate a partial lockdown scenario. This allows us to probe how strongly the survival chances of variants with different degrees of overdispersion are affected by this kind of mitigation strategy, and we see a moderate dependence. Superspreading variants are clearly hit harder by these restrictions, in terms of extinction risk.

In Figure 34, we take a different perspective and look at the relative ability of two differently-dispersed variants to spread *after* gaining a foothold, i.e. after having moved past the initial stage where stochastic extinction risk is high.

We do this by initially letting 1% of the population be infected with the ancestral variant ($k = 0.1$) and just 0.01% with an emerging, more homogeneous variant ($k = 0.2$). Initially, connectivity is restricted to 10 people, but once the emerging variant reaches a share of 20% of current infections – around day 65 – restrictions are lifted. This is modeled by transitioning to a homogeneous mixing contact structure (connectivity $\rightarrow \infty$). In panel A, the epidemic trajectories of each of the two strains are shown on a linear scale. The original strain is approximately stationary under the restricted social conditions, while both variants predictably surge when restrictions are lifted. The initial part of the trajectory is more clearly seen on a logarithmic scale, as presented in panel B. Here, one can clearly see that the emerging variant grows essentially exponentially even under partial lockdown conditions while the old variant doesn't grow appreciably. After the reopening, the two variants are seen to spread equally well. This latter point is more clearly seen in panel C, where the relative incidence of the two variants are shown - i.e. each variant's share of current infections. The emerging variant's share grows fast in the initial stage, but after society reopens, the new variant completely loses its competitive advantage and the relative incidence stabilizes at close

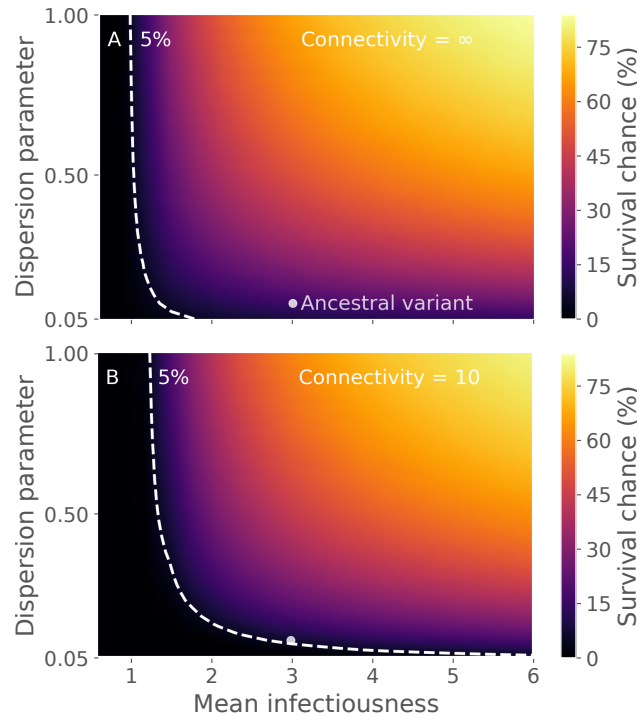


Figure 33: Survival chance of variants with different biological mean infectiousness and dispersion parameter. **Top:** Unmitigated scenario, modeled by a homogeneous mixing contact structure. **Bottom:** Partial lockdown scenario, modeled by limiting contacts to 10.

to 20%. This underscores that the fitness advantage of new variants can be enormously context dependent, and not necessarily reflective of a higher biological mean infectiousness. Consequently, this finding has implications for attempts at extrapolating the basic reproductive number of an emerging variant on the basis of measurements obtained under mitigation.

In order to systematically quantify the relative fitness of variants with different levels of overdispersion and mean infectiousness, we ran epidemics for variants with $k \in [0.05, 1.0]$ and mean infectiousness ranging from $1/3$ to 2 times that of the ancestral variant. In Figure 35A, the results for a connectivity of 50 are shown. Clearly, there is a strong relation between mean infectiousness (the horizontal axis) and observed reproductive number. However, it is also clear that the dispersion plays a substantial role. In panel B, where connectivity is restricted to 10, this effect becomes much stronger, as evidenced by the bending of the contour lines. The dashed white line in each plot indicates pathogens which spread as effectively as the ancestral variant. Remarkably, panel B shows that it is possible for a pathogen that has only *half* the mean infectiousness of the ancestral variant to spread just as efficiently, provided it is more homogeneous ($k \approx 1$).

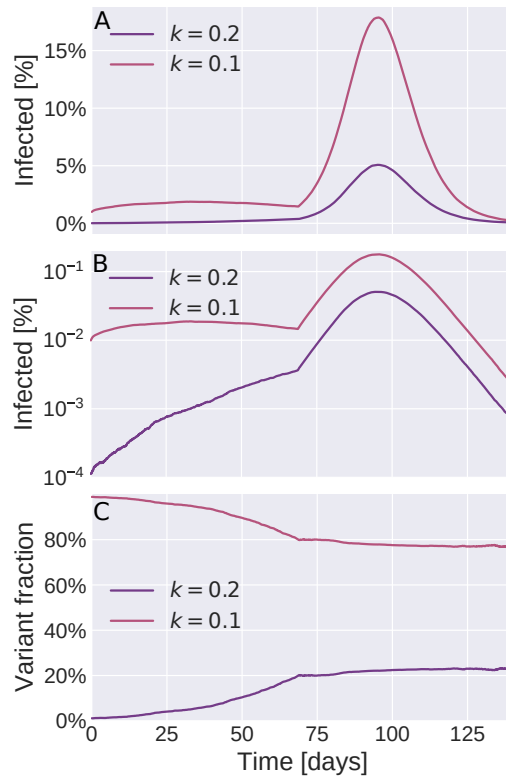


Figure 34: Competition of two variants with different levels of overdispersion. Initially, the heterogeneous ancestral variant ($k = 0.1$) makes up 99% of infections while an emerging, more homogeneously spreading variant ($k = 0.2$) makes up just 1% of infections. Initially, a partial lockdown is in place (modeled by an Erdos-Renyi network with 10 contacts/person). When the emerging variant makes up 20% of current infections (at $t \approx 65$ days), society is reopened (modeled by transitioning to a homogeneous mixing contact structure). **A)** Incidence of the two variants over time, linear scale. **B)** Incidence of the two variants over time, logarithmic scale. **C)** Relative incidence of the two variants. After reopening, the fractions stabilize, indicating that the two variants spread equally well under these conditions.

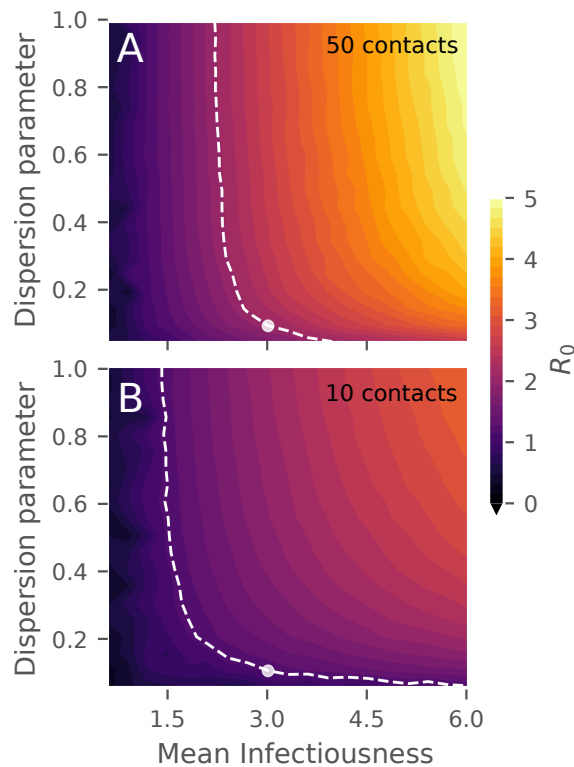


Figure 35: Fitness of variants measured by basic reproductive number R_0 . Variants have different mean infectiousness and level of dispersion k . **A)** Epidemic spread in an Erdős-Renyi network with a mean connectivity of 50. **B)** Epidemic spread in an Erdős-Renyi network with a mean connectivity of 10.

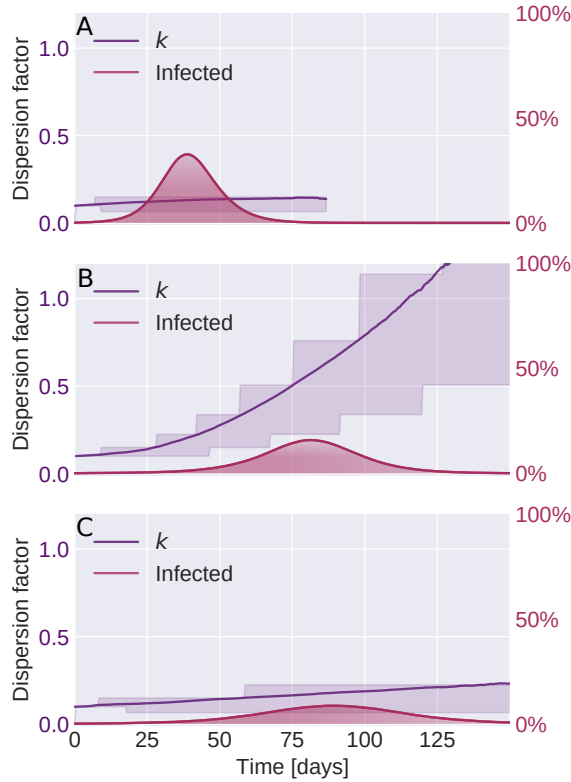


Figure 36: Evolution of the dispersion parameter k under different conditions. In each panel, the red curve shows the combined incidence of all strains while the purple curve shows the average value of k among currently infected individuals. The shaded purple area shows the 25% and 75% percentiles of the k distribution as a function of time. **A)** Unmitigated scenario. During the epidemic, a weak tendency towards increasing homogeneity (increasing k) is seen. **B)** Partial lockdown (connectivity = 15) scenario. The k value increases dramatically, indicating that homogeneous spreading is being strongly selected for. **C)** No social network restrictions (i.e. no lockdown), but transmissibility lowered by other means (corresponding to e.g. mask wearing) until initial growth rate matches that of panel A.)

We have seen that there are apparently large fitness advantages associated with lower levels of dispersion, and that this is even more true in (partial) lockdown-like scenarios. We have also seen preliminary evidence that overdispersion is not a fixed property and that the Alpha variant may show reduced overdispersion [53, 52]. It is thus natural to ask: if dispersion is allowed to mutate randomly, will it be driven in any particular direction? Using the evolution model described in the Methods section, we simulate the spread of a pathogen which initially has the properties of the ancestral variant ($k = 0.1$ and $R_0 = 3$ in an unmitigated scenario). In an unmitigated scenario (Figure 36A) this leads to a weak tendency towards increased homogeneity, with a k value that only grows slightly over the course of the epidemic. In Figure 36B, we run a similar scenario albeit with average connectivity restricted to 10. Suddenly, a dramatic increase in the average k value over time is seen. Since the mutations are completely random and unbiased at the microscopic scale, any observed drift is caused by a fitness advantage. Of course, one could object that the scenarios in panels A and B are not directly comparable since the epidemic in panel B is much less rapid, leaving the pathogen with many more generations in which to mutate. For this reason, we included a scenario where the infectiousness of the pathogen has been reduced such that the initial growth rate is identical to that of panel B, but without the connectivity restrictions. Here we see a slightly greater increase in k compared to panel A, but nowhere near as dramatic as in panel B. It is thus clear that it is the lockdown which exerts an evolutionary pressure on the overdispersion, driving the pathogen towards more homogeneous patterns of spreading.

3.2 CONTACT TRACING IN HETEROGENEOUS SOCIAL NETWORKS

In this section we use a real-life, temporally resolved contact network to assess the impact of network and contact heterogeneities on contact tracing as a mitigation strategy in a COVID-19-like disease.

The contact tracing algorithm used is a model of digital (app-based) contact tracing, which the underlying data set is well-suited to study, since it consists of proximity data similar to what many contact tracing apps rely on. This section is based on the (preprint) study Ref. [65].

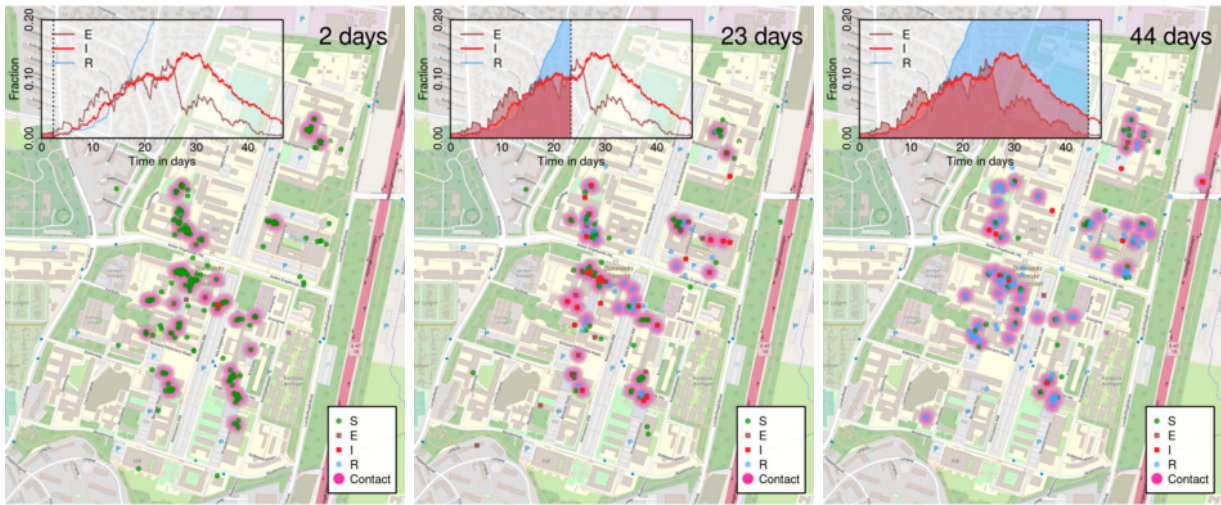


Figure 37: A single simulation of an outbreak on an empirical contact network. We show a zoom view of a small geographical part of the system, based on recorded GPS coordinates. Contacts (as defined in the Methods section) are represented by pink, the corresponding epidemic trajectory is shown in the inserts, at three different times during the outbreak.

3.2.1 METHODS

The dataset we work with consists of temporally resolved proximity data collected using smartphones distributed to a group of 1000 students at the Technical University of Denmark as part of the Copenhagen Networks Study [66, 67]. These smartphones had been pre-installed with an application which collected communication, location and proximity data in the form of call and messaging logs, geo-location by means of GPS as well social proximity using Bluetooth. Every five minutes, Bluetooth ports on all participating devices would open and scan for nearby devices. Bluetooth signal strength as well as GPS location would then be recorded. This signal strength can then in turn be used as an approximate proxy for distance. The data we used were collected in the period 2013-2015, i.e. pre-pandemic. As such they reflect usual day-to-day contact patterns in the absence of any disease awareness or non-pharmaceutical interventions.

In order to transform these Bluetooth data into a temporal contact network, we must define a notion of *contact* that is epidemiologically relevant. Such a definition depends on the disease in question. If fomite or environmental transmission is significant, a simple measure based on proximity time and distance is insufficient. For COVID-19, there is evidence that fomite transmission is only a minor contributing factor, and we have used a cut-off in (RSSI) signal strength of -85dBm which captures almost all $\leq 1\text{m}$ interactions while excluding most $\geq 3\text{m}$ interactions. This allows us to define a time-dependent social network with a time step size of 5 minutes.

The epidemic model consists of an agent-based SEPIR (Susceptible-Exposed-Presymptomatic-Infected-Recovered) model run on top of the time-dependent social network. The disease progression model itself is very similar to the one described in section 3.1.1. For more details, see Ref. [65]. As opposed to the previous agent-based modelling studies described in this thesis, we do not impose a specific contact dynamics here, since it is predetermined by the empirical contact network. When two persons are in contact, there is a fixed probability per unit of time for an infection to be transmitted, provided that one is infectious and the other susceptible. In addition to the five states (S, E, P, I and R), an individual may also be flagged as Quarantined (Q). This quarantine is assumed to be perfect, and while it is active the individual cannot infect anyone or be infected. The significance of the presymptomatic (P) state will become clear once the contact tracing algorithm

is described, but in the absence of contact tracing, an agent in the P state is treated similarly to one in the infected (I) state.

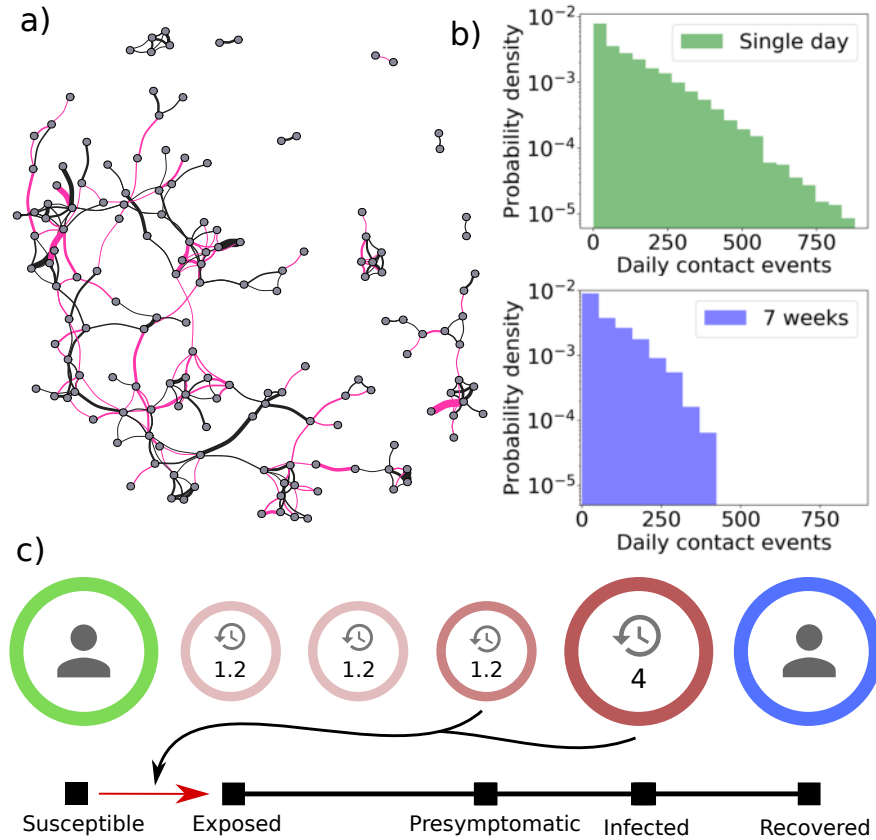


Figure 38: **a)** A small fragment of a cumulative one-week contact network. The thickness of each link indicates total time active. Only links with more than two hours of activity are included. Black links are recurring relative to the previous week while red links indicate new connections. **b)** Top: Histogram of contact events over a one-day period (semi-logarithmic). The mean is 131 and the coefficient of variation (CV) is 1.03. Bottom: Histogram of contact events over 7 weeks, rescaled to a daily rate (semi-logarithmic). The mean is 86 and the CV is 0.95. **c)** Disease progression model for the agents in the system.

The contact tracing (TTI) model. Our contact tracing model is based on a test-trace-isolate (TTI) scheme. It thus contains three elements: a regular testing scheme, a mechanism for tracing contacts of discovered infected persons and a set of rules for isolation of suspected (or confirmed) cases. Regular (“background”) testing happens at a constant rate r_{test} . Once a positive individual is found by regular testing, the contact tracing algorithm is initiated. First, the positive individual and their recent contacts are placed under quarantine for a specified time t_Q and tested once the quarantine has elapsed (before potentially being released from quarantine). The contact tracing algorithm is as follows:

- A list of contacts is kept for each individual. When the index case is tested positive, contacts more recent than a certain *retention time* are kept, while the rest are discarded. This retention time is set at 5 days in our simulations. A sensitivity analysis for this parameter is performed in the supporting information of Ref. [65]. We refer to the remaining contacts on the list as *traced individuals*.
- If a traced individual has been in (cumulative) contact with the index case for longer than a certain *contact threshold*, the traced individual is also placed in quarantine, with a minimum quarantine duration of t_Q .
- Once the quarantine duration t_Q has elapsed, the individual is tested and released if negative. Otherwise a new quarantine of duration t_Q is issued.

We generally measure the rate of testing in units of $1/\tau_I$, so a testing rate of e.g. 1 corresponds to having a 50% chance of being tested during the Infectious state (since the rate for leaving the state then equals the rate for being tested).

Perturbing heterogeneities. Our general objective is to assess the importance of different types of social contact heterogeneity for the effectiveness of contact tracing. To this end, we implement different ways of partially homogenizing contact patterns. By comparing the performance of the simple contact tracing algorithm on these partially homogenized networks with its performance on the *true* network, we can separately probe the importance of heterogeneities.

The first method is *edge swapping*. This algorithm, when performed on a contact network, preserves the degree distribution but destroys correlations in who contacts who. In other words, it is a way of probing the importance of the specific network structure, without conflating it with the degree distribution (number of contacts per person).

The second method, *randomization*, homogenizes the degree distribution as well as the network structure.

The edge swapping algorithm works as follows. The following steps are re-iterated many times, until each edge has been swapped multiple times on average:

- Randomly select two edges $A \leftrightarrow B$ and $C \leftrightarrow D$.
- Swap the chosen edges, leaving instead the edges $A \leftrightarrow C$ and $B \leftrightarrow D$.

Since any node which participates in this procedure merely has its edges *replaced* but never *destroyed*, the degree distribution remains intact.

The randomization algorithm is simpler yet. The following step is simply reiterated until each edge has been swapped multiple times on average:

- Select an edge $A \leftrightarrow B$ at random. Replace the endpoints A and B with nodes C and D chosen from the population at random.

This homogenizes the degree distribution as well as the network structure, but the overall level of social activity in the system is unchanged (edges are replaced with new edges, never destroyed).

Both algorithms are performed *within* each timestep, meaning that edges are swapped at times t_n and t_{n+1} independently of each other. Thus, the distribution of social contact *durations* is not preserved. For this reason, we also implemented a version of the edge swapping which is *duration preserving*. For the purposes of describing this algorithm, we introduce bit of notation. Let the collection of fixed-time contact networks at time steps $t = 1, 2, \dots, T$ be denoted by $G = \{G_t\}$. G_t is then the network of contacts at time step t , representing a five minute window. Duration-preserving edgeswapping requires a global (in time) view of connections, and cannot be performed on a single G_t . Before swapping, we run through the collection G and record the times at which each connection begins, labeling each with a persistent and unique connection ID. This allows the recognition of equivalent contacts across time steps. Once this has been done, the swapping algorithm proceeds very similarly to before:

- For each $t \in \{1, \dots, T\}$:
 - For each edge in G_t , check whether the contact was initiated at time t (i.e. is a new contact) or is a continuation of an existing one. If it is new, generate a connection ID L_i and add it to a list $L = \{L_i\}$.
 - Perform a random pairing of IDs in L , with the additional rule that only contacts of equal duration can be paired.
 - Perform the simple edge swapping algorithm described above on the pairings from the previous step.
 - Record swaps performed in this time step so that they can be consistently performed in succeeding time steps.

- Perform each swap recorded in previous time steps, unless it has elapsed. In the latter case, discard it.

3.2.2 RESULTS

Heterogeneity persists on outbreak time scale. We find that the distribution of daily contact events, as well as the distribution of contact events over 7 weeks both have a coefficient of variation close to 1 (1.03 and 0.95, respectively), consistent with an exponential distribution (see Figure 38). This shows that heterogeneity is pronounced and persists on a time scale relevant for the outbreak. It approximately represents a **quenched disorder**.

Social structure reduces epidemic severity. In the absence of mitigation, Figure 39 shows that the peak of the epidemic is much higher in both the randomized and the edge swapped networks than in the true network. This indicates that the network structure in itself is important in determining the epidemic peak.

The final size of the epidemic, on the other hand, is similar for the true and edge swapped networks, while it is much higher for the randomized network. This shows that heterogeneous social activity levels are important in reducing the final size of the epidemic, while social structure itself plays a smaller role.

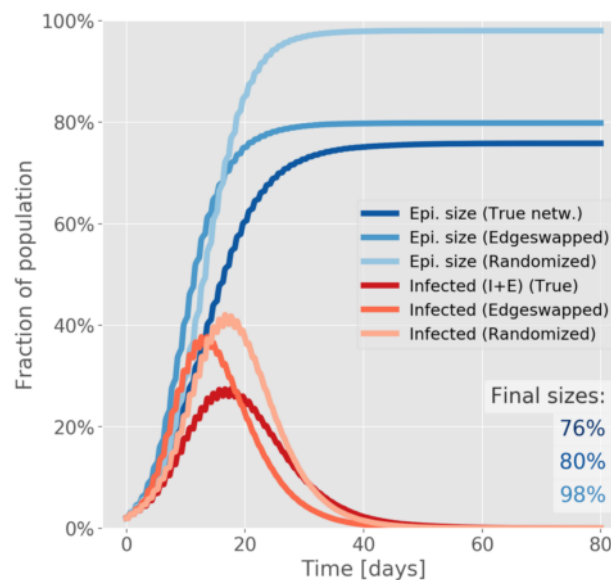


Figure 39: Trajectories of epidemics in the true empirical network as well as the edge swapped and randomized versions. Blue curves show cumulative epidemic sizes while red curves show current incidence.

The dependence of contact tracing on heterogeneity is controlled by the contact threshold. As shown in Figure 40, the relative effectiveness of contact tracing on the true and homogenized networks depends heavily on the contact *threshold*, i.e. the minimum duration of contact between an infected and a susceptible person before they are considered relevant for contact tracing. At a high threshold (125 minutes, panel c), the true network performs drastically better than all (partially) homogenized networks, including the duration-preserving edge swapped one. At shorter thresholds (15 minutes of cumulative contact, panel a), the story becomes more nuanced. Here, the true and edge swapped networks (whether duration-preserving or not) perform about equally well in terms of mitigative power. Furthermore, they all perform much better than the randomized network. However, it should be noted that the true network attains this mitigation with substantially lower levels of quarantine (panel b). In this sense, the contact tracing is more *targeted* in the true network, indicating lower socioeconomic cost.

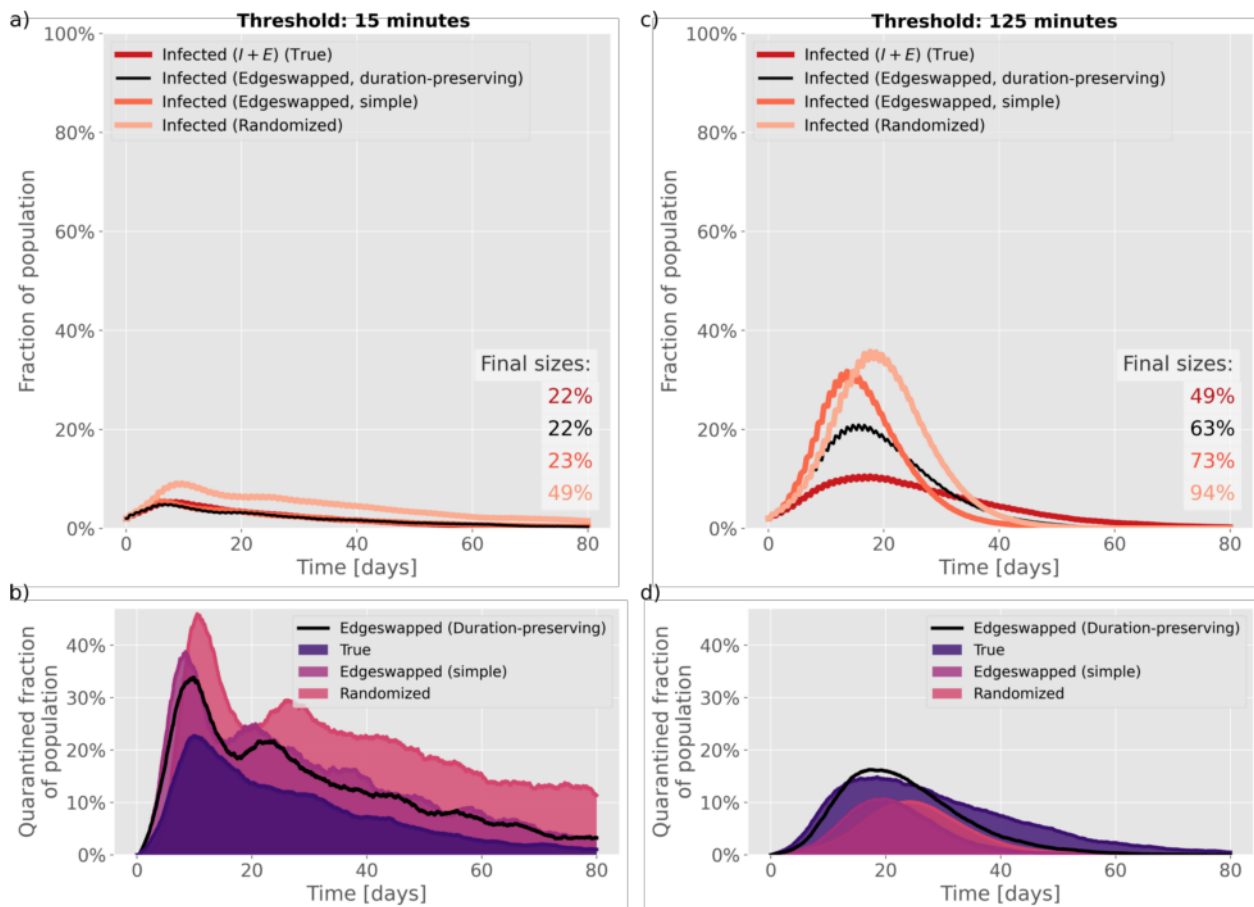


Figure 40: **a+c**) Epidemic trajectories at a contact threshold of either 15 (a) or 125 (c) minutes. Incidence curves for the true network as well as edge swapped, duration-preserving edgeswapped and randomized networks are shown. **b+d**) Quarantined fraction of the population as a function of time, for each of the four network types.

Contact tracing efficiency depends non-trivially on regular testing. As intuition would perhaps have it, the size of the epidemic is an increasing function of the contact threshold, while the total time spent in quarantine is a decreasing function (Figure 41b). When it comes to the rate of testing, the picture is more subtle (Figure 41a). For small values of the testing rate, the time spent in quarantine goes up when the testing rate is increased. This of course makes sense, since more cases are identified. However, at even higher testing rates, the mitigative effect begins to dominate, and the time that an average person spends in quarantine actually decreases. The net effect is that the function has a maximum at a testing rate of ~ 0.20 .

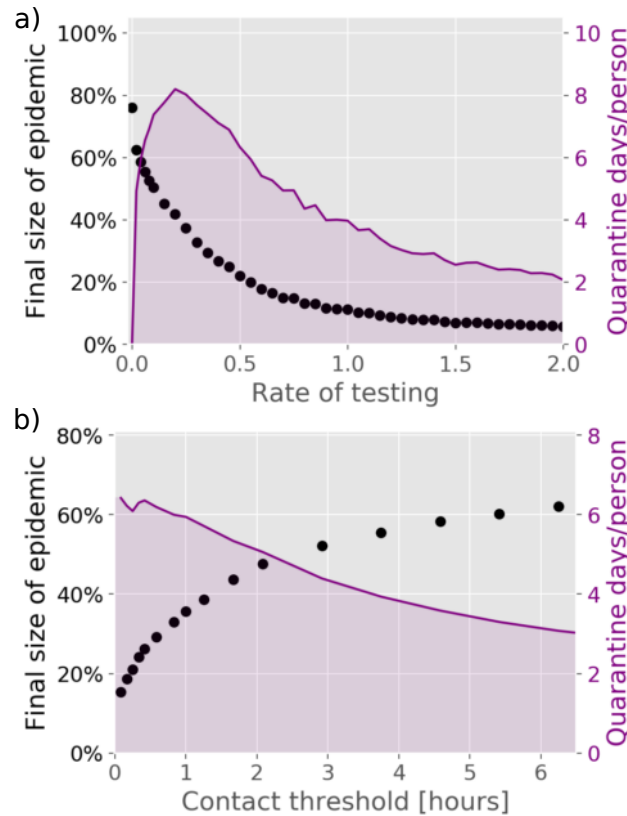


Figure 41: **a)** Dependence of effectiveness of contact tracing on the rate of testing (in the real, empirical network). Note that, while the final size of the epidemic is a monotonically decreasing function of regular testing rate, the total days spent in quarantine actually reaches a maximum around a rate of testing of 0.2, before declining. **b)** Contact tracing dependence on the contact threshold (in the real, empirical network).

3.3 DISCUSSION

In a broader perspective, the work outlined in sections 3.1 and 3.2 emphasizes the importance of taking stochasticity and heterogeneity into account when modeling disease spreading. Not only is it important to include a degree of randomness in spreading phenomena, but one must also realize that *noise is not just noise*. In the case of superspreading, we have seen that quenched disorder in transmission gives rise to a phenomenon which has profound impact on lockdown-type mitigation strategies which rely on broad reductions in contact numbers. However, more targeted interventions can also benefit from overdispersion if done correctly. In Ref. [68], the authors used a simple branching process model to estimate the impact of superspreading on backward (also known as “retrospective”) contact tracing and found that it, too, was enhanced. When an infected individual is discovered, this strategy relies on asking “Who was this person infected *by* and who else could *that* person have infected?” rather than “Who did this person infect?” as in forward contact tracing. Our simulations in [38] corroborate these findings. A simple way of estimating the relative efficacy of retrospective contact tracing is by measuring how many secondary cases each infected

person allows one to trace with and without superspreading. In a homogeneous mixing scenario, we find that the answer is 2.7 in the absence of overdispersion ($k = \infty$) and 24 with superspreading at a COVID-19-like level ($k = 0.1$). This is of course a best-case scenario. In practice, test-trace-isolate programmes based on retrospective contact tracing have further limitations, especially due to temporal constraints inherent to a disease with a relatively short generation time.

The importance of quenched noise is by no means limited to infectiousness and disease transmission. Our findings on the importance of (non-)repeated contacts in [37] were an example of quenched *social structure* versus homogeneous mixing. Network models often assume entirely static social networks – which may be reasonable over shorter time frames – while traditional aggregated compartmental models usually assume homogeneous mixing (which may be reasonable in certain environments). The truth necessarily lies somewhere in between, and in [65] we studied disease transmission on an empirical contact network precisely to disentangle these effects. As outlined above, we found that social activity patterns in the student population remained quite stable on the scale of a couple of months and that the different contact heterogeneities (network structure, differences in activity levels, differences in contact durations) each contributed to the effectiveness of the contact tracing strategy studied to varying degrees. In Ref. [69], the authors studied how “rewiring” of social networks over time may cause resurgences of an epidemic, again highlighting the importance of temporal correlations in social behaviour and the lack thereof.

The findings of section 3.2 highlight the importance of incorporating realistic contact heterogeneity into models of contact tracing if numerically credible conclusions are to be reached. While heterogeneity becomes especially crucial for mitigation strategies which rely directly on contact network structure, such as contact tracing, even unmitigated epidemics are significantly affected by social heterogeneities. Among the networks studied in this section, the *randomized* type closely mimics the homogeneous mixing seen in classical, aggregated S(E)IR models. As shown above, simulations on such a homogeneous network systematically and substantially overestimates the size of the epidemic relative to the empirical network. Some degree of social activity heterogeneity can actually be included in aggregated models, however only in a coarse-grained fashion. A recent study [70] did this and similarly found that epidemic size was reduced. The evident importance of heterogeneity and stochasticity, as well as the distinction between quenched and annealed noise are arguments in favour of agent based models.

3.4 PUBLICATIONS FOR CHAPTER 3

The third and final chapter of this thesis builds on the following manuscripts. The papers were written during this degree and I have not submitted them for any other academic degree.

1. K. Sneppen, **B. F. Nielsen**, R. J. Taylor, and L. Simonsen, “Overdispersion in COVID-19 increases the effectiveness of limiting nonrepetitive contacts for transmission control”, *Proceedings of the National Academy of Sciences* **118** (2021), no. 14.
2. **B. F. Nielsen**, L. Simonsen, and K. Sneppen, “COVID-19 superspreading suggests mitigation by social network modulation”, *Physical Review Letters* **126** (2021), no. 11, 118301.
3. **B. F. Nielsen**, A. Eilersen, L. Simonsen, and K. Sneppen, “Lockdowns exert selection pressure on overdispersion of SARS-CoV-2 variants”, *medRxiv* (2021).
4. **B. F. Nielsen**, K. Sneppen, L. Simonsen, and J. Mathiesen, “Social network heterogeneity is essential for contact tracing”, *medRxiv* (2020).
 - **Note:** The version included in this thesis is updated relative to the online preprint, and instead bears the title “Differences in social activity increase efficiency of contact tracing”.
5. S. Ørskov, **B. F. Nielsen**, S. Føns, K. Sneppen, and L. Simonsen, “The COVID-19 pandemic: Key considerations for the epidemic and its control”, *APMIS* (2021).

OVERDISPERSION IN COVID-19 INCREASES THE EFFECTIVENESS OF LIMITING NONREPETITIVE CONTACTS FOR TRANSMISSION CONTROL

Authors: Kim Sneppen¹, Bjarke Frost Nielsen¹, Robert J. Taylor² and Lone Simonsen².

¹ The Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark.

² Department of Science and Environment, Roskilde University, Roskilde, Denmark.

My contribution: Contributed to figure creation, data analysis and writing of the manuscript.

Publication status: Published in *Proceedings of the National Academy of Sciences of the United States of America (PNAS)* (2021).

Hyperlink(s): <https://doi.org/10.1073/pnas.2016623118>



Overdispersion in COVID-19 increases the effectiveness of limiting nonrepetitive contacts for transmission control

Kim Sneppen^{a,1}, Bjarke Frost Nielsen^a, Robert J. Taylor^b, and Lone Simonsen^b

^aNiels Bohr Institute, University of Copenhagen, 2100 København Ø, Denmark; and ^bDepartment of Science and Environment, Roskilde University, 4000 Roskilde, Denmark

Edited by Simon Asher Levin, Princeton University, Princeton, NJ, and approved March 8, 2021 (received for review August 5, 2020)

Increasing evidence indicates that superspreading plays a dominant role in COVID-19 transmission. Recent estimates suggest that the dispersion parameter k for severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is on the order of 0.1, which corresponds to about 10% of cases being the source of 80% of infections. To investigate how overdispersion might affect the outcome of various mitigation strategies, we developed an agent-based model with a social network that allows transmission through contact in three sectors: “close” (a small, unchanging group of mutual contacts as might be found in a household), “regular” (a larger, unchanging group as might be found in a workplace or school), and “random” (drawn from the entire model population and not repeated regularly). We assigned individual infectivity from a gamma distribution with dispersion parameter k . We found that when k was low (i.e., greater heterogeneity, more superspreading events), reducing random sector contacts had a far greater impact on the epidemic trajectory than did reducing regular contacts; when k was high (i.e., less heterogeneity, no superspreading events), that difference disappeared. These results suggest that overdispersion of COVID-19 transmission gives the virus an Achilles’ heel: Reducing contacts between people who do not regularly meet would substantially reduce the pandemic, while reducing repeated contacts in defined social groups would be less effective.

pandemic | overdispersion | mitigation strategies | superspreading | social networks

Countries worldwide have responded to the COVID-19 pandemic by implementing unprecedented “lockdown” strategies: closing schools and workplaces; closing or strictly regulating restaurants, bars, theaters, and other venues; and banning large gatherings. Such measures moderately reduced disease transmission in the 1918 Spanish influenza epidemic (1); however, in the COVID-19 pandemic, lockdowns have been highly effective, albeit at great cost to society (2). Not enough is known about which of the mitigation measures used during lockdowns is most effective. Understanding the relative contributions of reducing different types of contacts in different settings is essential for the current situation as well as for pandemic preparedness.

The occurrence of “superspreading events,” in which a large number of people are infected in a short time (often in a single location), is a well-documented aspect of the COVID-19 pandemic (3), from a string of superspreading events at fitness centers in Seoul, South Korea (4) to a wedding reception at the Big Moose Inn in Millinocket, ME at which at which over half the guests were infected (5).

Heterogeneity in transmission is well known in several infectious diseases (6–9), including the recent coronavirus threats severe acute respiratory syndrome (SARS) (10) and Middle East respiratory syndrome (MERS) (11). In 2005, Lloyd-Smith et al. (6) surveyed the importance of superspreading events across infectious diseases and pioneered the use of the “dispersion parameter” k to describe how the number of infections

generated by an individual is distributed around the mean, with lower values of k corresponding to a broader distribution.

Multiple studies have found that k for SARS-CoV-2 is on the order of 0.1, corresponding to ~10% of infected people causing 80% of new infections (12–15). This also implies that the majority of infected individuals cause less than one secondary infection and thus, cannot sustain the epidemic on their own should the superspreading events somehow be prevented.

Consistent with this, the household attack rate is low, as shown by several studies. In China, figures of 15 and 12% have been reported (13, 16), while a nationwide study from Denmark gave a household attack rate of 17% (17); in the context of a superspreading event in a South Korean call center, the household attack rate was 16% (18). The low household attack rate implies that most infected people do not even infect their household contacts. The overdispersion seen in SARS-CoV-2 stands in contrast to pandemic influenza, which was found to have a dispersion parameter of about $k = 1$ (19), so that 45% of infected people cause 80% of new infections.

Measurements of the level of transmission heterogeneity in COVID-19 have been based on several different methodologies, each having its own strengths and weaknesses. Perhaps the most direct measurement is by contact tracing (13). This method allows for a straightforward assessment of overdispersion but may be affected by biases inherent in contact tracing data, such as close contacts being more readily found or large outbreaks being more carefully investigated. Other studies have relied on aggregate incidence data (12, 15, 20) and even phylodynamic methods (14). These disparate studies found similar levels of

Significance

Evidence indicates that superspreading plays a dominant role in COVID-19 transmission, so that a small fraction of infected people causes a large proportion of new COVID-19 cases. We developed an agent-based model that simulates a superspreading disease moving through a society with networks of both repeated contacts and nonrepeated, random contacts. The results indicate that superspreading is the virus’ Achilles’ heel: Reducing random contacts—such as those that occur at sporting events, restaurants, bars, and the like—can control the outbreak at population scales.

Author contributions: K.S. and L.S. designed research; K.S. performed research; K.S., B.F.N., R.J.T., and L.S. analyzed data; and K.S., B.F.N., R.J.T., and L.S. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

This open access article is distributed under [Creative Commons Attribution License 4.0 \(CC BY\)](https://creativecommons.org/licenses/by/4.0/).

¹To whom correspondence may be addressed. Email: sneppen@nbi.ku.dk.

This article contains supporting information online at <https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2016623118/-DCSupplemental>.

Published March 19, 2021.

heterogeneity, increasing the robustness of the basic finding that overdispersion is high in COVID-19.

Given the importance of superspreading to COVID-19 transmission, modeling studies assessing the effect of different mitigation strategies would do well to take superspreading into account. Agent-based models, which set up a network of individual agents that interact according to defined rules, are well suited to exploring the impact of mitigation in the presence of superspreading. Like standard compartmental Susceptible, Exposed, Infected, Recovered (SEIR) models, they can reproduce the epidemic curves observed in a population in an unmitigated scenario. Unlike purely compartmental models, agent-based models can easily adjust individual infectivity and mimic repeated social interactions within defined groups, as might be found in households, schools, and workplaces. Agent-based models can also include different types of social interaction and phenomena such as a disease saturating some households or workplaces by infecting all susceptible agents.

We therefore developed an agent-based model with a social network structure to investigate how overdispersion might affect nonpharmaceutical mitigation efforts to control a superspreading disease such as COVID-19. In brief, we simulated epidemic trajectories in an agent-based model with a population of 1 million agents. Upon infection, agents transition from susceptible to exposed, infected, and recovered states (Fig. 1A); agents are on average infectious for 5.5 d. We allowed contacts of three types: close (within a small, unchanging group as might be found in a household or other close association), regular (within a larger, unchanging group as might be found in a workplace, school, extended family, or other social unit), and random (drawn randomly from the entire agent population and not repeated regularly) (Fig. 1B). We adjusted the contact rates to achieve a 1:1:1 ratio of contact time in the three sectors, consistent with survey data from Mossong et al. (21). Within the timescale set by the generation time of COVID-19, our close and regular networks can be considered constant. Contacts that occur less frequently belong to the random sector. To simulate superspreading, we assigned infectivity according to a gamma distribution with dispersion parameter $k = 0.1$ and adjusted the overall infectivity to produce an initial growth rate of 23% per day, as observed for COVID-19 in Europe and North America (22–24), which corresponded to a basic reproductive number of 2.5. In the unmitigated case, contacts were allowed in all three sectors; we then simulated two additional scenarios in which the regular and random contacts were restricted. These three scenarios were simulated under two conditions, with k set to infinity (no superspreading) and with k set to 0.1 (superspreading). The model is described in detail in *Methods*.

Our findings suggest that superspreading gives COVID-19 an Achilles' heel: Limiting contacts in the part of the social environment where many random contacts are encountered—and where superspreading events are most likely to occur—slows transmission dramatically and far more effectively than limiting contacts in social groups where people meet repeatedly, such as in the home, work, or school.

Results

We found that the presence of superspreading profoundly improves the impact of reducing random contacts in mitigating the epidemic. Regardless of whether superspreading is present in the model, the overall percentage of the population infected in a no mitigation scenario is 90% (Fig. 2). Thus, superspreading has hardly any effect on the trajectory of an unmitigated epidemic. Furthermore, comparing Fig. 2, it is clear that a mitigation strategy based on restricting regular contacts performs similarly in both the superspreading (Fig. 2B) and nonsuperspreading (Fig. 2A) scenarios. However, when a mitigation strategy based exclusively on restricting random contacts is employed in the superspreading scenario, the effect is dramatically enhanced:

The final epidemic size is just 15%, compared with 57% in the absence of superspreading.

We performed several sensitivity tests to investigate whether our findings were robust to changes in model parameters.

We varied the dispersion parameter k in the interval [0.05, 1.0] and found that as it increased, the effect of preventing random contacts gradually diminished (Fig. 3). This shows that the efficacy of random sector-based mitigation increases monotonically with the degree of superspreading. On the other hand, even partial mitigation of the random sector still had a considerable effect when $k = 0.1$ (*SI Appendix, Fig. S1*).

By adjusting the mean infection rate, we varied the initial epidemic growth rate from 16 to 30% per day (*SI Appendix, Fig. S2*), an interval that covers the range of premitigation growth rates observed in Europe and North America (22–24). We found, as expected, that a faster-growing epidemic is more difficult to mitigate; however, the enhanced effect of random sector mitigation when superspreading is present remains.

To assess the sensitivity of our results to the partitioning of the three social sectors, we varied the ratio of contacts in each sector from the base case of 1:1:1 to 2:2:1 for close, regular, and random contacts (*SI Appendix, Fig. S3A*) and increased the size of the groups from which regular and close contacts were drawn, respectively (*SI Appendix, Fig. S3B and C*). These variations had only a moderate negative effect on mitigation, reflecting that a mitigation strategy based on removing random contacts becomes relatively less effective if fewer random contacts are made in the premitigation scenario. In a related analysis, we analyzed the effect of introducing heterogeneity in the number of individuals with whom an agent interacts. We did this by letting half of the population spend only 1/6 of their contact time in the random sector while allowing the other half to spend 1/2 of their contact time interacting in the random sector. In this way, we maintained the overall activity in the random sector to be 1/3. The result was a moderate decrease in the degree of mitigation (*SI Appendix, Fig. S4*).

To determine the effect of increased heterogeneity in social activity, we exponentially distributed the overall contact time of individuals, so that some agents would make contact more frequently than others (*SI Appendix, Fig. S5*). This heterogeneity was found to decrease the epidemic size in general, similar to what Britton et al. (25) recently showed for COVID-19. Nonetheless, random sector-based mitigation remained by far the most effective.

Finally, we measured the distribution of the number of secondary infections arising in our simulations (*SI Appendix, Fig. S6*). This analysis is an important test of our model since it is crucial that the model reproduces the degree of transmission heterogeneity reported in the literature; the analysis also allows us to assess the degree of transmission heterogeneity introduced by the model's social structure alone. When we set the dispersion parameter for infectivity to $k = 0.1$ (our base superspreading scenario), the coefficient of variation (*CV*) of the observed distribution of secondary cases is 3.1, consistent with an observed k value of ~ 0.1 for a negative binomial distribution (6), indicating that the model has the desired level of transmission heterogeneity in our base superspreading scenario. When the distribution of infectiousness is taken to be homogeneous (i.e., the non-superspreading scenario [formally obtained at infinite k for infectivity]), the observed distribution of cases has a *CV* of 0.7, consistent with an observed k value of 3.3 for a corresponding negative binomial distribution. Thus, the social structure by itself contributes only very moderately to the transmission heterogeneity observed in our superspreading simulations.

Across the sensitivity analyses, our basic finding remains unchanged: In an epidemic driven by superspreading, restricting random nonrepeating contacts is far more effective than limiting the regular repeating contacts that occur in interconnected groups.

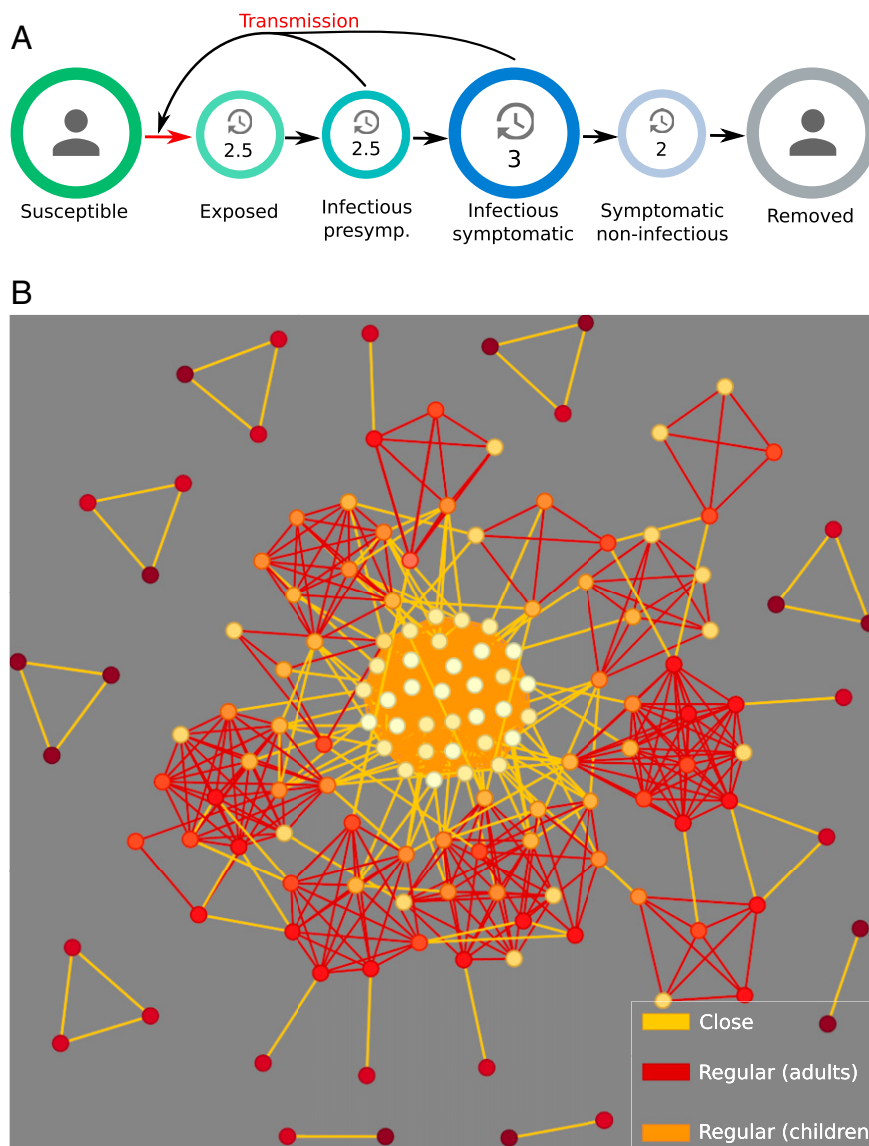


Fig. 1. (A) Schematic representation progression of disease in our agent-based model. Individual agents become infectious 2.5 d before symptom onset on average. Agents enter the recovered state after an average of 3 d of symptoms, giving an average total infectious period of 5.5 d. (B) Schematic representation of the connectivity between 150 agents. Individuals are represented as nodes, with shading indicating age (light = young, dark = older). Edges represent social connections, with bright yellow denoting close contacts, orange denoting regular contacts between adults, and red edges denoting regular contacts involving children. Random contacts are not pictured. The network diagram was generated by running our simulation on a smaller population of just 150 individuals, with the same rules for connectivity as in the full-scale simulations.

Discussion

Policy makers worldwide face excruciating choices as they seek to ease restrictions as much as possible without causing a surge in COVID-19 cases that would overwhelm health care systems, especially by exceeding available intensive care unit beds needed to keep critically ill COVID-19 patients alive. These policy choices must take new information into account as the pandemic unfolds.

Evidence is now overwhelming that superspreading plays a key role in COVID-19 transmission (12–15). Yet, models used to predict effects of mitigation strategies often do not consider this phenomenon (26–28). In this study, we built an agent-based model with an underlying social structure to take on this task.

Our results indicate that reducing random contacts has an out-sized effect in an epidemic characterized by superspreading; in the absence of superspreading, the same mitigation strategy is much less effective. This means that mitigation policies should focus on limiting contacts during activities that bring together large numbers

of people who would otherwise not routinely come into contact, such as at sporting events, restaurants, bars, weddings, funerals, and religious services; repeated contacts that occur in smaller social groups are much less important. If such gatherings cannot be avoided, steps such as wearing face masks and moving events outdoors might also help. Our results also suggest that in complex settings such as workplaces and schools, which have characteristics of both our regular and random sectors, preventing congregation of large groups of people who would otherwise rarely meet is important.

Why does our model suggest that the presence of superspreaders favors these policy choices? When random contacts are prevented, regular contacts become the main source of infection. However, because the number of possible connections is limited in a regular social unit, a highly infectious individual soon runs out of susceptible contacts. When random contacts are allowed, however, there is no such limitation because as far as the superspreading agent is concerned, every contact is new. It follows that an epidemic driven by superspreading is

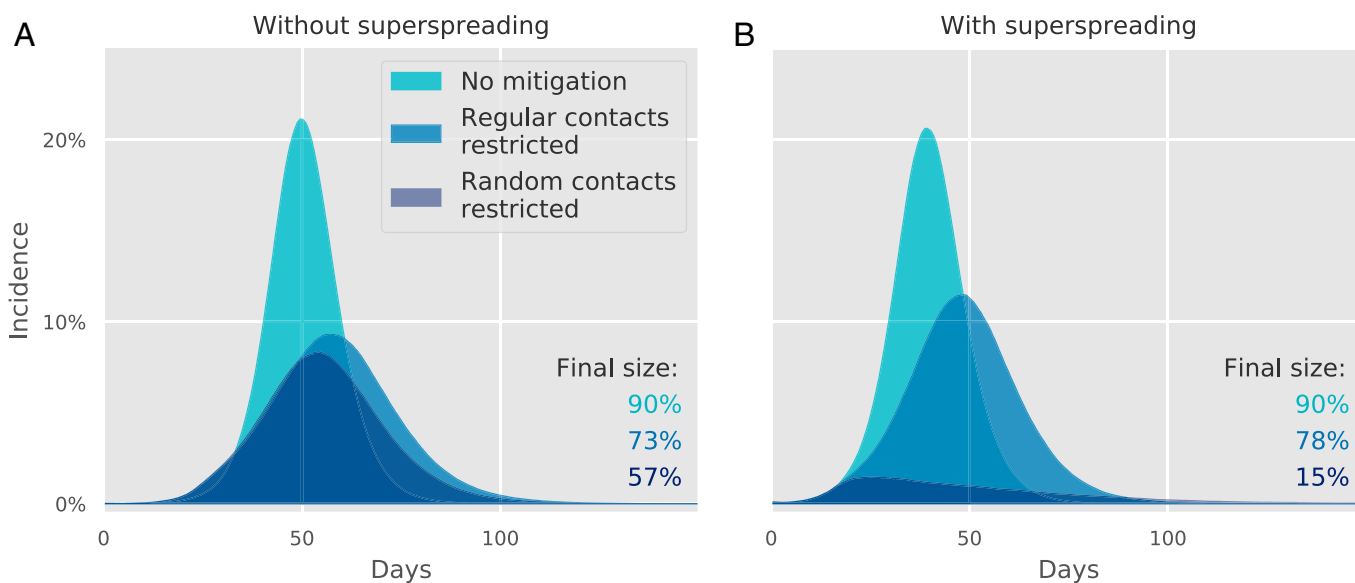


Fig. 2. The impact of mitigation on modeled incidence. Simulation of epidemic trajectories with mitigation starting when 1% of the population has been infected. In each panel, we show three trajectories corresponding to the unmitigated epidemic, the case where we completely restrict all regular contacts, and the case where we restrict all random contacts. When superspreading is not present (i.e., k is infinite; *A*), the effect of eliminating regular and random contacts is similar; however, when superspreading is a factor in transmission (i.e., when $k = 0.1$; *B*), the effect of eliminating random contacts is dramatically enhanced. We did not consider mitigation by limiting close contacts as this would not be a credible mitigation strategy.

fueled more by the diversity of contacts—the total number of different people encountered—and less by the duration of contacts—how long one spends with each. Thus, preventing random contacts in the model provides more benefit than preventing regular contacts.

It is worth noting that an equal ratio of contact time across sectors does not mean that the number of secondary infections is the same in each. Even when k is high so that superspreading is not present (Fig. 2*A*), about 40% of transmissions occur during random contacts because the saturation effect is small. When k is low and superspreading is present, this fraction increases to about 60%, the removal of which corresponds to a 2.5-fold reduction in the reproductive number of the disease—a reduction sufficient to mitigate the epidemic (Fig. 2*B*).

Our finding that the propagation of an overdispersed disease is more sensitive to the many random contacts (rather than the few but persistent regular contacts) is broadly applicable, regardless of the underlying biological mechanism. If, for example, one considers a disease where the high reproductive number of some individuals is the result of a prolonged infectious period, transmission would still be limited by the number of different persons an individual encounters. In our model, this number is set by the combined size of their close and regular contacts, when access to random contacts is restricted.

The most important limitation of our study is the model's simplicity compared with the complex reality of human society. Our social structure does not precisely reproduce the complex and fluid interactions of human societies. However, our division of contacts approximates the range of possible interactions, from familiar to random. We relegated all nonrepeating contacts to the random sector, so that contacts with known persons occurred only through two fixed social networks, one small and one somewhat larger. In the real world of large families, workplace cafeterias, school playgrounds, and neighborhood restaurants, many interactions in the random sector would be with familiar but rarely seen people such as old friends and extended family; likewise, some contacts with random people would occur in places dominated by repeat contacts with familiar people. We simply separated those into two artificially distinct spheres.

The mechanism that underlies superspreading is not understood, but relevant factors include both the rate at which an

infected person sheds the virus and the environment in which the virus is shed, including the density of people and their susceptibility. Behavior, including shouting or singing, can increase both the rate of viral shedding and the susceptibility to infection, and a gathering in a closed room with poor ventilation involves considerably higher risk than one outdoors (29, 30). Superspreading has been broadly categorized in three main categories: biological, behavioral/social, and opportunistic (31). However, these categories are not mutually exclusive, and superspreading is generally a question of means (high infectiousness) and opportunity (social and environmental context). In order for a superspreading event to occur, a highly infectious individual must have access to a large number of distinct contacts. In our model, the means is simulated by assigning a distribution of individual infectiousness from a gamma distribution. While we do not specifically model events, we do allow many contacts in the random sector, which allows some agents to cause large clusters of secondary infections.

Other recent studies modeling superspreading in COVID-19 have generally come to the conclusion that “cutting the tail” (i.e., targeted elimination of superspreaders) would be an effective

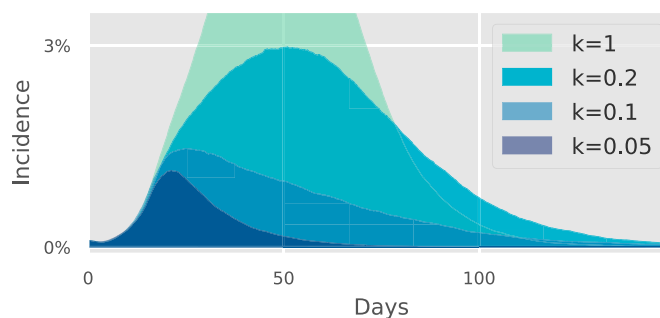


Fig. 3. Sensitivity of model results to dispersion factor k . When an epidemic size of 1% is reached, a mitigation scheme consisting of restricting all random contacts is initiated. We explore the epidemic trajectories for different values of the overdispersion factor k . As k decreases (i.e., transmission heterogeneity increases), eliminating random contacts has a progressively greater effect.

Table 1. Distribution of simulated population by age group (40), with conditional probabilities for relative social contact time (21)

Age (y)	Percentage of population	Relative social time per person
0–9	10.9	1.21
10–19	11.9	1.70
20–29	13.3	1.45
30–39	11.7	1.45
40–49	13.6	1.38
50–59	13.6	1.31
60–69	11.7	1.06
70–79	8.9	0.81
80+	4.3	0.81

means of mitigation (31–33). What is less clear is how to construct policies to accomplish that and how to identify the situations and modes of contact which are likely to lead to superspreading. By distinguishing between repeated and random contacts, our model points to a feasible population-wide mitigation strategy. This is not possible in well-mixed (32), branching process (31), or purely network-based models (33), which do not incorporate different types of social contacts.

The social network underlying our model is of the “small-world” variety (34), insofar as it is characterized by cliquishness and short typical distances between nodes. Thus, any given node in our model can typically be reached by moving through only a few close and regular units. Block et al. (27) recently used small-world networks to explore how mitigation strategies that alter typical nodal distance and cliquishness affect the epidemic trajectory. In the same vein, Leng et al. (35) studied the influence of social bubbles on mitigation efforts using an agent-based model with three levels of transmission: within households, between households in the same bubble, and lastly, community spread (akin to our random sector). However, none of these papers addressed the effect of superspreading on the mitigation strategies. Our results lend support to mitigations based on cutting links between cliques (27, 35) since the mixing of different close and regular groups occurs primarily through encounters in the random sector. Our work further shows that this kind of mitigation strategy is enhanced in a pandemic characterized by superspreading, as illustrated by Fig. 2B (compared with Fig. 2A).

Superspreading is a defining feature of the COVID-19 pandemic; a relatively small minority of the population causes the majority of infections, while most do not even infect people in their own household. As it is not possible to identify these superspreaders before transmission occurs, we here suggest an effective alternative strategy, namely that policies should aim to reduce contact diversity, rather than attempt to limit total contact time. This means that mitigation policies should focus on limiting activities that bring together many people who would otherwise not routinely come into contact.

Methods

We developed an age-stratified, agent-based model with three sectors of social contact through which the disease can be transmitted. Each agent is assigned to one close and one regular unit, within which contacts are repeated over time, and participates in random contacts drawn from the entire population.

Agents are stratified by age in 10-y intervals and assigned age-dependent social activity levels a_i , which are adjusted such that the observed contact

rates in an unmitigated scenario fit the age-dependent activity given in Table 1 (21). Close units have some properties of households: an average of 2.3 members, adults are in the same or adjacent age bands, and children are taken to be 20 to 40 y younger than adults in the same unit. The CV of the generated close contact network sizes is 0.59. This may be compared with The European Union Statistics on Income and Living Conditions Survey, which reports an average household size of 2.3 with a CV of 0.57 (36). Regular units have properties of workplaces and schools: Agents 20 to 70 y of age are assigned to a Poisson-distributed cluster with an average of eight agents. Agents under 20 y old are assigned a regular unit of 18 members. Each of these units is also assigned two adults aged 20 to 70. Agents older than 70 y are not assigned to a regular unit. Random contacts are chosen from the entire population at random for each infection attempt to simulate brief contacts without temporal correlation.

The progression of the disease is modeled in an SEIR framework, with agents passing through each stage at a rate determined by the average durations given in Fig. 1. The exposed state is subdivided into four stages, each of 1.25 d in length, with a constant probability rate for transitioning from one stage to the next. The first two of these stages comprise the gamma-distributed preinfectious state (average total duration: 2.5 d, SD: 1.8 d). The next two stages comprise the presymptomatic infectious state (average total duration: 2.5 d, SD: 1.8 d). This is followed by the infected state, in which agents are infectious and symptoms may be displayed [average total duration: 3 d (37, 38), SD: 3 d]. Agents then pass into the recovered state where they are no longer infectious. Simulations are run in a population of 1 million, randomly seeded with 100 infected agents. Agents are assigned a gamma-distributed infectivity $\beta \cdot s_i$, where s_i is drawn from a gamma distribution $P(s)$, proportional to $s^{k-1} \exp(-k s)$ with continuous $s > 0$ [Lloyd-Smith et al. (6)]. Here, k is the dispersion parameter, which determines the CV of the distribution according to $CV = 1/\sqrt{k}$. The rate constant β is calibrated to reproduce the observed initial exponential growth rate of 23% per day of an unmitigated COVID-19 epidemic (22–24).

In each time step of size Δt (of 30-min duration), each infected agent has an age-dependent probability for making contact to another agent; for each such contact, a contact partner is drawn from one of the three social sectors. The rate at which each of these sectors is chosen is based on a population-based survey of mixing patterns in eight European countries by Mossong et al. (21). That study found that the “home” sector made up 19 to 50% of all contacts, while the “work/school” sector accounted for 23 to 37%, and the remaining sectors amounted to 27 to 44%. For our model, we approximated this stratification by letting one-third of all contacts fall into each of the three sectors, for our base case. In *SI Appendix*, Fig. S2, we investigate the effect of varying these sector-specific social contact frequencies. Potential targets for infection are selected proportional to the age-dependent social activity listed in Table 1.

At each contact, the disease is transmitted with probability $P_t = \beta s_i \Delta t$. The time step length is chosen small enough to ensure that the probability of infection in any given time step is always less than one, even for the most infectious individuals. We simulate mitigation strategies by not permitting infection in a chosen fraction of contacts in one or more of the contact sectors. Mitigation is initiated when the infected population reaches 1% of the total. When mitigation by reduction of random contacts is performed, social networks are kept fixed, and the same numbers of contacts are removed in superspreading and nonsuperspreading scenarios to facilitate direct comparison.

To analyze the impact of heterogeneous social activity, we assigned each agent a separate activity parameter a_i selected from an exponential distribution (*SI Appendix*, Fig. S5). At each contact attempt from agent i to agent j , if $a_i < a_j$ then the contact proceeds as usual; however, if $a_i > a_j$, then the contact proceeds with a probability a_j/a_i . This procedure yields an exponential distribution of observed social activity, with more active agents being removed from the susceptible pool earlier in the epidemic.

Data Availability. Model code data have been deposited in GitHub (<https://github.com/NBIBioComplexity/SuperCoV>) (39).

ACKNOWLEDGMENTS. We thank Raul Donangelo, Viggo Andreasen, and Andreas Eilersen for enlightening discussions and multiple corrections and suggestions to the manuscript.

- R. J. Hatchett, C. E. Mecher, M. Lipsitch, Public health interventions and epidemic intensity during the 1918 influenza pandemic. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 7582–7587 (2007).
- S. Flaxman et al., Imperial College COVID-19 Response Team, Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature* **584**, 257–261 (2020).
- K. Swinkels, SARS-CoV-2 Superspreading Events Database (2020). <https://kmswinkels.medium.com/covid-19-superspreading-events-database-4c0a7aa2342b>. Accessed 13 October 2020.

- S. Jang, S. H. Han, J. Y. Rhee, Cluster of Coronavirus disease associated with fitness dance classes, South Korea. *Emerg. Infect. Dis.* **26**, 1917–1920 (2020).
- P. Mahale et al., Multiple COVID-19 outbreaks linked to a wedding reception in rural Maine—August 7–September 14, 2020. *MMWR Morb. Mortal. Wkly. Rep.* **69**, 1686–1690 (2020).
- J. O. Lloyd-Smith, S. J. Schreiber, P. E. Kopp, W. M. Getz, Superspreading and the effect of individual variation on disease emergence. *Nature* **438**, 355–359 (2005).
- R. A. Stein, Super-spreaders in infectious diseases. *Int. J. Infect. Dis.* **15**, e510–e513 (2011).

8. R. M. May, R. M. Anderson, Transmission dynamics of HIV infection. *Nature* **326**, 137–142 (1987).
9. M. E. J. Woolhouse *et al.*, Heterogeneities in the transmission of infectious agents: Implications for the design of control programs. *Proc. Natl. Acad. Sci. U.S.A.* **94**, 338–342 (1997).
10. S. Riley *et al.*, Transmission dynamics of the etiological agent of SARS in Hong Kong: Impact of public health interventions. *Science* **300**, 1961–1966 (2003).
11. A. J. Kucharski, C. L. Althaus, The role of superspreading in Middle East respiratory syndrome coronavirus (MERS-CoV) transmission. *Euro Surveill.* **20**, 14–18 (2015).
12. A. Endo, S. Abbott, A. J. Kucharski, S. Funk; Centre for the Mathematical Modelling of Infectious Diseases COVID-19 Working Group, Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China. *Wellcome Open Res.* **5**, 67 (2020).
13. Q. Bi *et al.*, Epidemiology and transmission of COVID-19 in 391 cases and 1286 of their close contacts in Shenzhen, China: A retrospective cohort study. *Lancet Infect. Dis.* **20**, 911–919 (2020).
14. D. Miller *et al.*, Full genome viral sequences inform patterns of SARS-CoV-2 spread into and within Israel. *Nature* **11**, 5518 (2020).
15. M. S. Y. Lau *et al.*, Characterizing superspreading events and age-specific infectiousness of SARS-CoV-2 transmission in Georgia, USA. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 22430–22435 (2020).
16. Q. L. Jing *et al.*, Household secondary attack rate of COVID-19 and associated determinants in Guangzhou, China: A retrospective cohort study. *Lancet Infect. Dis.* **20**, 1141–1150 (2020).
17. F. P. Lyngse *et al.*, COVID-19 transmission within Danish households: A nationwide study from lockdown to reopening. *medRxiv* [Preprint] (2020). 10.1101/2020.09.09.20191239 (Accessed 17 March 2021).
18. S. Y. Park *et al.*, Coronavirus disease outbreak in call center, South Korea. *Emerg. Infect. Dis.* **26**, 1666–1670 (2020).
19. C. Fraser, D. A. T. Cummings, D. Klinkenberg, D. S. Burke, N. M. Ferguson, Influenza transmission in households during the 1918 pandemic. *Am. J. Epidemiol.* **174**, 505–514 (2011).
20. J. B. Kirkegaard, K. Sneppen, Variability of individual infectiousness derived from aggregate statistics of COVID-19. *medRxiv* [Preprint] (2021). 10.1101/2021.01.15.21249870 (Accessed 17 March 2021).
21. J. Mossong *et al.*, Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Med.* **5**, e74 (2008).
22. A. Remuzzi, G. Remuzzi, COVID-19 and Italy: What next? *Lancet* **395**, 1225–1228 (2020).
23. Our World in Data, GitHub - Owid/Covid-19-Data: Data on COVID-19 (Coronavirus) Cases, Deaths, Hospitalizations, Tests. GitHub (2020). <https://github.com/owid/covid-19-data>. Accessed 1 May 2020.
24. N. Afshordi, B. Holder, M. Bahrami, D. Lichtblau, Diverse local epidemics reveal the distinct effects of population density, demographics, climate, depletion of susceptibles, and intervention in the first wave of COVID-19 in the United States. *medRxiv* [Preprint] (2020). 10.1101/2020.06.30.20143636 (Accessed 17 March 2021).
25. T. Britton, F. Ball, P. Trapman, A mathematical model reveals the influence of population heterogeneity on herd immunity to SARS-CoV-2. *Science* **369**, 846–849 (2020).
26. N. G. Davies, A. J. Kucharski, R. M. Eggo, A. Gimma, W. J. Edmunds; Centre for the Mathematical Modelling of Infectious Diseases COVID-19 working group, Effects of non-pharmaceutical interventions on COVID-19 cases, deaths, and demand for hospital services in the UK: A modelling study. *Lancet Public Health* **5**, e375–e385 (2020).
27. P. Block *et al.*, Social network-based distancing strategies to flatten the COVID-19 curve in a post-lockdown world. *Nat. Hum. Behav.* **4**, 588–596 (2020).
28. K. Prem *et al.*; Centre for the Mathematical Modelling of Infectious Diseases COVID-19 Working Group, The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: A modelling study. *Lancet Public Health* **5**, e261–e270 (2020).
29. Centers for Disease Control and Prevention, Scientific Brief: SARS-CoV-2 and Potential Airborne Transmission (2020). <https://www.cdc.gov/coronavirus/2019-ncov/more/scientific-brief-sars-cov-2.html>. Accessed 19 November 2020.
30. S. Asadi *et al.*, Effect of voicing and articulation manner on aerosol particle emission during human speech. *PLoS One* **15**, e0227699 (2020).
31. B. M. Althouse *et al.*, Superspreading events in the transmission dynamics of SARS-CoV-2: Opportunities for interventions and control. *PLoS Biol.* **18**, e3000897 (2020).
32. M. P. Kain, M. L. Childs, A. D. Becker, E. A. Mordecai, Chopping the tail: How preventing superspreading can help to maintain COVID-19 control. *Epidemics* **34**, 100430 (2021).
33. F. Wong, J. J. Collins, Evidence that coronavirus superspreading is fat-tailed. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 29416–29418 (2020).
34. D. J. Watts, S. H. Strogatz, Collective dynamics of ‘small-world’ networks. *Nature* **393**, 440–442 (1998).
35. Trystan Leng *et al.*; Centre for Mathematical Modelling of Infectious Disease 2019 nCoV Working Group, The effectiveness of social bubbles as part of a Covid-19 lockdown exit strategy, a modelling study. *Wellcome Open Res* **5**, 213, 10.12688/wellcomeopenres.16164.1 (2020).
36. Eurostat, EU statistics on income and living conditions microdata 2004–2019, release 2 in 2020. <https://doi.org/10.2907/EUSILC2004-2019V.1>. Accessed 17 March 2021.
37. X. He *et al.*, Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nat. Med.* **26**, 672–675 (2020).
38. J. Griffin *et al.*, Rapid review of available evidence on the serial interval and generation time of COVID-19. *BMJ Open* **10**, e040263 (2020).
39. K. Sneppen, B. F. Nielsen, SuperCoV. GitHub. <https://github.com/NBIBioComplexity/SuperCoV>. Deposited 27 January 2021.
40. Statistics Denmark, Population in Denmark (May 1, 2020). <https://www.dst.dk/en/Statistik/emner/befolkning-og-valg/befolkning-og-befolkningsfremskrivning/folketal>. Accessed 19 November 2020.

COVID-19 SUPERSPREADING SUGGESTS MITIGATION BY SOCIAL NETWORK MODULATION

Authors: Bjarke Frost Nielsen¹, Lone Simonsen² and Kim Sneppen¹.


¹ The Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark.

² Department of Science and Environment, Roskilde University, Roskilde, Denmark.

My contribution: Contributed to conceptualization and development, programming of the model, simulations, analytical calculations, creating figures as well as writing of the manuscript.

Publication status: Published in *Physical Review Letters* (2021). Highlighted as *Editor's Suggestion* and featured in *Physics* with accompanying article "Heterogeneity Matters When Modeling COVID-19".

Hyperlink(s): <https://doi.org/10.1103/PhysRevLett.126.118301>

COVID-19 Superspreading Suggests Mitigation by Social Network ModulationBjarke Frost Nielsen^{1,*}, Lone Simonsen^{2,†} and Kim Sneppen^{1,‡}¹*Niels Bohr Institute, University of Copenhagen, Blegdamsvej 17, 2100 Copenhagen, Denmark*²*Department of Science and Environment, Roskilde University, Universitetsvej 1, 4000 Roskilde, Denmark* (Received 1 October 2020; revised 6 January 2021; accepted 22 January 2021; published 15 March 2021)

Although COVID-19 has caused severe suffering globally, the efficacy of nonpharmaceutical interventions has been greater than typical models have predicted. Meanwhile, evidence is mounting that the pandemic is characterized by superspreading. Capturing this phenomenon theoretically requires modeling at the scale of individuals. Using a mathematical model, we show that superspreading drastically enhances mitigations which reduce the overall personal contact number and that social clustering increases this effect.

DOI: [10.1103/PhysRevLett.126.118301](https://doi.org/10.1103/PhysRevLett.126.118301)

During the ongoing COVID-19 pandemic, news stories have frequently appeared detailing spectacular events where single individuals—so-called *superspreaders*—have infected a large number of people within a short time frame [1–3]. By now, there is substantial evidence that these are not just singular events but that they reflect a marked transmission heterogeneity [4–6], a signature feature of the disease. In a well-mixed population, such heterogeneity has little bearing on the trajectory of an epidemic, but, when public sphere contacts are restricted, heterogeneity takes on a decisive role, as shown in Ref. [7]. In this Letter, we investigate the effects of transmission heterogeneity—i.e., superspreading—on mitigation strategies which rely on a general reduction in social network size and probe the influence of social clustering on such interventions.

The origins of superspreading can be diverse, depending on the characteristics of the pathogen in question. Superspreading events may occur due to circumstances and behavior as well as biology. Even medical procedures, such as intubation and bronchoscopy, which facilitate the production of aerosols [8], can lead to superspreading events in respiratory diseases. However, the most straightforward model of superspreading is that some individuals simply shed the virus to a much greater extent than the average infected person. For COVID-19, this “*biological superspreader*” phenomenon has some traction and is supported by the observation that household transmission is limited, despite the relatively high *average* infectiousness of COVID-19 [9–11].

Superspreading is not a phenomenon which is particular to SARS-CoV-2 but has been observed in connection with several other pathogens, including coronaviruses such as SARS [12,13] and MERS [14], as well as in diseases such as measles [15] and Ebola virus disease [16,17]. Pandemic influenzas such as the 1918 Spanish flu, on the other hand, are believed to be far more “democratic” [18]. The heterogeneity of transmission is usually quantified using the Gamma distribution [15]. This is the origin of the *dispersion parameter* or *k value*, which determines the fraction of infectious individuals who account for the majority of infections (Fig. 1). Smaller *k* means greater heterogeneity—in fact, when *k* is small ($|k| \ll 1$), it approximates the fraction of infected individuals who give rise to 80% of infections. For COVID-19, which is believed to have a *k* value of perhaps 0.1 [4–6], the most infectious 10% of individuals thus cause approximately 80% of infections.

The fundamental difference between a homogeneously spreading disease and a highly heterogeneous one is reflected in the infection networks they give rise to, as visualized in Fig. 1. When only a small fraction of individuals cause the bulk of infections, a reduction in social network connectivity amounts to decreasing the likelihood that a superspreader infects another superspreader and thus propagates the disease. Consequently, in a network characterized by superspreading [Fig. 1(a)], the outbreak can be stopped by cutting only a few select edges. Not so for the network in Fig. 1(c).

In this Letter, we present a model of superspreading phenomena which assumes that the driving force is a biological heterogeneity in infectiousness. We implement this as an agent-based model with contact networks and are also able to capture much of the phenomenology in analytical formulas. In the model, *N* agents are placed as the nodes in a contact network. We investigate different types of network, but our base case is the Erdős-Renyi

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

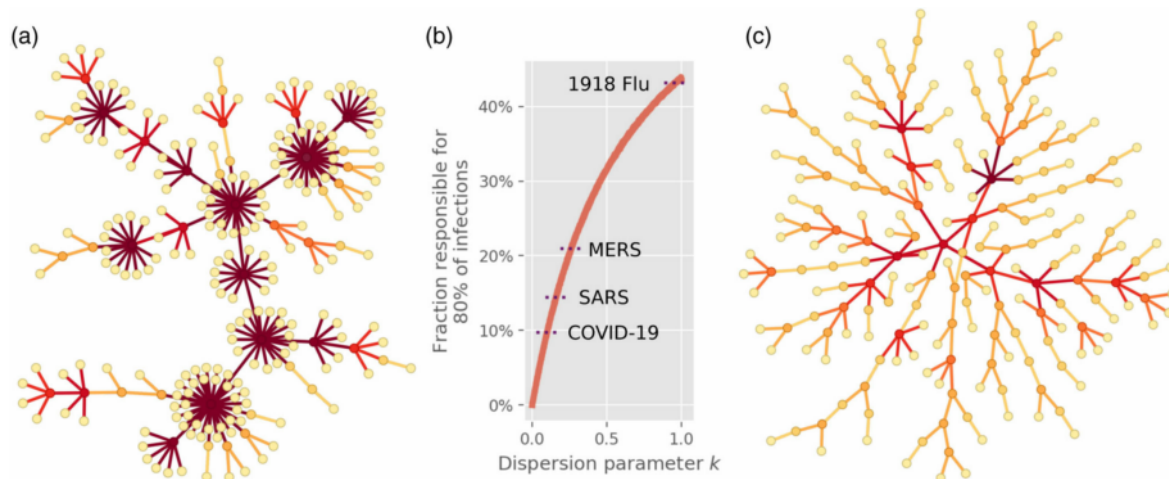


FIG. 1. The characteristics of superspreading. (a) Simulated infection network characterized by superspreading, with a dispersion parameter $k = 0.1$, within what has been observed for COVID-19 [4,5]. Superspreaders appear as hubs, while most individuals are “dead ends,” meaning that they do not transmit the disease. The epidemic mainly grows by spreading from one superspreader to the next. (b) The dispersion parameter k provides a measure of superspreading, with lower k values corresponding to a greater heterogeneity. With a k value of 0.1 for COVID-19, approximately 10% of the population has the infectiousness to cause 80% of transmission. SARS and MERS are also characterized by a significant heterogeneity [14,15], while pandemic influenza is believed to be more homogeneous [18]. (c) Simulated infection network without superspreading (all individuals have equal infectiousness). Here, most individuals spread the disease to a few others, leading to a branched structure.

network, which is characterized by a Poisson degree distribution and an absence of clustering.

At initialization, the infectiousness of each individual is drawn from a Gamma distribution [15]. As such, it is an innate property of each individual. The possible states of each individual are susceptible, exposed, infected, and recovered (for details, see Supplemental Material [19], which includes Refs. [20–27]). At each time step, each individual randomly selects one of its contacts to interact with, meaning that only a subset of the network is active at any given time. While a link between an infectious and a susceptible individual is active, there is a constant probability of infection per unit of time, as determined by the individual infectiousness.

This basic setup also lends itself to analytic calculations, as long as saturation effects can be ignored. Consider a single infected person who has c contacts, who are all assumed susceptible. First, the infectiousness r of the individual is drawn from a gamma distribution $P_I(r)$ with dispersion parameter k and mean μ . The distribution of the reproductive number R of an individual with a *known* infectiousness r and degree (i.e., connectivity) c is given by

$$P(R; r, c) = \binom{c}{R} (1 - e^{-r/c})^R (e^{-r/c})^{(c-R)}. \quad (1)$$

Taking the variability in infectiousness into account, the overall distribution of R becomes

$$P(R; c) = \int_{r=0}^{r=\infty} dr P_I(r) P(R; r, c). \quad (2)$$

In the limit of infinite connectivity, corresponding to a well-mixed population, this becomes a negative binomial distribution. That particular case has been studied in Ref. [15]. Given a contact network and a corresponding degree distribution $P_C(c)$ —for example, a Poisson distribution in the case of an Erdős-Renyi network—the connectivities can be summed over to yield a distribution of individual reproductive numbers, $P(R) = \sum_c P_C(c) P(R; c)$.

As reflected in the equations above, the *actual* number of secondary infections depends not only on biological infectiousness. In Figs. 2(a) and 2(b), we use this analytical framework to explore how the number of personal contacts affects the resultant distribution of infections. Without superspreading [Fig. 2(a)], a reduction in the contact number has a very modest effect and the distributions overlap. When the heterogeneity is at a COVID-like level [Fig. 2(b)], it is quite a different story. Here, a decrease in mean connectivity has a considerable effect, and mitigation suddenly looks feasible. Previously, another mitigation strategy which benefits from superspreading was suggested by Ref. [15], with the crucial difference that it relies on prior identification and targeting of superspreaders, in contrast to the broad reduction in mean connectivity explored here.

To quantify the sensitivity of the epidemic to social network size, we consider the basic reproductive number

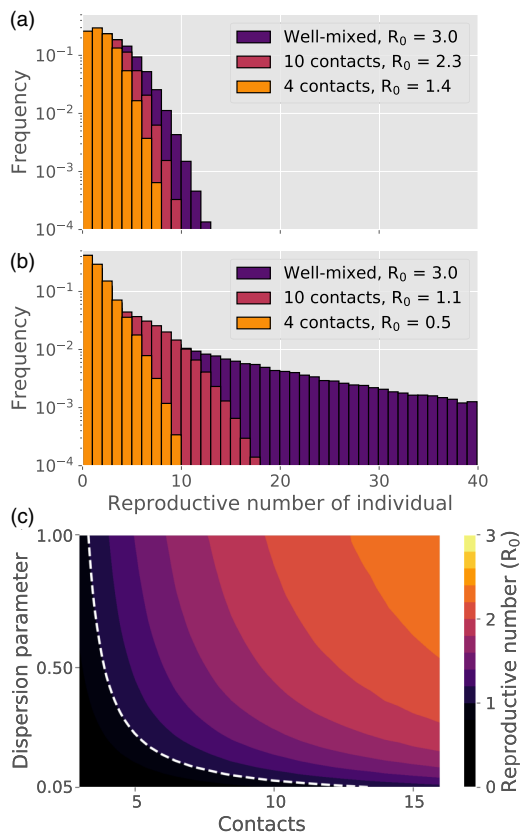


FIG. 2. The reproductive number. Distributions of individual reproductive number R and value of R_0 for different dispersion parameters and number of social contacts during an infectious period. This figure is based on the analytical framework described in the main text. See Supplemental Material [19] for details on the calculation. (a) Distribution of R for a disease where all individuals have equal infectiousness. (b) Distribution of R for a disease characterized by superspreading (dispersion parameter $k = 0.1$). (c) Basic reproductive number R_0 as a function of social connectivity and dispersion. The dashed line represents $R_0 = 1$. These calculations take into account the Poisson distributed contact number and the fact that each infectious person will have one insusceptible person in their network (the individual from whom the infection originated), even when computing the *basic* reproductive number. Details on an analytic computation of R_0 for *fixed* (δ -distributed) contact number can be found in Supplemental Material [19].

R_0 , meaning the average number of infections that each infected person causes in a situation where all contacts are still susceptible. In Fig. 2(c), the R_0 is given as a function of the dispersion parameter k and the average contact number. The epidemic is evidently much more sensitive to reductions in contact numbers when the transmission heterogeneity is high. A mitigation in which the average number of contacts goes from being unrestricted, down to about 10, causes a reduction in R_0 which lowers both the *peak* and *total* number of persons infected during the course of the epidemic (the *attack rate*). The overall trajectory of a

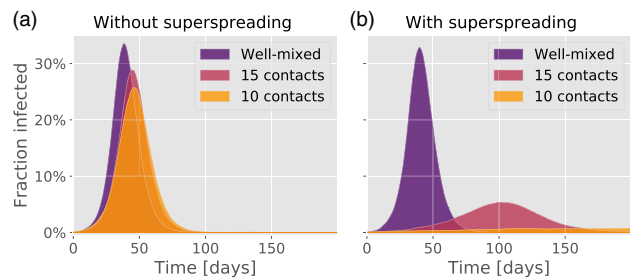


FIG. 3. The epidemic trajectory of a heterogeneous disease is highly sensitive to mitigation. Epidemic trajectories as a function of the number of people that each person interacts with during an infectious period. (a) Time evolution in the absence of any infection heterogeneity. (b) Time evolution for a disease with dispersion parameter $k = 0.1$, roughly representative of COVID-19.

homogeneous disease is largely unaffected by social connectivity [Fig. 3(a)], whereas a heterogeneous epidemic is very sensitive [Fig. 3(b)]. We find a particularly large sensitivity to a reduction of contact number from 15 down to 10 [Fig. 3(b)], indicating a critical threshold for disease spreading, in line with the threshold indicated by the dashed curve in Fig. 2(c).

Crucially, a reduction in contact *time* is not necessary when the disease is characterized by superspreading. What counts is rather a reduction in contact *diversity*, meaning the number of different persons with whom you come into contact during the time you are infectious [7]. This differs fundamentally from SIR models, where contact time and diversity are not differentiated between [28]. In our model, a reduction in the size of an individual's social circle is not accompanied by a reduction in contact time, since the same number of contact events is maintained, with each remaining person being contacted more often. Thus, a mildly infectious individual will not experience appreciable saturation by a reduction in contact diversity, whereas a superspreader will be highly limited by the resultant local saturation.

So far, our analysis has been based on the Erdős-Renyi network, which is largely devoid of clusters. This was chosen as a clean setting in which to probe how social connectivity affects superspreading. However, any realistic social network will involve clusters of people who know each other [29–31]—after all, your colleagues know each other as well as knowing you. It is thus natural to ask whether such *cliquishness* impacts superspreading. In Fig. 4, we compare a cluster-free network to one characterized by a high degree of clustering [32]. See Supplemental Material [19] for the algorithm used to generate this network.

The attack rate of the disease is clearly lowered by clustering, in general (Fig. 4), but the effect is especially significant when heterogeneity is high. The mechanism behind this is that of *local saturation*. If a superspreader

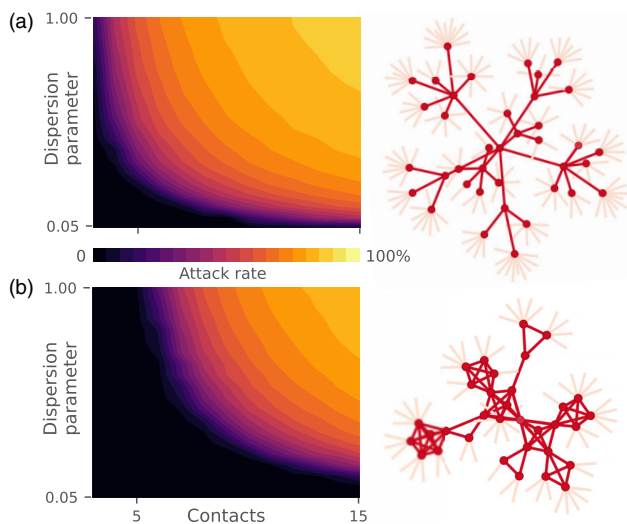


FIG. 4. Final attack rate (total fraction of the population infected) as a function of network connectivity and transmission heterogeneity. In (a), we investigate an Erdős-Renyi network, with the same degree distribution as in Fig. 2. (b) explores a network where each person is assigned to two groups of people, leading to a highly clustered network. The black regions indicate conditions where the disease cannot spread in the population. On the right-hand side, small fragments of the networks in question are shown. Each of the two contour plots in this figure are based on 1500 runs of the model. A detailed description of the algorithm used to generate the clustered network is included in Supplemental Material [19].

infects a significant portion of his network, there is a risk that one of these individuals will turn out to be another superspreader. However, *if* there is clustering, a large part of this *second* superspreader’s network will already have been exposed, and the second superspreader does comparatively little harm.

In the literature, there exists ample evidence that *social* heterogeneity, implemented either through wide distributions of social activity [28] or through social networks with broad degree distributions [23–26], has a significant effect on the course of an epidemic. Notably, epidemics tend to attain lower final sizes in networks with broad degree distributions [23,24,26,33,34] and in clustered networks [26]. While the effects of varying degree distributions as well as clustering were explored in Ref. [26], mitigation by contact network reduction was not investigated, and no variation in individual infectiousness was assumed. As we have shown, the effects of biological superspreading on mitigation are profound in networks with a representative mean. In Supplemental Material [19], we simulate an epidemic on a much more socially heterogeneous (fat-tailed) network based on data from Ref. [27] and find that our conclusions are robust to alterations in the degree distribution.

Beyond the mitigation strategies discussed here, which rely on broad reductions in contact numbers, more targeted

strategies are possible—most prominently, test-trace-isolate (TTI) strategies. While an in-depth treatment of the implications of superspreading for TTI strategies is beyond the scope of this Letter, our simulations do imply that backward contact tracing (see, e.g., [35]) is more effective in the presence of superspreading. When encountering an infected individual, this strategy relies on asking “Who was this person infected by, and who else might *that* person have infected?” rather than simply asking “Who might this person have infected?,” as one would in forward tracing schemes. We can estimate the efficacy of backward tracing in our simulations by measuring how many secondary cases each infected person allows one to trace, with and without superspreading. In a well-mixed scenario, we find the answer to be 2.7 without superspreading ($k = \infty$) and 24 with COVID-like superspreading ($k = 0.1$). Of course, such a backward contact tracing scheme may run into practical limitations, especially regarding the temporal constraints arising from a disease with a relatively short generation time. Nevertheless, these results seem to indicate that transmission heterogeneity may profoundly influence TTI mitigation strategies as well.

Superspreading is now a well-established phenomenon for a number of diseases [15], including COVID-19 [4,5]. In spite of this, the extent to which circumstance and person-specific properties contribute to the observed overdispersion in COVID-19 is still not clear. Superspreading can also have a social component, exemplified by highly social individuals, who come into contact with a large number of people in a limited time frame. However, such individuals would also be *superreceivers*, a trait which impacts the epidemic even in the absence of mitigation [36,37]. In any case, ability as well as opportunity is necessary for superspreading to occur. In our model, we have focused on interindividual variation in ability to produce and transmit virus. This simplification is supported by cases of one person infecting many people at different times and locations [38] and by the observation that most infected people do not even infect their spouse [9–11]. However, more complex models could incorporate realistic *social* heterogeneity as well as large temporal variations in viral load [39,40]—effects which we have not probed. Furthermore, studies which address event-driven superspreading as well as contact tracing in the presence of superspreaders are also needed.

Regardless of the origin of superspreading, we emphasize the particular fragility of a disease in which a major part of infections are caused by the minority. If this is the case, the disease is vulnerable to mitigation by reducing the number of *different* people that an individual meets within an infectious period. The significance is clear: Everybody can still be socially active but generally only with relatively few—on the order of ten persons. Importantly, our study further demonstrates that repeated contact with *interconnected* groups (such as at a workplace or in friend groups) is

comparatively less damaging than repeated contacts with independent people.

We thank Robert J. Taylor, Andreas Eilersen, Gorm G. Jensen, and Julius B. Kirkegaard for enlightening discussions. Our research has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program, Grant Agreement No. 740704, as well as from The Carlsberg Foundation, Grant No. 61114.

*bjarkenielson@nbi.ku.dk

†lonesimo@ruc.dk

‡sneppen@nbi.dk

- [1] L. Hamner, High SARS-CoV-2 attack rate following exposure at a choir practice—Skagit County, Washington, March 2020, *Morb. Mortal. Wkly. Rep.* **69**, 606 (2020).
- [2] S. Kumar, S. Jha, and S. K. Rai, Significance of super spreader events in COVID-19, *Indian Journal of public health : quarterly Journal of the Indian Public Health Association* **64**, 139 (2020).
- [3] Y. Liu, R. M. Eggo, and A. J. Kucharski, Secondary attack rate and superspreading events for SARS-CoV-2, *Lancet* **395**, e47 (2020).
- [4] D. Miller, M. A. Martin, N. Harel, T. Kustin, O. Tirosh, M. Meir, N. Sorek, S. Gefen-Halevi, S. Amit, O. Vorontsov *et al.*, Full genome viral sequences inform patterns of SARS-CoV-2 spread into and within Israel, *Nat Commun.* **11**, 5518 (2020).
- [5] A. Endo, S. Abbott, A. J. Kucharski, S. Funk *et al.*, Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China, *Wellcome Open Res.* **5**, 67 (2020).
- [6] M. S. Lau, B. Grenfell, M. Thomas, M. Bryan, K. Nelson, and B. Lopman, Characterizing superspreading events and age-specific infectiousness of SARS-CoV-2 transmission in Georgia, USA, *Proc. Natl. Acad. Sci. U.S.A.* **117**, 22430 (2020).
- [7] K. Sneppen, R. J. Taylor, and L. Simonsen, Impact of superspreaders on dissemination and mitigation of COVID-19, medRxiv <https://doi.org/10.1101/2020.05.17.20104745> (2020).
- [8] T. R. Frieden and C. T. Lee, Identifying and interrupting superspreading events—implications for control of severe acute respiratory syndrome coronavirus 2, *Emerging Infect. Dis.* **26**, 1059 (2020).
- [9] Q. Bi, Y. Wu, S. Mei, C. Ye, X. Zou, Z. Zhang, X. Liu, L. Wei, S. A. Truelove, T. Zhang *et al.*, Epidemiology and transmission of COVID-19 in 391 cases and 1286 of their close contacts in Shenzhen, China: A retrospective cohort study, *Lancet Infect. Dis.* **20**, 911 (2020).
- [10] S. Y. Park, Y.-M. Kim, S. Yi, S. Lee, B.-J. Na, C. B. Kim, J.-i. Kim, H. S. Kim, Y. B. Kim, Y. Park *et al.*, Coronavirus disease outbreak in call center, South Korea, *Emerging Infect. Dis.* **26**, 1666 (2020).
- [11] Q.-L. Jing, M.-J. Liu, Z.-B. Zhang, L.-Q. Fang, J. Yuan, A.-R. Zhang, N. E. Dean, L. Luo, M.-M. Ma, I. Longini *et al.*, Household secondary attack rate of COVID-19 and associated determinants in Guangzhou, China: A retrospective cohort study, *Lancet Infect. Dis.* **20**, 1141 (2020).
- [12] J. Rabout, A. Shigayeva, A. McGeer, E. Bontovics, M. Chapman, D. Gravel, B. Henry, S. Lapinsky, M. Loeb, L. C. McDonald *et al.*, Risk factors for SARA transmission from patients requiring intubation: A multicentre investigation in Toronto, Canada, *PLoS One* **5**, e10717 (2010).
- [13] Z. Shen, F. Ning, W. Zhou, X. He, C. Lin, D. P. Chin, Z. Zhu, and A. Schuchat, Superspreading SARS events, Beijing, 2003, *Emerging Infect. Dis.* **10**, 256 (2004).
- [14] A. J. Kucharski and C. L. Althaus, The role of superspreading in Middle East respiratory syndrome coronavirus (MERS-CoV) transmission, *Eurosurveillance* **20**, 21167 (2015).
- [15] J. O. Lloyd-Smith, S. J. Schreiber, P. E. Kopp, and W. M. Getz, Superspreading and the effect of individual variation on disease emergence, *Nature (London)* **438**, 355 (2005).
- [16] C. L. Althaus, Ebola superspreading, *Lancet Infect. Dis.* **15**, 507 (2015).
- [17] M. S. Lau, B. D. Dalziel, S. Funk, A. McClelland, A. Tiffany, S. Riley, C. J. E. Metcalf, and B. T. Grenfell, Spatial and temporal dynamics of superspreading events in the 2014–2015 West Africa Ebola epidemic, *Proc. Natl. Acad. Sci. U.S.A.* **114**, 2337 (2017).
- [18] C. Fraser, D. A. Cummings, D. Klinkenberg, D. S. Burke, and N. M. Ferguson, Influenza transmission in households during the 1918 pandemic, *American Journal of Epidemiology* **174**, 505 (2011).
- [19] See Supplemental Material at <http://link.aps.org/supplemental/10.1103/PhysRevLett.126.118301> for model description and sensitivity analyses.
- [20] A. Remuzzi and G. Remuzzi, COVID-19 and Italy: What next?, *Lancet* **395**, 1225 (2020).
- [21] Our World In Data and European Centre for Disease Prevention and Control, COVID-19-Data (Deaths), <https://github.com/owid/covid-19-data/> (2020).
- [22] M. A. Billah, M. M. Miah, and M. N. Khan, Reproductive number of coronavirus: A systematic review and meta-analysis based on global level evidence, *PLoS One* **15**, e0242128 (2020).
- [23] M. Barthélemy, A. Barrat, R. Pastor-Satorras, and A. Vespignani, Dynamical patterns of epidemic outbreaks in complex heterogeneous networks, *J. Theor. Biol.* **235**, 275 (2005).
- [24] J. Hindes and I. B. Schwartz, Epidemic Extinction and Control in Heterogeneous Networks, *Phys. Rev. Lett.* **117**, 028302 (2016).
- [25] S. Bansal, B. T. Grenfell, and L. A. Meyers, When individual behaviour matters: Homogeneous and network models in epidemiology, *J. R. Soc. Interface* **4**, 879 (2007).
- [26] E. M. Volz, J. C. Miller, A. Galvani, and L. A. Meyers, Effects of heterogeneous and clustered contact patterns on infectious disease dynamics, *PLoS Comput. Biol.* **7**, e1002042 (2011).
- [27] L. Danon, J. M. Read, T. A. House, M. C. Vernon, and M. J. Keeling, Social encounter networks: Characterizing Great Britain, *Proc. R. Soc. B* **280**, 20131037 (2013).
- [28] T. Britton, F. Ball, and P. Trapman, A mathematical model reveals the influence of population heterogeneity on herd immunity to SARA-CoV-2, *Science* **369**, 846 (2020).

- [29] D. J. Watts and S. H. Strogatz, Collective dynamics of 'small-world' networks, *Nature (London)* **393**, 440 (1998).
- [30] J. Davidsen, H. Ebel, and S. Bornholdt, Emergence of a Small World from Local Interactions: Modeling Acquaintance Networks, *Phys. Rev. Lett.* **88**, 128701 (2002).
- [31] M. Rosvall and C. T. Bergstrom, Maps of random walks on complex networks reveal community structure, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 1118 (2008).
- [32] A. Eilersen and K. Sneppen, Estimating cost-benefit of quarantine length for COVID-19 mitigation, medRxiv <https://doi.org/10.1101/2020.04.09.20059790> (2020).
- [33] M. E. J. Newman, Spread of epidemic disease on networks, *Phys. Rev. E* **66**, 016128 (2002).
- [34] R. Pastor-Satorras and A. Vespignani, Epidemic Spreading in Scale-Free Networks, *Phys. Rev. Lett.* **86**, 3200 (2001).
- [35] A. Endo *et al.*, Implication of backward contact tracing in the presence of overdispersed transmission in COVID-19 outbreaks, *Wellcome Open Res.* **5** (2020).
- [36] B. F. Nielsen, K. Sneppen, L. Simonsen, and J. Mathiesen, Social network heterogeneity is essential for contact tracing, medRxiv <https://doi.org/10.1101/2020.06.05.20123141> (2020).
- [37] A. V. Tkachenko, S. Maslov, A. Elbanna, G. N. Wong, Z. J. Weiner, and N. Goldenfeld, Persistent heterogeneity not short-term overdispersion determines herd immunity to COVID-19, *arXiv:2008.08142*.
- [38] Y. Shen, C. Li, H. Dong, Z. Wang, L. Martinez, Z. Sun, A. Handel, Z. Chen, E. Chen, M. H. Ebell *et al.*, Community outbreak investigation of SARS-CoV-2 transmission among bus riders in Eastern China, *JAMA Intern. Med.* **180**, 1665 (2020).
- [39] L. Zou, F. Ruan, M. Huang, L. Liang, H. Huang, Z. Hong, J. Yu, M. Kang, Y. Song, J. Xia *et al.*, SARS-CoV-2 viral load in upper respiratory specimens of infected patients, *N. Engl. J. Med.* **382**, 1177 (2020).
- [40] J. B. Kirkegaard, J. Mathiesen, and K. Sneppen, Airborne pathogens in a heterogeneous world: Superspreading & mitigation, medRxiv (to be published).

LOCKDOWNS EXERT SELECTION PRESSURE ON OVERDISPERSION OF SARS-CoV-2 VARIANTS

Authors: Bjarke Frost Nielsen¹, Andreas Eilersen¹, Lone Simonsen² and Kim Sneppen¹.

¹ The Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark.

² Department of Science and Environment, Roskilde University, Roskilde, Denmark.

My contribution: Contributed to conceptualization and development, programming of the model, simulations and creating figures as well as writing of the manuscript.

Publication status: Available as a preprint on the medRxiv website, submitted June 30th, 2021.

Hyperlink(s): <https://doi.org/10.1101/2021.06.30.21259771>

Lockdowns exert selection pressure on overdispersion of SARS-CoV-2 variants

Bjarke Frost Nielsen^a, Andreas Eilersen^a, Lone Simonsen^b, and Kim Sneppen^{a,2}

^aNiels Bohr Institute, University of Copenhagen, Blegdamsvej 17, 2100 Copenhagen, Denmark.; ^bDepartment of Science and Environment, Roskilde University, Universitetsvej 1, 4000 Roskilde, Denmark.

Preprint dated July 8, 2021

The SARS-CoV-2 ancestral strain has caused pronounced superspreading events, reflecting a disease characterized by overdispersion, where about 10% of infected people causes 80% of infections. New variants of the disease have different person-to-person variations in viral load, suggesting for example that the Alpha (B.1.1.7) variant is more infectious but relatively less prone to superspreading. Meanwhile, mitigation of the pandemic has focused on limiting social contacts (lockdowns, regulations on gatherings) and decreasing transmission risk through mask wearing and social distancing. Using a mathematical model, we show that the competitive advantage of disease variants may heavily depend on the restrictions imposed. In particular, we find that lockdowns exert an evolutionary pressure which favours variants with lower levels of overdispersion. We find that overdispersion is an evolutionarily unstable trait, with a tendency for more homogeneously spreading variants to eventually dominate.

Overdispersion | Evolution | Superspreading | Non-pharmaceutical interventions

One of the major features of the coronavirus pandemic has been overdispersion in transmission, manifesting itself as superspreading. There is evidence that around 10% of infected individuals are responsible for 80% of new cases (1–4). This means that some individuals have a high personal reproductive number, while the majority hardly infect at all. A recent study has shown this is reflected in the distribution of viral loads which is extremely wide, with just 2% of SARS-CoV-2 positive individuals carrying 90% of the virus particles circulating in communities (5). Overdispersion is in fact a key characteristic of certain diseases (6–8). However, this is by no means a universal signature of infectious respiratory diseases. Pandemic influenza, for example, is characterized by a much more homogeneous transmission pattern (9–11).

As an emerging virus evolves, its transmission patterns may change and it may become more or less prone to superspreading. The Alpha (B.1.1.7) variant of SARS-CoV-2 has been reported to be ~ 50% more transmissible than the ancestral SARS-CoV-2 virus under varying degrees of lockdown (12–14). Meanwhile, others have shown that the Alpha variant possesses a higher average viral load and a reduced variability between infected persons, compared to the ancestral strain (15, 16). It remains to be seen how this reduced variability affects the transmission patterns of the virus.

The altered viral load distributions seen in persons infected with the Alpha variant have also been investigated at the level of individual mutations. The spike protein of the Alpha variant prominently features the N501Y substitution (asparagine replaced by tyrosine at the 501 position) as well as the Δ H69/V70 deletion (histidine and valine deleted at the 69 and 70 positions). Investigators found that the viral

load is, on average, three times as great for the Alpha variant compared with the ancestral strain (16). Furthermore, viral load distributions in samples taken from persons infected with a variant with the Δ H69/V70 show a lower variance, whether or not they also have tyrosine at the 501 position. However, the difference in variance was most pronounced for those samples which had the deletion as well as the 501Y mutation. Similarly, an analysis of samples with the N501Y mutation show that they have a higher median viral load as well as a substantially diminished variance compared to those without it. Using data from Ref. (15), we calculate that the viral loads in samples of the Alpha variant are associated with a lower coefficient of variation of approximately 2, compared to 4 for the ancestral strain. Importantly, the exact relation between viral load and infectiousness is not well understood; however, a higher viral load is logically expected to increase the risk of disease transmission. By this logic, the decreased variability in the viral load for the Alpha variant may translate into a reduced overdispersion in transmission.

In this paper, we use a mathematical model to study the competition between idealized variants which differ in their level of overdispersion (k) and their mean infectiousness. Our focus is on exploring whether overdispersion confers any evolutionary (dis)advantages, and whether non-pharmaceutical interventions which restrict social network size and transmissibility change the fitness landscape for variants with varying degrees of overdispersion. While it is evident that a higher mean infectiousness confers an evolutionary advantage to an emerging pathogen, it is not *a priori* obvious if a competitive

Significance

One of the most important and complex properties of viral pathogens is their ability to mutate. The SARS-CoV-2 pandemic has been characterized by overdispersion – a propensity for superspreading, which means that around 10% of those who become infected cause 80% of infections. However, evidence is mounting that this is not a stable property of the virus and that the Alpha variant spreads more homogeneously. We use a mathematical model to show that lockdowns exert a selection pressure, driving the pathogen towards more homogeneous transmission. In general, we highlight the importance of understanding how non-pharmaceutical interventions exert evolutionary pressure on pathogens. Our results imply that overdispersion should be taken into account when assessing the transmissibility of emerging variants.

The authors declare no competing interests.

²To whom correspondence should be addressed. E-mail: sneppen@nbi.dk

61 advantage can be gained by specifically altering the *variability*
62 in infectiousness (while keeping transmissibility unchanged).
63 Our recent studies have shown that the presence of overdispersion
64 makes a pandemic far more controllable than influenza
65 pandemics when mitigating by limiting non-repetitive contacts
66 (17) and personal contact network size (18). We therefore spec-
67 ulate that restrictions which alter social contact structure may,
68 conversely, provide a fitness advantage to variants with more
69 homogeneous transmission, and may thus play a role in viral
70 evolution.

71 Across several diseases, individual variations in infectious-
72 ness have been approximated by a Gamma distribution (6)
73 characterized by a certain mean value and a dispersion pa-
74 rameter known as k , which is related to the coefficient of
75 variation (CV) through $CV = 1/\sqrt{k}$. In the simplest of cases
76 (a well-mixed population), infection attempts are modeled as
77 a constant-rate (Poisson) process, which leads to a personal
78 reproductive number which follows a negative binomial distri-
79 bution. The dispersion parameter k characterizes the degree of
80 transmission heterogeneity; a *lower* k corresponds to greater
81 heterogeneity. For small values of k , it approximately corre-
82 sponds to the fraction of infected individuals responsible for
83 80% of new infections. The value for the SARS-CoV-2 ancestral
84 virus is around 10%, corresponding to a k -value of approxi-
85 mately 0.1. Other coronaviruses are also prone to superspread-
86 ing, with the k -values of SARS-CoV-1 and MERS estimated
87 at 0.16 (6) and 0.26 (19), respectively. To explore questions of
88 how such overdispersion affects fitness and pathogen evolution,
89 we use an agent-based model of COVID-19 spreading in a
90 social network, as originally developed in Ref. (18).

91 Overdispersion in personal reproductive number – i.e. sup-
92 erspreading – is a phenomenon that requires *means* (biological
93 infectiousness) as well as *opportunity* (social context). Super-
94 spreading can have diverse origins, ranging from purely be-
95 havioural to biological (8, 20). However, a recent meta-review
96 (21) compared the transmission heterogeneity of influenza
97 A (H1N1), SARS-CoV-1 and SARS-CoV-2 and found that
98 higher variability in respiratory viral load was closely associ-
99 ated with increased transmission heterogeneity. This suggests
100 that biological aspects of individual diseases are decisive in
101 determining the level of overdispersion, and thus the risk of
102 superspreading.

103 Initial survival of variants

104 The words *fitness* and *competitive advantage* may take on
105 several meanings in an evolutionary context. For our purposes,
106 it is especially important to distinguish between the ability
107 of a pathogen to *avoid stochastic extinction* and to *reproduce*
108 *effectively* in a population.

109 To quantify the ability to avoid stochastic extinction we
110 use a branching process to simulate an outbreak of a variant
111 with a given level of overdispersion in a naive population. We
112 then record whether it survives beyond the first 10 genera-
113 tions of infections, as a measure of the ability of that variant to take
114 hold. Repeating these simulations multiple times allows us
115 to compute the survival chance of each variant as a function
116 of its infectiousness and overdispersion, in the absence and
117 presence of mitigation (Fig. 1). Since we are dealing with a
118 few related quantities, some definitions must be made. By
119 the *basic reproductive number* (R_0) we mean the average num-
120 ber of new infections which each infected person gives rise to

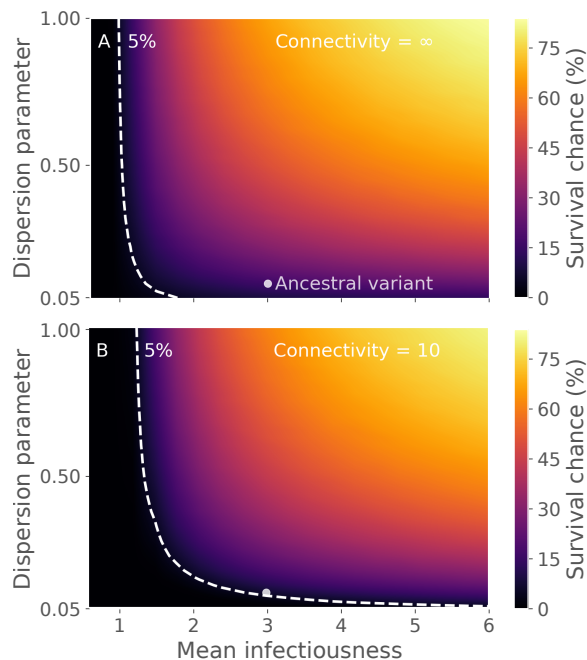


Fig. 1. Initial survival chance depends strongly on overdispersion and moderately on lockdown status. **A)** The epidemic spreads in an unrestricted setting (homogeneous mixing contact structure) **B)** The epidemic spreads in a situation with limited social connectivity (modeled as an Erdos-Renyi network of average connectivity 10). The survival chance is computed by simulating several outbreaks, each starting from a single infected individual in a susceptible population. This initial individual is infected with a variant of a given overdispersion. For each outbreak, the variant is recorded as having *survived* if it does not go extinct within 10 generations. The dashed white line indicated parameters for which the variant has a 5% chance of surviving. The biological mean infectiousness (horizontal axis) has been scaled such that it equals the basic reproductive number (R_0) in the homogeneous mixing scenario of panel A. For details on these calculations, see the Materials and Methods section.

121 *when all contacts are susceptible*. This is in contrast to the
122 effective reproductive number (known variously as R , R_t and
123 R_e), which is affected by population immunity. Note that R_0
124 as well as R_e are context dependent, since behaviour (and
125 mitigation strategies) will affect e.g. the number of contacts
126 that a person has and thus the reproductive number. Another
127 parameter entirely is the (*biological*) *mean infectiousness*, by
128 which we mean the rate at which transmission occurs *when*
129 an infected person is in contact with a susceptible person. This is
130 a property of the disease and not of the social environment. In
131 Fig. 1, the independent variables are thus the mean infectious-
132 ness and the dispersion parameter, both of which are assumed to
133 be properties of the disease. The details of the calculation
134 can be found in the Materials and Methods section.

135 In the unmitigated scenario (Fig. 1A), the procedure is rel-
136 atively straightforward. A single infected individual is initially
137 introduced, with a personal reproductive number z drawn from
138 a negative binomial distribution $P_{NB}[Z; R_0, k]$ with mean value
139 R_0 and dispersion parameter k . Thus, this individual gives
140 rise to z new cases, and the algorithm is reiterated for each of
141 these subsequent infections.

142 In the case of a lockdown scenario, in terms of restrictions
143 of the number of social contacts (Fig. 1B), the algorithm is
144 slightly more involved. In this case, a *degree* c (the number of
145 contacts) is first drawn from a degree distribution (in this case

146 a Poisson distribution, to mimic an Erdős-Renyi network). A
 147 biological reproductive number ξ (the *infectiousness*) is then
 148 drawn from a Gamma distribution with mean value R_0 and
 149 dispersion parameter k . The actual personal reproductive
 150 number z is then drawn from the distribution

$$151 \quad P(z; \xi, c) = \binom{c}{z} (1 - e^{-\xi/c})^z (e^{-\xi/c})^{(c-z)}. \quad [1]$$

152 This reflects that the personal reproductive number z is, natu-
 153 rally enough, limited by the number of distinct social contacts
 154 c . This algorithm is then reiterated for each of the z new
 155 cases.

156 Similar results can be obtained analytically by considering
 157 the probability that an infection chain dies out in infinite
 158 time. Let that probability be d and let $p_i, i \in \{0, 1, \dots\}$ be
 159 the distribution of personal reproductive number (i.e. p_i is
 160 the probability that a single infected individual will infect i
 161 others). Then the extinction risk d is the sum:

$$162 \quad d = p_0 + p_1 d + p_2 d^2 + \dots \quad [2]$$

163 where the first term on the right hand side is the extinction
 164 risk due to the index case producing no new infections, the
 165 second term is the case where the index case gives rise to
 166 one branch of infections which then dies out (this being the
 167 reason for the single factor of d in the second term) and so on.
 168 Since each new branch exists independently of the other, the
 169 extinction events are independent and the probabilities may
 170 be combined by simple multiplication as in Eq. Eq. (2).

171 We find that the survival chance depends very strongly
 172 on overdispersion (Fig. 1), with more homogeneous variants
 173 ($k \sim 1$) having a good chance of survival while highly overdis-
 174 persed variants ($k \leq 0.1$) are very unlikely to survive beyond
 175 10 generations. This finding fits well with the general pat-
 176 tern of overdispersed spreading, namely that many individuals
 177 hardly become infectious while a few pass the disease onto
 178 many others. The uneven distribution of infectiousness makes
 179 heterogeneous diseases more fragile in the early stages of an
 180 epidemic, and thus more prone to stochastic extinction.

181 For the case of homogeneous mixing (Fig 1A) and the num-
 182 ber of generations tending to infinity, Lloyd-Smith et al (6)
 183 performed a similar calculation using the generating function
 184 method described in Eq. 2. For a disease with $R_0 = 3$ and a k
 185 value of 0.16 (similar to what they estimated for SARS-CoV-1),
 186 the survival chance was found to be 24%. Our model yields
 187 the same figure in the unmitigated connectivity $\rightarrow \infty$ limit.

188 To assess the effect of lockdown-like non-pharmaceutical
 189 interventions on the initial survival chances of a pathogen, we
 190 performed an analogous computation in a socially restricted
 191 setting (Fig. 1B). Compared with the unmitigated scenario of
 192 Fig. 1A, it can be seen that the mitigation has an effect on the
 193 survival chance, affecting highly overdispersed variants (small
 194 k) much more than their more homogeneous counterparts
 195 (with the same mean infectiousness). This result is parallel
 196 to the effect of lockdown-like interventions on the *competitive*
 197 *advantage* of a variant, which we explore in the next section.

198 In Ref. (20), the authors study stochastic extinction of
 199 a superspreading disease under a targeted intervention they
 200 call *cutting the tail*. They introduce a cutoff value N_{cutoff}
 201 for the personal reproductive number, and if a person has a
 202 personal reproductive number $z \geq N_{\text{cutoff}}$, a new z is drawn
 203 until one below the threshold is obtained. Since the disease is

highly heterogeneous, this process is analogous to "removing"
 a potential superspreading event and replacing it with a much
 lower personal reproductive number (typically $z = 0$). This is
 exactly why the intervention is rightly called *targeted*. Their
 approach is thus based on viewing superspreading entirely as
 an event-based phenomenon, where one can directly remove
 superspreading events above some threshold size, and instead
 let the individuals take part in other less risky events. Our
 approach, on the other hand, assumes superspreading to be
 due to a combination of high individual biological infectious-
 ness and opportunity, e.g. a large number of social contacts.
 These two viewpoints are complementary in obtaining a com-
 prehensive description of superspreading phenomena, rather
 than mutually exclusive (17).

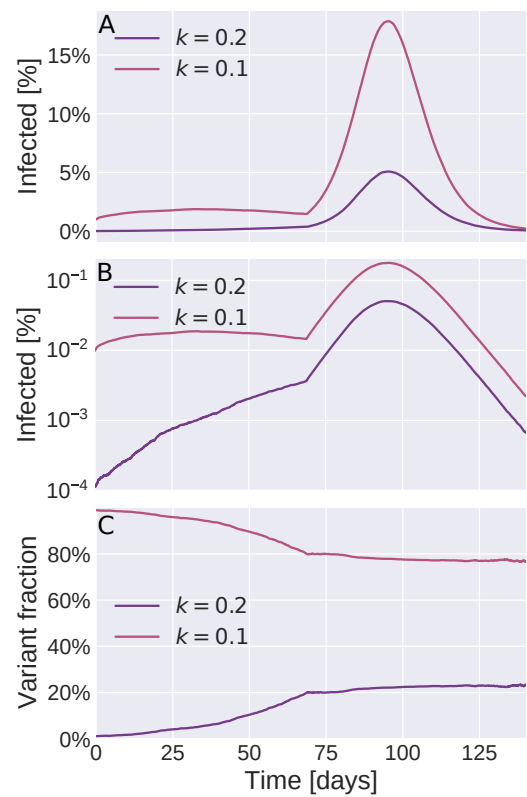


Fig. 2. Simulations of the emergence of a new variant. An initially dominant ("ancestral") strain with dispersion parameter $k = 0.1$ (red) has initially infected 1% of the population. The figure follows the emergence of a new variant (purple), which has the same biological mean infectiousness, but is more homogeneous ($k = 0.2$). Initially, 0.01% of the population is infected with the emerging variant. The two variants exhibit perfect cross-immunity. The initial scenario is a partially locked-down society (modeled as an Erdős-Renyi network with 10 contacts/person). When the new variant reaches 20% of all current infections (around day 65), the lockdown is completely lifted (modeled by a homogeneous mixing contact structure with the same *total* social time available per person). **A**) Incidence of each strain as a function of time since the new variant was introduced. Notice that the new variant spreads approximately exponentially until day 65 (see also panel B), whereas the ancestral strain stays at about 1% incidence. When restrictions are lifted, both surge. **B**) Same data as panel A, but plotted on a logarithmic scale. In this plot, exponential growth shows up as a straight line, and it is thus clear that the new variant spreads approximately exponentially during the lockdown phase. **C**) The relative proportions of the old and new variants. In the locked-down society, the new variant has a distinct fitness advantage, as revealed by its increasing share of infections. Once restrictions are lifted around $t = 65$ days, the fitness advantage is lost and the two variants spread equally well.

Competitive advantage is determined by context

We now turn to the competition between two variants which have already managed to gain a foothold, and so have moved past the initial risk of stochastic extinction. This is a separate aspect of “fitness”, distinct from the initial survival ability described in the last section. Fig. 2 explores the competition between two strains which differ only in their level of overdispersion. The ancestral variant has a broad infectiousness distribution ($k = 0.1$) while the other – the *new variant* – is more narrowly distributed ($k = 0.2$). In the initial partial lockdown scenario, each person is only allowed contact with 10 others. At first, the fraction of infections due to the new variant is observed to grow rapidly. When it reaches a 20% share of active infections, around day 65, the lockdown is lifted (simulated by a shift to a homogeneous mixing contact structure). Naturally, this more permissive contact structure causes a surge in both variants (Fig. 2c). However, the fraction of infections owing to *each* variant suddenly stabilizes, indicating that the more homogeneous new variant has lost its competitive advantage in the unmitigated scenario.

This sudden loss of competitive advantage demonstrates conceptually that the fitness of variants with different patterns of overdispersion depends on context, in the form of non-pharmaceutical interventions or the absence thereof. To quantify this dependence, we separately simulate the spread of several pathogen variants, each with its own specified mean infectiousness and dispersion parameter k , and measure the resulting basic reproductive numbers. In each case we let the pathogen spread in an Erdős-Renyi network with a mean connectivity of either 10 or 50, to simulate scenarios with either a restricted or fairly open society. The results are shown in Fig. 3, where the competitive (dis)advantage of each variant is plotted as a function of its a given biological mean infectiousness and dispersion. The infectiousness is given relative to the SARS-CoV-2 ancestral strain which is set to average infectiousness = 1 and has dispersion $k = 0.1$. This average infectiousness of 1 corresponds to a basic reproduction number of $R_0 = 3$ in a well-mixed scenario, representative of COVID-19 (22). In the socially restricted case with only 10 contacts, the competitive advantage depends strongly on the dispersion parameter, as evidenced by the contour lines in Fig. 3A. The dashed white contour in the figure indicates variants which spread *as well* as the ancestral strain. Concretely, a variant with just half the biological infectiousness of the ancestral strain has no substantial competitive disadvantage, provided it is sufficiently homogeneous ($k \gtrsim 1.0$). In the more socially connected scenario (Fig. 3B), the competitiveness of a strain is observed to depend less strongly on dispersion, and is primarily determined by biological mean infectiousness. Viewed more broadly, these results imply that an observed increase in R_0 for an emerging variant may be due to a *combination* of changes in transmission patterns (k) and biological mean infectiousness

So far, our focus has been on mitigation strategies which rely on reductions in contact network. However, even when societies reopen by allowing contact with an increased number of individuals, non-pharmaceutical interventions which decrease transmission risk per encounter may be in force. These may include face masks and regular testing. In the Supporting Information, we show that interventions which decrease the transmission risk per encounter (i.e. per unit of

contact time) in fact decrease the competitive advantage of more homogeneous variants. These types of interventions thus have essentially the opposite effect, relative to strategies which reduce social connectivity.

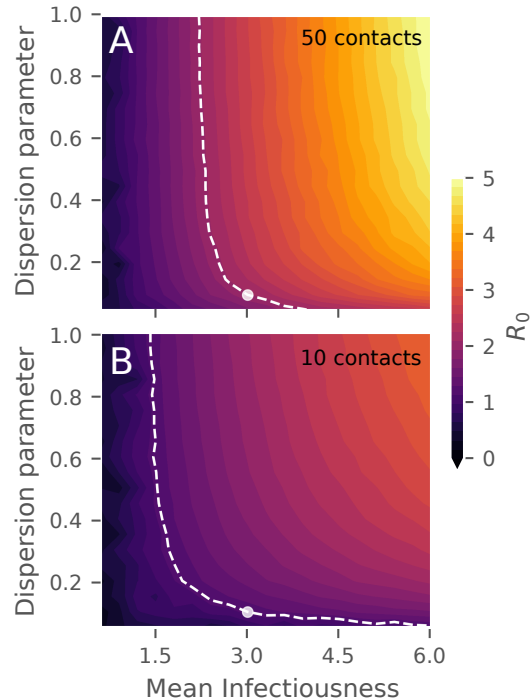


Fig. 3. Relative fitness of variants. The color indicates the basic reproductive number that each variant exhibits under the given circumstances. The dashed white line indicates variants which have the same fitness as the ancestral strain, which is estimated to have $k = 0.1$. The biological mean infectiousness (horizontal axis) has been scaled such that it equals the basic reproductive number (R_0) in a homogeneous mixing scenario. **A)** Spread of the disease in a connectivity 10 Erdős-Renyi network, corresponding to a partial lockdown. **B)** Spread of the disease in a connectivity 50 Erdős-Renyi network, corresponding to a mostly open society.

Interventions exert selection pressure

As the observed differences in the viral load distributions of the Alpha (B.1.1.7.) variant and the ancestral strain suggest, overdispersion is not a fixed property, but rather one that may evolve over time. Furthermore, the SARS-CoV-2 pathogen has been estimated to mutate at a rate of approximately 2 substitutions per genome per month (23), translating to about one mutation per three transmissions. In Fig. 4, we explore the consequences of overdispersion as an evolving feature of the pathogen. In these simulations, the virus has a mutation probability of 1/3 at each transmission. When it mutates, the overdispersion factor is either increased (by a factor of 3/2) or decreased (by a factor of 2/3). Thus, we assume no drift on the microscopic scale, but one may arise macroscopically due to selection pressure from the environment. It should of course be noted that while the assumed mutation rate is realistic for SARS-CoV-2, many mutations will be neutral and only very few mutations will affect transmission dynamics. As such, the present model will likely overestimate the *magnitude* of the drift in overdispersion. It is however conceptually robust – decreasing the mutation rate merely slows down the drift, but the tendency remains.

In our simulations, we find that there is always a tendency

306 for overdispersion to decrease (i.e. for the k value to *increase*),
 307 leading to more homogeneous disease transmission. This makes
 308 sense, since we have already established that heterogeneous
 309 disease variants are more likely to undergo stochastic extinc-
 310 tion (Fig. 1) and that they have a competitive disadvantage
 311 as soon as contact structures are anything but well-mixed
 312 (Fig. 3). In the absence of any interventions, the tendency
 313 to evolve towards homogeneity is quite weak (Fig. 4A), but
 314 when a partial lockdown is instituted, the picture changes
 315 dramatically and the k value increases exponentially. The
 316 conclusion is thus that lockdowns exert a selection pressure on
 317 the virus when it comes to overdispersion, towards developing
 318 a less superspreading-prone phenotype.

319 One may of course object that the scenarios of Fig. 4A (un-
 320 restricted spread) and 4B (partial lockdown) are not directly
 321 comparable, since the epidemic in 4A unfolds much more
 322 rapidly. For this reason, we have included the scenario shown
 323 in 4C, where the transmission rate per encounter has been
 324 lowered, but social structure is unrestricted. The transmission
 325 rate is lowered such that the *initial* daily growth rates in Fig.
 326 4B and 4C are identical (11%/day averaged over the first 14
 327 days). This slightly increases the growth of k over the course of
 328 the epidemic, but to a much lower level than in the lockdown
 329 scenario, demonstrating that it is indeed the restriction of
 330 social network that provides the selection pressure driving k
 331 upwards.

332 Discussion

333 With this paper we have demonstrated that the relative success
 334 and survival of mutants of a superspreading disease depends on
 335 the type of mitigation strategies employed within a population.
 336 The choice of a certain mitigation strategy may well amount to
 337 selecting the next dominant variant. If, for example, a simple
 338 lockdown is enacted while still allowing people to meet within
 339 restricted social groups, the evolution of more homogeneously
 340 spreading disease variants may become favoured.

341 The spreading of an emerging virus in a human society is
 342 a complex phenomenon, where the actual reproductive number
 343 depends on sociocultural factors, mitigation policies and
 344 self-imposed changes in the behaviour of citizens as awareness
 345 grows in the population. The spread of a disease such as
 346 COVID-19 cannot simply be characterized by a single fitness
 347 quantity like the basic reproductive number R_0 , but will also
 348 depend on the heterogeneities of transmission patterns within
 349 the population. If schools are open, mutants which spread
 350 more easily among children may be selected for, whereas rapid
 351 self-isolation of infected individuals may tend to favor vari-
 352 ants which temporally separate disease transmission from the
 353 development of symptoms. We have focused on modeling the
 354 evolutionary effects of biological superspreading in the context
 355 of mitigations such as lockdowns which have been implemented
 356 globally during the COVID-19 pandemic. We found that such
 357 lockdowns will favour the emergence of homogeneously spread-
 358 ing variants over time.

359 Our findings also have implications for the assessment of
 360 new variants. They highlight the importance of taking overdis-
 361 persion into account when evaluating the transmissibility of an
 362 emerging variant. We have shown that the disease can spread
 363 more effectively not only by increasing its biological mean
 364 infectiousness, but also by changing its pattern of transmission
 365 to become more homogeneous. Practically, this means that

transmission data obtained under even partial lockdown can
 lead to an overestimation of the transmissibility of an emerging
 variant. We thus call for an increased focus on measuring the
 overdispersion of variants, as this may be critical for estimat-
 ing the reproductive number of new variants. These estimates
 in turn determine the required vaccination levels to reach herd
 immunity.

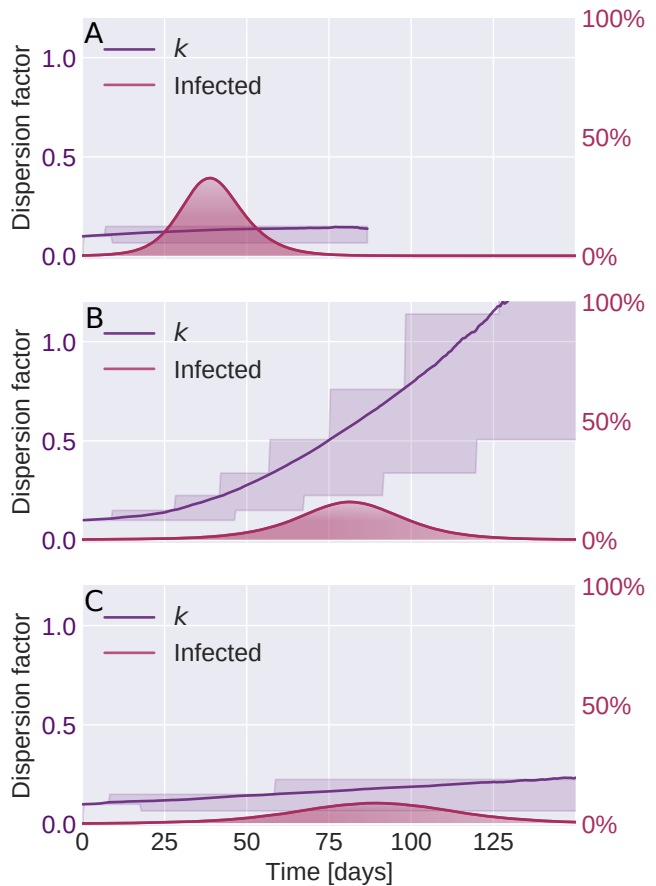


Fig. 4. Evolution of overdispersion is driven by imposed restrictions. In these simulations, random mutations occur which alter the level of transmission overdispersion in a non-directed fashion. However, external evolutionary pressures are seen to drive the disease towards developing more homogeneous spreading patterns. The filled red curve shows the combined incidence of all strains. The purple curve shows the average dispersion factor k in the infected population (with higher k corresponding to a more homogeneous infectiousness). The shaded purple area shows the 25% and 75% percentiles of the distribution of dispersion factors in the infected population. **A)** The pathogen evolves in an open society with no restrictions imposed (homogeneous mixing contact structure). **B)** Partial lockdown, with an average social network connectivity restricted to 15 persons. **C)** No restrictions on social network, but infectiousness lowered by other means (e.g. face masks).

Materials and Methods

We use an individual-based (or agent-based) network model of disease transmission as originally developed in Ref. (18). In this section, we present only a brief overview of the basic model, and refer to Ref. (18) for a more detailed description. We then go on to describe in detail the simulations and calculations which are particular to this manuscript.

The disease progression model consists of four overall states, Susceptible, Exposed, Infected and Recovered. The exposed state has an average duration of 2.4 days and is subdivided into two consecutive states with exponentially distributed waiting times (i.e.

having constant probability rate for leaving the state) of 1.2 days each, thus constituting a gamma distributed state when viewed as a whole. The infectious state is divided into two states as well, of 1.2 and 5 days in duration, respectively.

Each individual in the model is associated with a fixed social network. Only a subset of edges are activated in each timestep, to simulate a contact event. In the simulations of this work, we always use either an Erdős-Renyi network with finite mean connectivity, or a homogeneous-mixing contact structure, which is also obtainable as the infinite connectivity limit of an Erdős-Renyi network.

When an edge connecting a susceptible and an infectious individual is active, there is a certain probability per unit of time for disease transmission to occur. This rate is determined by the individual infectiousness r_i of the infectious agent, which is drawn from a gamma distribution with dispersion parameter k before the individual has become infectious. As such, the infectiousness for any given individual is assumed constant throughout the infectious stage of the disease. The infectiousness distribution determines an upper bound on size Δt of the timesteps in the model, since the inequality $r_i \cdot \Delta t < 1$ must hold for all agents. A timestep of size $\Delta t = 30\text{min}$ was used throughout, since this was sufficient to ensure that the inequality was satisfied.

Below we go into more detail as to how the simulations involving multiple strains were performed.

Stochastic extinction. The stochastic extinction (or, conversely, survival) plots of Figure 1 in the main text rely entirely on a branching process algorithm with sampling of probability distributions with an analytic description. In practice, we have performed the computation by numerical sampling.

In each generation of the epidemic, the computation is reiterated. Without loss of generality, we therefore here describe a single generation which initially has I infected individuals. Note that for the initial generation, $I = 1$ infected individuals.

- For $i \in \{1, \dots, I\}$:
 - Draw individual infectiousness ξ_i from Gamma distribution $P_\xi(\xi; k, \mu)$
 - Draw number of contacts c from a Poisson distribution with a given mean connectivity.
 - Given number of contacts c , draw personal reproductive number z_i from the distribution Eq. (3)

$$P_z(z; \xi, c) = \binom{c}{z} (1 - e^{-\xi/c})^z (e^{-\xi/c})^{(c-z)}. \quad [3]$$

- Let the number of newly infected be $I = \sum_i z_i$ and repeat the algorithm with this new value of I .

If the number of infected I ever drops to zero, the outbreak is said to have undergone stochastic extinction in that generation. By performing multiple such branching process simulations for each value of the parameters μ (mean infectiousness) and k (dispersion factor) we build up a statistic of the survival chance of each specific variant. To generate Figure 1, this is repeated for two different values of the mean connectivity c .

Two-strain competition simulations. In Fig. 2, two strains spread simultaneously in the population of $N = 10^6$ individuals. Initially, 0.99% of the population are infected with the heterogeneous "old" variant ($k = 0.1$), while 0.01% are infected with the more homogeneous "new" variant ($k = 0.2$). Once a person with a given variant infects a susceptible individual, the characteristics of the variant are passed on to the newly infected individual, such that the infectiousness of this person is drawn from a Gamma distribution with dispersion parameter k set by the variant. In other words, these simulations assume that no further mutations affecting overdispersion occur, allowing us to track solely the competition of two differently-dispersed variants within a population.

Evolutionary model. In Fig. 4, we allow the pathogen to stochastically mutate upon transmission, with the mutations affecting the degree of overdispersion. In the simulations, the pathogen mutates on average once for each new host it is transmitted to (i.e. with

mutation probability $p = 1/3$) and the mutations are assumed to always affect overdispersion, by either increasing the k value by a factor of $3/2$ (i.e. $k \rightarrow 3k/2$) or decreasing it by a factor of $2/3$ (i.e. $k \rightarrow 2k/3$). On a microscopic level, the dispersion level thus performs an unbiased (multiplicative) random walk. The value of this step-size parameter is arbitrarily chosen, and as such the simulations can only be regarded as qualitative and conceptual. However, although no intrinsic bias is built into the mutation mechanism, external selection pressures may drive the level of overdispersion in the population up or down, as is explored in Fig. 4.

In Fig. 4C, the average infectiousness of the strain is lowered so as to produce an initial growth rate that is identical to that of 4A, namely 11% per day in the first 14 days of the epidemic.

ACKNOWLEDGMENTS. We thank Robert J. Taylor, and Julius B. Kirkegaard for enlightening discussions. Our research has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme, grant agreement No. 740704, as well as from the Carlsberg Foundation under its Semper Ardens programme (grant # CF20-0046).

1. D Miller, et al., Full genome viral sequences inform patterns of SARS-CoV-2 spread into and within Israel. *Nat. communications* **11**, 1–10 (2020).
2. A Endo, S Abbott, AJ Kucharski, S Funk, et al., Estimating the overdispersion in covid-19 transmission using outbreak sizes outside china. *Wellcome Open Res.* **5**, 67 (2020).
3. C Pozderac, B Skinner, Superspreading of sars-cov-2 in the usa. *Plos one* **16**, e0248808 (2021).
4. JB Kirkegaard, K Sneppen, Variability of individual infectiousness derived from aggregate statistics of covid-19. *medRxiv* **0** (2021).
5. Q Yang, et al., Just 2% of sars-cov-2-positive individuals carry 90% of the virus circulating in communities. *Proc. Natl. Acad. Sci.* **118** (2021).
6. JO Lloyd-Smith, SJ Schreiber, PE Kopp, WM Getz, Superspreading and the effect of individual variation on disease emergence. *Nature* **438**, 355–359 (2005).
7. AP Galvani, RM May, Dimensions of superspreading. *Nature* **438**, 293–295 (2005).
8. ME Woolhouse, et al., Heterogeneities in the transmission of infectious agents: implications for the design of control programs. *Proc. Natl. Acad. Sci.* **94**, 338–342 (1997).
9. C Fraser, DA Cummings, D Klinkenberg, DS Burke, NM Ferguson, Influenza transmission in households during the 1918 pandemic. *Am. journal epidemiology* **174**, 505–514 (2011).
10. J Brugger, CL Althaus, Transmission of and susceptibility to seasonal influenza in switzerland from 2003 to 2015. *Epidemics* **30**, 100373 (2020).
11. MG Roberts, H Nishiura, Early estimation of the reproduction number in the presence of imported cases: pandemic influenza h1n1-2009 in new zealand. *PLoS one* **6**, e17835 (2011).
12. MS Graham, et al., Changes in symptomatology, reinfection, and transmissibility associated with the sars-cov-2 variant b. 1.1. 7: an ecological study. *The Lancet Public Heal.* **6**, e335–e345 (2021).
13. E Volz, et al., Assessing transmissibility of sars-cov-2 lineage b. 1.1. 7 in england. *Nature* **593**, 266–269 (2021).
14. NG Davies, et al., Estimated transmissibility and impact of sars-cov-2 lineage b. 1.1. 7 in england. *Science* **372** (2021).
15. M Kidd, et al., S-variant SARS-CoV-2 lineage B.1.1.7 is associated with significantly higher viral loads in samples tested by ThermoFisher TaqPath RT-qPCR. *The J. infectious diseases* **223** (2021).
16. T Golubchik, et al., Early analysis of a potential link between viral load and the N501Y mutation in the SARS-COV-2 spike protein. *medRxiv* **0** (2021).
17. K Sneppen, BF Nielsen, RJ Taylor, L Simonsen, Overdispersion in COVID-19 increases the effectiveness of limiting nonrepetitive contacts for transmission control. *Proc. Natl. Acad. Sci.* **118** (2021).
18. BF Nielsen, L Simonsen, K Sneppen, COVID-19 superspreading suggests mitigation by social network modulation. *Phys. Rev. Lett.* **126**, 118301 (2021).
19. A Kucharski, CL Althaus, The role of superspreading in middle east respiratory syndrome coronavirus (mers-cov) transmission. *Eurosurveillance* **20**, 21167 (2015).
20. BM Althouse, et al., Superspreading events in the transmission dynamics of sars-cov-2: Opportunities for interventions and control. *PLoS biology* **18**, e3000897 (2020).
21. PZ Chen, et al., Heterogeneity in transmissibility and shedding sars-cov-2 via droplets and aerosols. *Elife* **10**, e65774 (2021).
22. A Billah, M Miah, N Khan, Reproductive number of coronavirus: A systematic review and meta-analysis based on global level evidence. *PLOS ONE* **15**, 1–17 (2020).
23. M Worobey, et al., The emergence of sars-cov-2 in europe and north america. *Science* **370**, 564–570 (2020).

DIFFERENCES IN SOCIAL ACTIVITY INCREASE EFFICIENCY OF CONTACT TRACING

Authors: Bjarke Frost Nielsen¹, Kim Sneppen¹, Lone Simonsen² and Joachim Mathiesen¹.

¹ The Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark.

² Department of Science and Environment, Roskilde University, Roskilde, Denmark.

My contribution: Contributed to conceptualization and development, programming of the model, performing simulations and creating figures as well as writing of the manuscript.

Publication status: Available as a preprint on the medRxiv website under the title “Social network heterogeneity is essential for contact tracing”, submitted June 5th, 2020.

Hyperlink(s): <https://doi.org/10.1101/2020.06.05.20123141>

Differences in social activity increase efficiency of contact tracing

Bjarke Frost Nielsen,¹ Kim Sneppen,¹ Lone Simonsen,² and Joachim Mathiesen^{1,*}

¹*Niels Bohr Institute, University of Copenhagen, Blegdamsvej 17, 2100 Copenhagen.*

²*Department of Science and Environment,
Roskilde University, 4000 Roskilde, Denmark.*

(Dated: August 10, 2021)

Digital contact tracing has been suggested as an effective strategy for controlling an epidemic without severely limiting personal mobility. Here, we use smartphone proximity data to explore how social structure affects contact tracing of COVID-19. We model the spread of COVID-19 and find that the effectiveness of contact tracing depends strongly on social network structure and heterogeneous social activity. Contact tracing is shown to be remarkably effective in a workplace environment and the effectiveness depends strongly on the minimum duration of contact required to initiate quarantine. In a realistic social network, we find that forward contact tracing with immediate isolation can reduce an epidemic by more than 70%. In perspective, our findings highlight the necessity of incorporating social heterogeneity into models of mitigation strategies.

I. INTRODUCTION

For diseases which are primarily transmitted in spatial proximity, contact patterns invariably play a central role in the course of an epidemic [1, 2]. For the purposes of modeling infectious diseases, contact patterns can be represented by a network where each individual is a node and spatial proximity between individuals is represented by time-dependent edges. Nonetheless, well-mixed compartmental models remain the typical approach to modeling epidemics [3–6]. Even such models, which do not incorporate a network structure, make assumptions about the underlying social contact patterns. In well-mixed models, the assumption is that mixing patterns are homogeneous *inside* sub-populations [7–13]. Although interaction rates *between* sub-populations can be adjusted, well-mixed models may fail to predict the evolution of an epidemic when social interactions are spatiotemporally restricted [14–16], as in real contact networks.

An oft-taken approach to modelling of contact tracing schemes is branching process simulation [17–19]. In these models, the outbreak is modeled generation by generation and the susceptible population is usually taken to be constant in size, rendering the models most useful for studying

* To whom correspondence should be addressed, mathies@nbi.ku.dk

21 early outbreaks. Such models have clear advantages in terms of mathematical tractability, but
22 lack the (disease and social) dynamics which is the main focus of this study. Social interactions
23 tend to follow a characteristic pattern of spatiotemporal correlation, where you meet the same
24 people at specific times during a week, at work or at home. At the same time, social activity varies
25 significantly from person to person. This correlation increases transmission heterogeneity, i.e. the
26 tendency of cases to occur in clusters.

27 During the COVID-19 pandemic, contact tracing has been the center of much attention due
28 to its promises of epidemic control without severely restricting mobility [20–25]. As a mitigation
29 strategy, contact tracing relies directly on the contact network structure and may benefit from
30 clustering of cases [26]. In order to assess contact tracing strategies, detailed information on
31 contact networks is indispensable, and the usual well-mixed approach is inadequate – more so than
32 when modelling unmitigated spreading [27, 28].

33 In this paper, we utilize Bluetooth proximity data obtained from a cohort of university students
34 at a large European university (see *Methods* for details). In most studies of this nature, mobility
35 data collected from mobile phones rely on spatial locations derived from estimated distances to cell
36 towers, GPS coordinates [29] or the proximity to known Wi-Fi access points. Whereas this kind
37 of data is useful for studies of aggregate mobility [30], the accuracy is typically not sufficient to
38 infer epidemiologically relevant social proximity between individuals. In contrast, the Bluetooth
39 data that we consider here can identify social proximity with a high spatial resolution ($<1\text{m}$). In
40 addition, our data has a high temporal resolution ($<5\text{mins}$), meaning that brief encounters (indi-
41 viduals passing by each other) can be distinguished from longer meetings. A high spatiotemporal
42 resolution is necessary to faithfully simulate disease transmission through a social network, since
43 diseases (such as COVID-19) may be less likely to transmit during short encounters or between
44 individuals separated by more than a few meters [31, 32]. The upper limit for the range of our
45 Bluetooth data is approximately 15 meters [33]. We also note that our data are similar in nature
46 to those collected by contact tracing smartphone applications [34].

47 Like all real-life proximity data, the data set used in this study comprises just a section of the
48 complete contact network of each participant. However, our data still display a well-defined and
49 robust heterogeneity which is the object of our study. We further note that contact heterogeneity
50 is pronounced despite the fact that our participant group is homogeneous in age and occupation,
51 and would be treated as undifferentiated in typical epidemiological models.

52 To study the effects of contact heterogeneity on an epidemic, we simulate the propagation of
53 COVID-19 on the empirical contact network, and compare with artificially homogenized versions

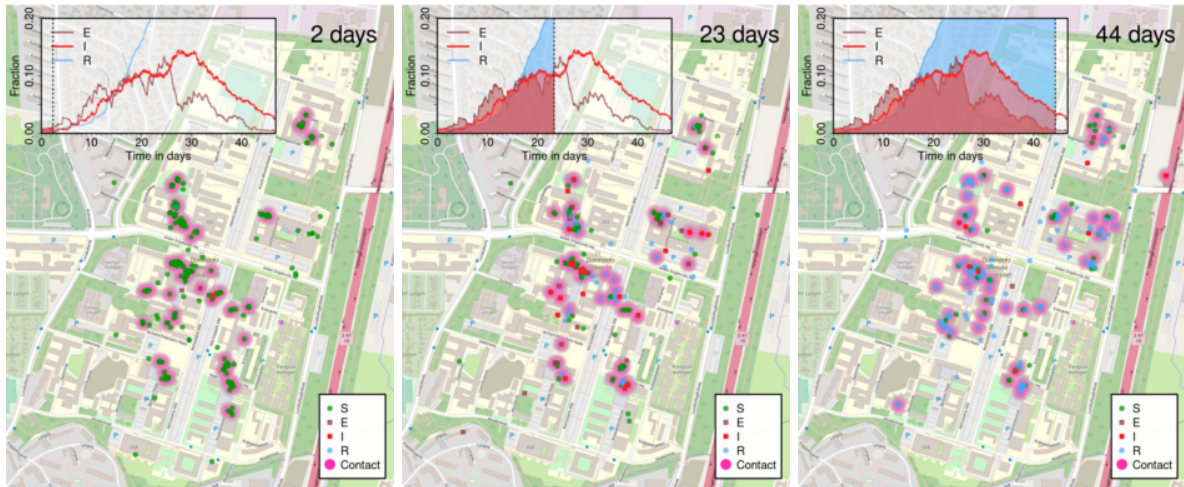


FIG. 1. Simulating the spread of COVID-19 on the contact network. Here, a zoom view on the geographical positions of a few individuals (based on GPS coordinates) during a typical work day and for a representative run of the epidemic model. Regions of contact (defined by signal strength exceeding the -85dBm cutoff) are shown as diffuse clouds of pink. Snapshots shown are at day 2, 23 and 44 of the outbreak.

54 of this network. This allows us to maintain certain features of the network (size, average contact
 55 rate) while altering others (network structure and degree distribution). We can then separately
 56 study how these features affect the outcomes of the epidemic, with and without mitigation. For
 57 that purpose, we introduce three degrees of heterogeneity: **i)** the true (observed) network. **ii)**
 58 an edge swapped version of the network [35], which retains heterogeneity in activity levels but
 59 homogenizes the network structure and edge correlations, and **iii)** a randomized network, which
 60 retains only the overall (mean) contact frequency, but eliminates heterogeneity.

61 Our main question is if contact tracing of COVID-19 is affected by the variation in individual
 62 social activity levels and by the structure of the social network itself. Our contact tracing algorithm
 63 has two key parameters, the probability for a symptomatic individual to undergo testing and the
 64 maximum duration of social proximity to an exposed individual allowed, before a self-quarantine
 65 is triggered. The latter is especially useful, since it is a directly controllable parameter when e.g.
 66 designing contact tracing smartphone applications [34].

68

II. METHODS

69 We use temporally resolved social proximity data collected using smartphones distributed to
 70 1000 participants (undergraduate students at the Technical University of Denmark [36, 37]). The
 71 smartphones were equipped with an application that collected days communication in the form of call

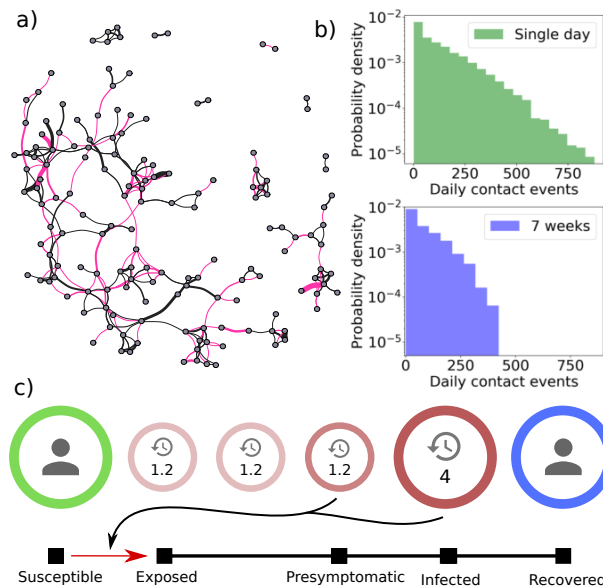


FIG. 2. **a)** A small subset of a contact network for one week. Link thickness indicates the cumulative contact time, with links with less than 2 hours cumulative activity being omitted. Black lines represent the links recurring from the previous week, whereas the red lines are new links. **b)** Top: Histogram of contact events over a single day (semi-logarithmic plot). The coefficient of variation is $c_V = 1.03$ and the mean is $\mu = 131$. Bottom: Histogram of contact events over a seven week period, divided by the number of days to obtain an average daily rate (semi-logarithmic plot). Here, $c_V = 0.95$ and $\mu = 86$. Both plots show a marked heterogeneity, demonstrating that contact heterogeneity is approximately a quenched disorder on the timescale of a few weeks. **c)** Our agent-based model of COVID-19 spreading on a contact network. Individuals in the **S**usceptible state may be exposed by those in the **P**resymptomatic as well as **I**nfected states. The **E**xposed-**P**resymptomatic triplet of states together comprise the gamma-distributed incubation period.

72 and text messaging logs, geo-location (GPS coordinates) and social proximity data using the Blue-
 73 tooth port. Every 5 minutes, all smartphones in the study scanned for nearby devices included in
 74 the study, and recorded Bluetooth signal strength as well as the GPS coordinates of the phone.
 75 The data we consider were collected over a period of two years, 2013-2015.

76 The approximate distance between participants can be inferred from the strength (RSSI) of
 77 the Bluetooth signal transmitted between devices. The signal strength can resolve distances in
 78 the range of ≤ 1 meter to approximately 10-15 meters [33]. To prepare our data for modeling of
 79 disease transmission, the collected RSSI values are related to an epidemiologically relevant notion
 80 of *contact*. The definition of a *contact* depends on the disease in question and its dominant mode(s)
 81 of transmission. If environmental transmission is significant, a simple short-distance cutoff would

82 be incorrect, and simple proximity data would be insufficient. However, for SARS-CoV-2, there
 83 is evidence that transmission by fomite is minor [38]. Our transmission model assumes that the
 84 transmission risk of COVID-19 increases sharply as interpersonal distance is decreased below 1-2m
 85 [32, 39–42]. Thus, we define two individuals to be in social contact whenever the Bluetooth signal
 86 strength between their respective devices exceeds -85dBm . This definition of contact captures
 87 essentially all $\leq 1\text{m}$ interactions while excluding a large portion of the 3m interactions and above
 88 [33].

89 From the social contacts, we can create a well-defined time-dependent contact network where
 90 individuals are represented by nodes and social contact by time-dependent links, similar in nature
 91 to the network used in [43]. The link activity, i.e. the contact between individuals, is resolved in
 92 temporal windows of 5 minutes. This time-dependent contact network is the basis for our modeling
 93 of the transmission of COVID-19.

94 We model the spread of COVID-19 by an agent-based model (where the study participants
 95 serve as the agents) with five states: **S**usceptible to the disease, **E**xposed, **P**re-symptomatic (but
 96 infectious), **I**nfected (possibly with symptoms) and **R**ecovered/Removed. In the absence of contact
 97 tracing (described below), the P and I states are identical, in that an individual in one of these
 98 states can infect others. Aside from these mutually exclusive states, persons can also be flagged as
 99 **Q**uarantined. In Fig. 1 an example trajectory is shown, together with a closeup of the university
 100 campus. The disease progression model is illustrated in Fig. 2. The transmission routine works by
 101 assuming a constant pairwise infection rate between individuals, when they are in contact. When
 102 a susceptible person comes into contact with a person in the **I** or **P** state, there is a probability p_{inf}
 103 of transmission of the disease in each 5-minute window. The basic model (without contact tracing)
 104 thus has four parameters: The transmission probability upon contact p_{inf} , and three time-scales
 105 characterizing the exposed, presymptomatic and infected states, τ_E , τ_P and τ_I .

106 As shown in Fig. 2, we assume the incubation time to be gamma-distributed with a mean of
 107 3.6 days, of which the last 1.2 days comprise the presymptomatic infectious state. The infectious
 108 state, where symptoms may be displayed, is set at four days. The last remaining parameter of the
 109 disease model, the transmission probability in each window of time, is fitted to reproduce a daily
 110 growth rate of 23% in the early epidemic, based on estimates from [44, 45]. This gives a basic
 111 reproductive number of $R_0 = 2.8$ when simulated on the empirical social network. Note that this is
 112 the pre-mitigation value, which fits well with the reproductive number obtained in a recent review
 113 [46].

114 By employing two different ways of *shuffling* the network connections (edges), we study both

115 the effects of heterogeneity in activity levels (social contact time) and in the network structure.
 116 The first method, *edge swapping*, preserves the degree of connectivity of each person (node), while
 117 destroying any network structure arising from e.g. group formation and spatial preferences [35].
 118 The second method, *randomization*, preserves only the overall connectivity level in each window
 119 of time, but homogenizes the number of contacts for each person.

120 The edge swapping procedure works as follows. Given a contact network at an instant of time
 121 (representing, in our case, a 5 minute time window), we iterate the following steps:

- 122 • Select two edges at random. Denote the pairs of connected nodes $A \leftrightarrow B$ and $C \leftrightarrow D$, respec-
 123 tively.
- 124 • Swap the chosen edges such that the connected pairs are now $A \leftrightarrow C$ and $B \leftrightarrow D$.

125 This is repeated until each edge in the system has been swapped several times, on average. Since
 126 no node loses or gains an edge by this procedure, the degree distribution is unchanged. Thus,
 127 the heterogeneity in social activity levels is preserved as well. However, since edge swapping is
 128 performed independently during each time step, the durations of contacts are not preserved. A
 129 10 minute contact is thus treated as two 5-minute contacts and each undergoes swapping inde-
 130 pendently. In the supplemental material, we describe a duration-preserving variation on the edge
 131 swapping algorithm.

132 The randomization procedure is simpler, and each iteration proceeds as follows:

- 133 • Select an edge at random. Rewire the edge by replacing its endpoints with two nodes, chosen
 134 at random from the entire system.

135 As with the edge swapping procedure, this is repeated until each edge of the network has been
 136 swapped several times, on average. Since edges are only rewired, and not created or destroyed, the
 137 overall connectivity of the network is preserved.

138

Contact tracing

139 The contact tracing scheme consists of two parts: *regular testing* of symptomatic individuals
 140 (with a constant rate of testing r_{test}) and the contact tracing algorithm itself, which is activated
 141 once an individual tests positive. Once a positive individual is found by regular testing, their
 142 recent contacts are put in quarantine for a specified time and tested once the quarantine period
 143 has elapsed (before potential release). In other words, the contact tracing scheme proceeds as
 144 follows:

- 145 • For each individual, a list of contact events is kept. When a person (the ‘index case’) is
146 tested positive, all contacts older than 5 days (the *retention time*) are discarded, the index
147 case is quarantined for 5 days.
- 148 • If a traced individual has been in contact with the index case for longer than a certain
149 cumulative *contact threshold*, the traced individual is also quarantined for 5 days.
- 150 • After the quarantine period has elapsed, the individual is tested. If negative, the individual
151 is released. Otherwise a new 5-day quarantine is issued.

152 The quarantine is assumed to be instantaneous and a quarantined person is assumed to have no
153 contact with others. We assume that regular testing happens at a constant rate when an individual
154 is in the symptomatic infected state. This rate of testing r_{test} is measured in units of $1/\tau_I$, the
155 rate at which an individual leaves the infected state. Thus a rate of testing of e.g. 1 corresponds
156 to a 50% chance of being tested while infected. Note that the simple algorithm used here is
157 non-recursive. This choice was made to simplify the analysis, i.e. to facilitate the comparison of
158 contact tracing in networks with different types of heterogeneity. For an exploration of the impact
159 of recursive vs. standard contact tracing, we refer to Refs. [47, 48].

160 The minimum quarantine time is set at 5 days in our simulations, as suggested by [25], but we
161 have performed a sensitivity analysis (see Supplemental Material) which shows that, while there
162 is still some benefit, the marginal effect of increasing the quarantine time decreases above five
163 days. We also performed a sensitivity analysis for the retention time, i.e. the maximum age of
164 contact events deemed relevant when performing contact tracing. It is clear that including contacts
165 which occurred long ago will lead to many unnecessary quarantines, but also that it may increase
166 epidemic control. Our sensitivity analysis shows that the total time spent in quarantine depends
167 only weakly on the retention time, but indicates that 5 days is a reasonable trade-off. See the
168 Supplemental Material for details.

169 III. RESULTS

170 The distribution of the number of daily contact events for each person in the study is found
171 to closely follow an exponential distribution (Fig. 2b), with a coefficient of variation of 1.03 and
172 a mean of 131. This reflects a marked heterogeneity in activity levels. When we consider the
173 distribution over a 7-week window, a significant degree of contact heterogeneity is retained, albeit
174 with some attenuation. Here the coefficient of variation is 0.95, still close to the value for an

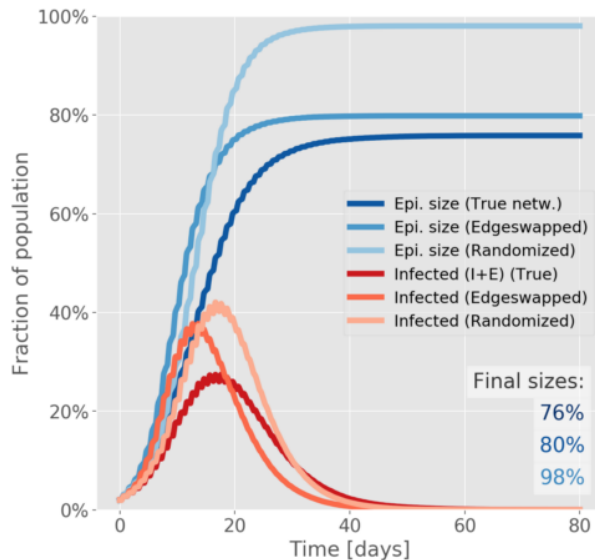


FIG. 3. **The effects of social heterogeneity on an unmitigated epidemic.** The red curves show the incidence, measured as the sum of exposed and infectious individuals (whether symptomatic or not). The blue curves indicate the attack rate, i.e. the cumulative fraction of the population who have been exposed to the disease. In both cases, the curves correspond to the *true*, *edge swapped* and *randomized* networks, in order of increasing brightness. Each trajectory represents an average of 50 simulations.

175 exponential distribution, and the mean is 86. It is clear that extreme social behaviour becomes less
 176 frequent over the longer time-window, reflecting that individuals do not participate in larger social
 177 events every single day. The mean value of 86 corresponds to individuals being socially inactive on
 178 34% of workdays.

179 **Social structure reduces the epidemic severity.** To assess the impact of social hetero-
 180 geneity on an unmitigated epidemic, we compare the simulated evolution of COVID-19 on three
 181 different contact networks (Fig. 3): The true (unshuffled) network, the edge swapped and the fully
 182 randomized network where each person is assigned an average contact frequency. Each trajectory
 183 is averaged over 50 runs, each similar in nature to the one shown in the inserts of Fig. 1.

184

185 We find that the final size of the epidemic (the total number of exposed individuals) is very
 186 sensitive to heterogeneity in social activity, but not to the network structure. Heterogeneity in
 187 social activity prevents the disease from spreading to all parts of the network, with the total
 188 fraction exposed reaching 76% in the true network and 98% in the randomized network. The edge
 189 swapped network, on the other hand, results in an epidemic size similar to the true network, despite
 190 the homogenization of social network structure caused by this procedure.

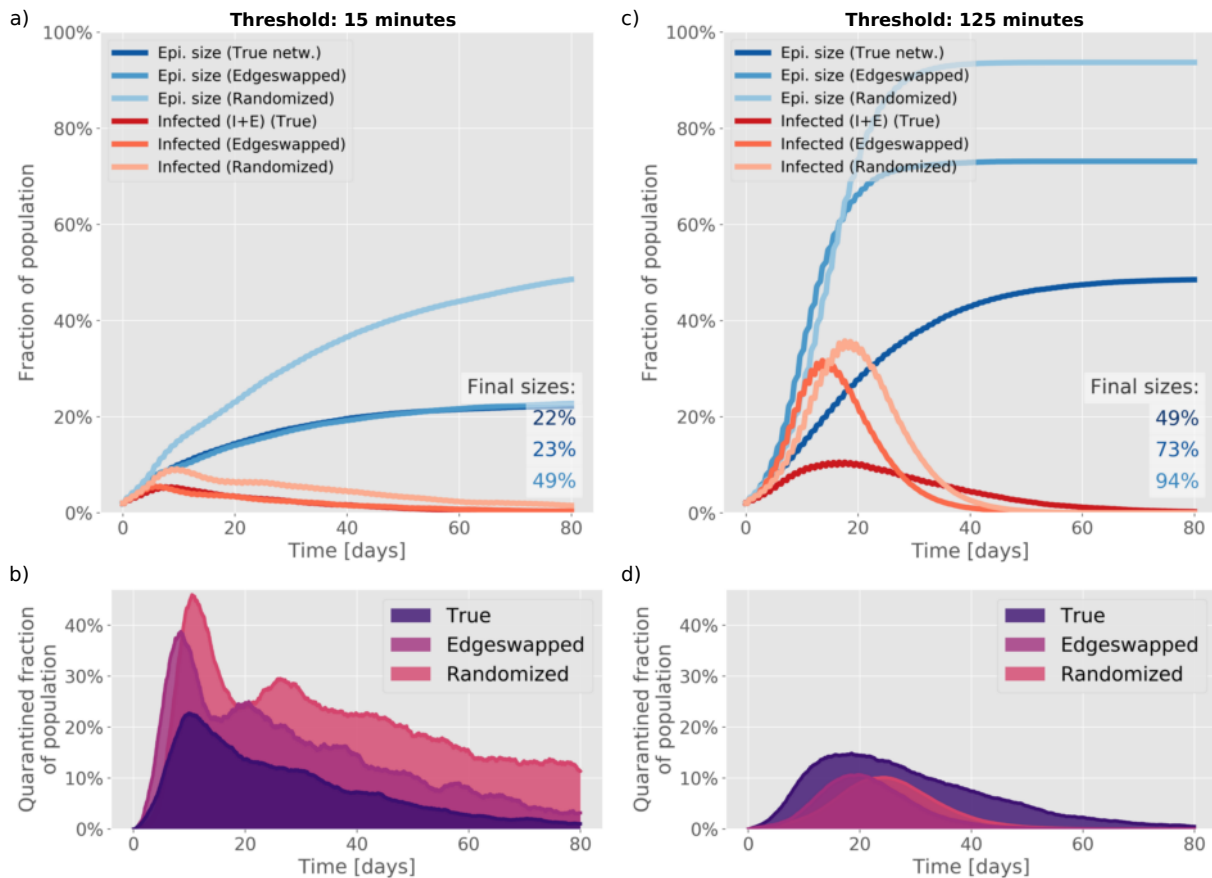


FIG. 4. **The effects of social heterogeneity on contact tracing at different thresholds.** Comparison of exposed + presymptomatic + infected (red) and recovered (blue) individuals in the three networks types. The testing rate is set at 0.5 times the rate for leaving the symptomatic infectious stage, giving a 25% probability of being tested while infected.

191 The epidemic *peak*, on the other hand, is quite sensitive to the social structure. The peak height
 192 increases by 10 percentage points when social network structure is destroyed, whereas eliminating
 193 the differences in social activity levels as well causes a further increase of just 4 percentage points.
 194 Furthermore, the heterogeneous activity leads to a faster initial growth of the epidemic, reaching the
 195 peak earlier. The mechanism behind this is that highly socially active individuals are more likely
 196 to contract as well as transmit the disease, meaning that they dominate the early epidemic.[49]

197 **Tracing depends on heterogeneity in a contact threshold-sensitive fashion** Contact
 198 tracing is most effective on the true social network, and performs poorly on the randomized network
 199 (Fig. 4), regardless of the contact threshold. The relative efficiency on the edge swapped network,
 200 however, depends quite strongly on the contact threshold, i.e. on how much cumulative contact
 201 time with a known infected person is allowed before triggering a quarantine. With a fairly short

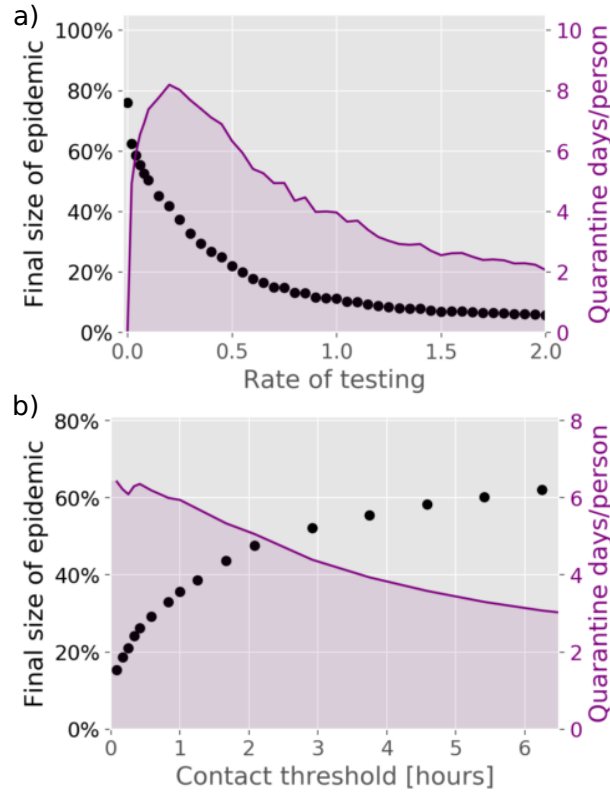


FIG. 5. Contact tracing effectiveness. Disease parameters are identical to those of Fig. 3. a) Rate of testing vs final size of epidemic and average number of days spent in quarantine per person. The contact threshold is set at 15 minutes. The rate of testing is measured in units of the rate for leaving the infected state, meaning that a rate of testing of 1 corresponds to a 50% chance of being tested during the infectious period. b) Contact threshold vs final size of epidemic and average number of days spent in quarantine per person. The rate of testing is set at 0.5 times the rate for leaving the symptomatic infectious stage, giving a 25% probability of being tested while infected. For each value of the parameter, 50 simulations were run.

202 contact threshold of 15 minutes (Fig. 4a), contact tracing on the edge swapped and true network
 203 are both highly effective, resulting in a final epidemic sizes of 22-23%. With a higher contact
 204 threshold of 125 minutes (Fig. 4c+d), contact tracing is much less effective in general, but now
 205 both of the homogenized networks perform much worse than the true network. This finding owes
 206 to the fact that repeated contacts are less frequent in the homogenized networks. It also explains
 207 the fact that the average quarantine time is much lower in the homogenized networks at high
 208 thresholds (Fig. 4d), since very few infected contacts are traced. A higher contact threshold thus
 209 has the advantage of reducing the overall time spent in quarantine (Fig4b+d) but results in a
 210 reduced epidemic control.

211 For contact tracing to be effective at higher contact thresholds, a substantial degree of temporal

212 correlation in contact dynamics is necessary. Both edge swapping and randomization reduce the
213 temporal correlation. To quantify this, we find that the median fraction of long contacts (of at
214 least 60 minutes cumulative duration) which are repeated from one week to the next is 30% in
215 the real network, while edge swapping and randomizing reduces this number to zero. Evidently,
216 repeated contacts are necessary for tracing to be effective at higher thresholds.

217 The edge swapping procedure reduces temporal correlation in two ways, which may be studied
218 separately. Firstly, it destroys correlations in social structure (“who meets who”), i.e. the identities
219 of contact partners are randomized, reducing the occurrence of repeated contacts as described
220 above. Secondly, the procedure destroys the duration distribution of contact *durations*, breaking
221 e.g. a 10 minute contact into two uncorrelated 5 minute contacts. In the Supplemental Material, we
222 describe an alternative edge swapping algorithm which preserves the durations of contacts, while
223 still swapping the individuals. This allows us to study the importance of the contact duration
224 distribution and the existence of repeated contacts separately. At high contact tracing thresholds
225 (125 minutes) We find that even duration-preserving edge swapping reduces the mitigative effect
226 of contact tracing relative to the true network. The reduction is not as strong as with the simple
227 edge swapping algorithm, leading us to conclude that there are two effects at play: destroying
228 the duration distribution leads to poorer performance, but simply randomizing the identities of
229 contacts while preserving degree and duration distributions has a detrimental effect in and of
230 itself. Conversely, one may conclude that repeated contacts (temporal correlations in the identities
231 of contact partners) as well as an inhomogeneous contact duration distribution are important
232 features which improve the effectiveness of contact tracing. Our finding that contact tracing on
233 the randomized network performs poorer than the edge swapped version is in good agreement with
234 the findings of [50], which shows theoretically that the presence of highly connected *hubs* in a social
235 network improves contact tracing.

236 Due to the limited temporal resolution of the Bluetooth proximity data, obtained only at 5 minute
237 intervals, the fidelity at very short contact tracing thresholds of e.g. 5 minutes will be lower. In
238 the Supplemental Material we explore the results obtainable at these lower contact thresholds, and
239 present a theoretical argument for the observed patterns. We find that mitigation by test-trace-
240 isolate (TTI) is always more *efficient* on the true network, in the sense of preventing more cases
241 per day of quarantine, but that the randomized network may in fact lead to a *lower* final attack
242 rate when contact thresholds tend to zero. The effect is due to the number of quarantines triggered
243 in the random network diverging in this limit.

244 **The rate of testing and contact threshold** The regular testing considered in our contact
 245 tracing algorithm is determined by a rate of testing which reflects several real-world factors not
 246 individually modeled here, such as general test capacity, symptom development and willingness to
 247 participate in testing. In Fig. 5a, we explore the influence of the rate of testing on the final size of
 248 the epidemic and the average time spent in quarantine. As one would expect, the quarantine time
 249 vanishes at very low rates of testing, where the epidemic size is maximal. Whereas the epidemic
 250 size is a decreasing function of testing, the quarantine time does not display a simple monotonic
 251 response to an increase in testing. Rather, it attains a maximum at 10% probability of being
 252 tested, followed by a gradual decline. This highlights the importance that changes in the testing
 253 strategy should go hand-in-hand with considerations of the nontrivial influence on the quarantine
 254 time. As such, it is possible to achieve a lower total quarantine time by increasing testing levels,
 255 simply due to the improved epidemic control.

256 If the aim is to keep the final size of the epidemic below for example 25%, our results show
 257 that a contact threshold of less than 30min is necessary (Fig. 5b). Note that the concurrent
 258 implementation of other mitigation strategies such as social distancing or limits on gathering will
 259 increase this critical threshold.

260 IV. DISCUSSION

261 In order to assess and credibly model the effectiveness of mitigation strategies, it is necessary
 262 to know which idealizations can be safely made, and which complexities must be retained in
 263 models. The present work shows that realistic social structure is an indispensable complexity
 264 when attempting to model contact tracing strategies and predict their effectiveness.

265 Although the social proximity data used in this study do not represent the social activity in a
 266 complex society, it exhibits a relevant level of social heterogeneity, which is stable over timescales
 267 long enough that it can influence epidemic dynamics. In this sense, the data can serve as a
 268 valuable model system in which to evaluate the impact of heterogeneity on disease propagation and
 269 mitigation of epidemics. We have found that social activity levels are exponentially distributed in
 270 this cohort, something which is consistent with observations of [7], who find a coefficient of variation
 271 of 0.8 for social contacts, for persons aged 20-30. The person-specific social activity exhibited in
 272 our data remains consistent over time, with both the 1-day and the 7-week activity patterns having
 273 coefficients of variations close to 1, representing a quenched disorder on the relevant timescale.

274 Even in the absence of mitigation, the social heterogeneity exhibited by our cohort significantly

275 affects the epidemic trajectory. However, not all outcomes are affected similarly. The epidemic peak
276 height is found to be sensitive to the social structure, while the final size of the epidemic is primarily
277 affected by heterogeneity in social *time*. The isolated sensitivity of attack rates to heterogeneous
278 social activity, i.e. differences in contact *time*, can be studied in a well-mixed compartmental model,
279 as was recently done [51], underscoring that the effect is not only present in structured networks.
280 The influence of social structure, however, is a more complex phenomenon and requires network
281 models, either synthetic or using observational social network data [1]. The effects of network
282 structure in epidemic spreading were previously studied by Barthélemy et al. using synthetic
283 social networks, which were however assumed static [49]. They found the mechanism to be a
284 hierarchical progression, with more well-connected individuals being infected early on, and more
285 sparsely connected nodes being affected later in the epidemic, if at all. However, this mechanism
286 depends on connectivity being a *quenched* variable, i.e. one that sticks to each individual over
287 time. We find that this condition is satisfied, at least on a timescale of a few months.

288 The sizable effects of social structure and heterogeneous activity seen in this study has implications
289 for epidemiological modelling in general. Due to their lack of social structure, traditional well-
290 mixed S(E)IR models would overestimate the severity of the epidemic, or, conversely, lead to an
291 underestimation of transmission risk when fitted to an observed epidemic trajectory. In a previous
292 modelling study [52], it was shown that heterogeneity in the *susceptibility* of individuals likewise
293 reduces the overall severity.

294 Once contact tracing is implemented, the effect of social heterogeneity becomes more complex.
295 We found that social structure and heterogeneous activity levels substantially increase the efficiency
296 of contact tracing. However, when the contact tracing threshold is low, heterogeneity in activity
297 levels alone improves effectiveness substantially, and network structure alone has less of an effect.
298 When the contact tracing threshold is high, both social network structure and heterogeneous
299 activity levels are necessary for efficient tracing. Furthermore, we found that the presence of
300 heterogeneity in contact duration improves the efficiency of contact tracing in itself. Our findings
301 also highlight that neither a quenched nor an annealed view of contact networks are sufficient for
302 modelling the spread of a disease such as COVID-19, since no clear separation of scales is present.
303 Important network dynamics takes place on time-scales shorter than an infectious period, while
304 some aspects of network structure and social activity are stable on timescales corresponding to
305 several generations of the disease. While many previous approaches have relied on such a separation
306 of time scales, sophisticated analytic frameworks for epidemic spreading on time-varying networks
307 have been proposed in recent years, allowing for e.g. continuously varying networks [53, 54].

308 The two central parameters of our contact tracing algorithm, the *rate of testing* and the *contact*
309 *threshold*, are not on equal footing. The rate of testing is influenced both by factors which are
310 within our control, such as the overall availability of testing, and by factors which are essentially
311 intrinsic to SARS-CoV-2, such as the rate at which symptoms develop. The contact threshold, on
312 the other hand, is a fully controllable parameter and essentially constitutes a design decision when
313 e.g. developing contact tracing applications [34]. Our results indicate that the contact threshold
314 must be kept quite low (< 30 minutes) if relatively efficient control (reducing epidemic final size by
315 about two thirds) is to be attained in an otherwise unmitigated epidemic. We find that the strength
316 of mitigation depends strongly on the rate of testing. This is expected since the ability to trace
317 contacts depends on the chance of identifying at least one case in the infection chain by regular
318 testing. What is perhaps less obvious is that the total quarantine time has a nontrivial (inverted
319 U-shaped) dependence on the rate of testing. As the rate is increased from 0, the quarantine
320 time increases. However, once an appreciable level of epidemic control has been achieved through
321 contact tracing, it begins to decline, with the peak value being attained at a rate corresponding to
322 a 10% probability of being tested while infected.

323 In this study, we have only considered forward contact tracing, where the primary objective is to
324 track down individuals who might have been infected by the index case. However, other schemes
325 exist, and two recent papers which came out after the initial publication of this manuscript have
326 shown that backwards contact tracing has an advantage in scenarios with highly clustered cases
327 [55, 56], i.e. where the transmission dynamics is overdispersed such that a few individuals cause
328 a high number of secondary infections while the majority cause few. Such clustering may arise by
329 several mechanisms of biological as well as social origin [57]. Recently, several studies have found
330 that COVID-19 transmission is in fact highly heterogeneous [57–62]. While we have focused on
331 the impact of social heterogeneity on mitigation by contact tracing, a recent study showed that
332 heterogeneity in biological infectiousness has a considerable impact on the feasibility of COVID-19
333 mitigation strategies which rely on contact network reduction [63], such as lockdowns.

334 While our study has highlighted the importance of network and activity heterogeneity for the
335 efficiency of contact tracing, some previous studies have highlighted other network measures, such as
336 *degree*, *betweenness* and *reach* as useful in further targeting contact tracing [29, 64]. It is our opinion
337 that there is still much to be learned about the usage of network data for the improvement of contact
338 tracing – and in order to identify the relevant mechanisms, modeling studies are indispensable.

339 In conclusion, heterogeneity in social activity makes mitigation by contact tracing much more
340 effective. If only more frequent contacts can be traced, social network structure becomes important

341 as well. It is thus important that realistic social heterogeneity and structure be taken into account
 342 when modeling contact tracing, as failure to do so may lead to underestimation of its effectiveness.

343

V. ACKNOWLEDGMENTS

344 We thank Gaute Linga, Kristian S. Olsen, Andreas Eilersen and Julius B. Kirkegaard for enlight-
 345 ening discussions. Our research has received funding from the European Research Council (ERC)
 346 under the European Union's Horizon 2020 research and innovation programme, grant agreement
 347 No. 740704.

348

Author Contributions

349 BFN, KS, LS, JM Wrote the paper; BFN analyzed the data and performed the simulations;
 350 BFN, KS, JM Developed the model; BFN, KS, LS, JM designed the study.

-
- 351 [1] S. Eubank, C. Barrett, R. Beckman, K. Bisset, L. Durbeck, C. Kuhlman, B. Lewis, A. Marathe,
 352 M. Marathe, and P. Stretz, Detail in network models of epidemiology: are we there yet?, *Journal of*
 353 *biological dynamics* **4**, 446 (2010).
- 354 [2] S. Bansal, B. T. Grenfell, and L. A. Meyers, When individual behaviour matters: homogeneous and
 355 network models in epidemiology, *Journal of the Royal Society Interface* **4**, 879 (2007).
- 356 [3] W. O. Kermack, A. G. McKendrick, and G. T. Walker, A contribution to the mathematical theory of
 357 epidemics, *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical*
 358 *and Physical Character* **115**, 700 (1927).
- 359 [4] Norwegian Institute of Public Health, Coronavirus modelling at the NIPH, [https://www.fhi.no/en/
 360 id/infectious-diseases/coronavirus/coronavirus-modelling-at-the-niph-fhi/](https://www.fhi.no/en/id/infectious-diseases/coronavirus/coronavirus-modelling-at-the-niph-fhi/) (2020), [On-
 361 line; accessed 28-May-2020].
- 362 [5] L. Peng, W. Yang, D. Zhang, C. Zhuge, and L. Hong, Epidemic analysis of covid-19 in china by
 363 dynamical modeling (2020), arXiv:2002.06563 [q-bio.PE].
- 364 [6] N. Ferguson, D. Laydon, G. Nedjati Gilani, N. Imai, K. Ainslie, M. Baguelin, S. Bhatia, A. Boonyasiri,
 365 Z. Cucunuba Perez, G. Cuomo-Dannenburg, *et al.*, Report 9: Impact of non-pharmaceutical inter-
 366 ventions (NPIs) to reduce COVID19 mortality and healthcare demand, Imperial College COVID-19
 367 Response Team (2020).
- 368 [7] J. Mossong, N. Hens, M. Jit, P. Beutels, K. Auranen, R. Mikolajczyk, M. Massari, S. Salmaso, G. S.
 369 Tomba, J. Wallinga, *et al.*, Social contacts and mixing patterns relevant to the spread of infectious

- diseases, PLoS medicine **5** (2008).
- [8] P. Klepac, S. Kissler, and J. Gog, Contagion! the BBC Four Pandemic—the model behind the documentary, *Epidemics* **24**, 49 (2018).
- [9] P. Klepac, A. J. Kucharski, A. J. Conlan, S. Kissler, M. Tang, H. Fry, and J. R. Gog, Contacts in context: large-scale setting-specific social mixing matrices from the BBC pandemic project, *medRxiv* (2020).
- [10] L. Pellis, S. Cauchemez, N. M. Ferguson, and C. Fraser, Systematic selection between age and household structure for models aimed at emerging epidemic predictions, *Nature communications* **11**, 1 (2020).
- [11] K. Prem, Y. Liu, T. W. Russell, A. J. Kucharski, R. M. Eggo, N. Davies, S. Flasche, S. Clifford, C. A. Pearson, J. D. Munday, *et al.*, The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: a modelling study, *The Lancet Public Health* (2020).
- [12] M. J. Keeling, E. Hill, E. Gorsich, B. Penman, G. Guyver-Fletcher, A. Holmes, T. Leng, H. McKimm, M. Tamborrino, L. Dyson, and M. Tildesley, Predictions of covid-19 dynamics in the uk: short-term forecasting and analysis of potential exit strategies, *medRxiv* 10.1101/2020.05.10.20083683 (2020).
- [13] Y. Huang, X. Cai, B. Zhang, G. Zhu, T. Liu, P. Guo, J. Xiao, X. Li, W. Zeng, J. Hu, *et al.*, Spatiotemporal heterogeneity of social contact patterns related to infectious diseases in the guangdong province, china, *Scientific reports* **10**, 1 (2020).
- [14] S. Bansal, J. Read, B. Pourbohloul, and L. A. Meyers, The dynamic nature of contact networks in infectious disease epidemiology, *Journal of biological dynamics* **4**, 478 (2010).
- [15] M. J. Keeling, The effects of local spatial structure on epidemiological invasions, *Proceedings of the Royal Society of London. Series B: Biological Sciences* **266**, 859 (1999).
- [16] P. Block, M. Hoffman, I. J. Raabe, J. B. Dowd, C. Rahal, R. Kashyap, and M. C. Mills, Social network-based distancing strategies to flatten the covid-19 curve in a post-lockdown world, *Nature Human Behaviour* , 1 (2020).
- [17] M. Barlow, A branching process with contact tracing, *arXiv preprint arXiv:2007.16182* (2020).
- [18] C. M. Peak, L. M. Childs, Y. H. Grad, and C. O. Buckee, Comparing nonpharmaceutical interventions for containing emerging epidemics, *Proceedings of the National Academy of Sciences* **114**, 4023 (2017).
- [19] J. Müller and V. Hösel, Contact tracing & super-spreaders in the branching-process model, *arXiv preprint arXiv:2010.04942* (2020).
- [20] M. J. Keeling, T. D. Hollingsworth, and J. M. Read, The efficacy of contact tracing for the containment of the 2019 novel coronavirus (covid-19)., *medRxiv* 10.1101/2020.02.14.20023036 (2020).
- [21] R. M. Anderson, H. Heesterbeek, D. Klinkenberg, and T. D. Hollingsworth, How will country-based mitigation measures influence the course of the covid-19 epidemic?, *The Lancet* **395**, 931 (2020).
- [22] J. Hellewell, S. Abbott, A. Gimma, N. I. Bosse, C. I. Jarvis, T. W. Russell, J. D. Munday, A. J. Kucharski, W. J. Edmunds, F. Sun, *et al.*, Feasibility of controlling covid-19 outbreaks by isolation of cases and contacts, *The Lancet Global Health* (2020).

- 406 [23] A. J. Kucharski, P. Klepac, A. Conlan, S. M. Kissler, M. Tang, H. Fry, J. Gog, J. Edmunds, C. C.-. W.
407 Group, *et al.*, Effectiveness of isolation, testing, contact tracing and physical distancing on reducing
408 transmission of sars-cov-2 in different settings, medRxiv (2020).
- 409 [24] L. Ferretti, C. Wymant, M. Kendall, L. Zhao, A. Nurtay, L. Abeler-Dörner, M. Parker, D. Bonsall, and
410 C. Fraser, Quantifying sars-cov-2 transmission suggests epidemic control with digital contact tracing,
411 Science **368** (2020).
- 412 [25] A. Eilersen and K. Sneppen, Estimating cost-benefit of quarantine length for covid-19 mitigation,
413 medRxiv 10.1101/2020.04.09.20059790 (2020).
- 414 [26] I. Z. Kiss, D. M. Green, and R. R. Kao, Disease contact tracing in random and clustered networks,
415 Proceedings of the Royal Society B: Biological Sciences **272**, 1407 (2005).
- 416 [27] M. Gasperek, M. Racko, and M. Dubovsky, A stochastic, individual-based model for the evaluation
417 of the impact of non-pharmacological interventions on covid-19 transmission in slovakia, medRxiv
418 10.1101/2020.05.11.20096362 (2020).
- 419 [28] A. Aleta, D. Martin-Corral, A. P. y Piontti, M. Ajelli, M. Litvinova, M. Chinazzi, N. E. Dean, M. E.
420 Halloran, I. M. Longini Jr, S. Merler, *et al.*, Modelling the impact of testing, contact tracing and
421 household quarantine on second waves of covid-19, Nature Human Behaviour **4**, 964 (2020).
- 422 [29] M. Serafino, H. S. Monteiro, S. Luo, S. D. Reis, C. Igual, A. S. L. Neto, M. Travizano, J. S. Andrade Jr,
423 and H. A. Makse, Superspreading k-cores at the center of covid-19 pandemic persistence, arXiv preprint
424 arXiv:2103.08685 (2021).
- 425 [30] S. Chang, E. Pierson, P. W. Koh, J. Gerardin, B. Redbird, D. Grusky, and J. Leskovec, Mobility
426 network models of covid-19 explain inequities and inform reopening, Nature , 1 (2020).
- 427 [31] P. Doung-Ngern, R. Suphanchaimat, A. Panjangampatthana, C. Janekrongtham, D. Ruampoom,
428 N. Daochaeng, N. Eungkanit, N. Pisitpayat, N. Srisong, O. Yasopa, *et al.*, Case-control study of use of
429 personal protective measures and risk for sars-cov 2 infection, thailand, Emerging Infectious Diseases
430 **26**, 2607 (2020).
- 431 [32] D. K. Chu, E. A. Akl, S. Duda, K. Solo, S. Yaacoub, H. J. Schünemann, A. El-harakeh, A. Bognanni,
432 T. Lotfi, M. Loeb, *et al.*, Physical distancing, face masks, and eye protection to prevent person-to-
433 person transmission of sars-cov-2 and covid-19: a systematic review and meta-analysis, The Lancet
434 **395**, 1973 (2020).
- 435 [33] V. Sekara and S. Lehmann, The strength of friendship ties in proximity sensor data, PLOS ONE **9**, 1
436 (2014).
- 437 [34] Apple Inc., Building an App to Notify Users of COVID-19 Exposure, https://developer.apple.com/documentation/exposurenotification/building_an_app_to_notify_users_of_covid-19_exposure
438 https://developer.apple.com/documentation/exposurenotification/building_an_app_to_notify_users_of_covid-19_exposure
439 exposure (2020), [Online; accessed 31-May-2020].
- 440 [35] S. Maslov and K. Sneppen, Specificity and stability in topology of protein networks, Science **296**, 910
441 (2002).

- 442 [36] A. Stopczynski, V. Sekara, P. Sapiezynski, A. Cuttone, M. M. Madsen, J. E. Larsen, and S. Lehmann,
443 Measuring large-scale social networks with high resolution, *PloS one* **9**, e95978 (2014).
- 444 [37] A. Mollgaard, I. Zettler, J. Dammeyer, M. H. Jensen, S. Lehmann, and J. Mathiesen, Measure of node
445 similarity in multilayer networks, *PloS one* **11** (2016).
- 446 [38] M. U. Mondelli, M. Colaneri, E. M. Seminari, F. Baldanti, and R. Bruno, Low risk of sars-cov-2
447 transmission by fomites in real-life conditions, *The Lancet Infectious Diseases* (2020).
- 448 [39] N. R. Jones, Z. U. Qureshi, R. J. Temple, J. P. Larwood, T. Greenhalgh, and L. Bourouiba, Two metres
449 or one: what is the evidence for physical distancing in covid-19?, *bmj* **370** (2020).
- 450 [40] Scientific Advisory Group on Emergencies, Transmission of SARS-CoV-2 and Mitigat-
451 ing Measures, [https://assets.publishing.service.gov.uk/government/uploads/system/
452 uploads/attachment_data/file/892043/S0484_Transmission_of_SARS-CoV-2_and_Mitigating_
453 Measures.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/892043/S0484_Transmission_of_SARS-CoV-2_and_Mitigating_Measures.pdf) (2020), [Online; accessed 14-Dec-2020].
- 454 [41] World Health Organization, Q&A on coronaviruses (COVID-19), [https://www.who.int/
455 emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/
456 q-a-coronaviruses](https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/q-a-coronaviruses) (2020), [Online; accessed 28-May-2020].
- 457 [42] Centers for Disease Control and Prevention, How COVID-19 Spreads, [https://www.cdc.gov/
458 coronavirus/2019-ncov/prevent-getting-sick/how-covid-spreads.html](https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/how-covid-spreads.html) (2020), [Online; ac-
459 cessed 28-May-2020].
- 460 [43] A. Stopczynski, A. S. Pentland, and S. Lehmann, Physical proximity and spreading in dynamic social
461 networks, arXiv preprint arXiv:1509.06530 (2015).
- 462 [44] Our World In Data and European Centre for Disease Prevention and Control, covid-19-data (Deaths),
463 <https://github.com/owid/covid-19-data/> (2020).
- 464 [45] A. Remuzzi and G. Remuzzi, Covid-19 and italy: what next?, *The Lancet* (2020).
- 465 [46] M. A. Billah, M. M. Miah, and M. N. Khan, Reproductive number of coronavirus: A systematic review
466 and meta-analysis based on global level evidence, *PloS one* **15**, e0242128 (2020).
- 467 [47] L. Baumgarten and S. Bornholdt, Epidemics with asymptomatic transmission: Sub-critical phase from
468 recursive contact tracing, arXiv preprint arXiv:2008.09896 (2020).
- 469 [48] D. Klinkenberg, C. Fraser, and H. Heesterbeek, The effectiveness of contact tracing in emerging epi-
470 demics, *PloS one* **1**, e12 (2006).
- 471 [49] M. Barthélemy, A. Barrat, R. Pastor-Satorras, and A. Vespignani, Dynamical patterns of epidemic
472 outbreaks in complex heterogeneous networks, *Journal of theoretical biology* **235**, 275 (2005).
- 473 [50] A. Reyna-Lara, D. Soriano-Paños, S. Gómez, C. Granell, J. T. Matamalas, B. Steinegger, A. Arenas,
474 and J. Gómez-Gardeñes, Virus spread versus contact tracing: Two competing contagion processes,
475 *Physical Review Research* **3**, 013163 (2021).
- 476 [51] T. Britton, F. Ball, and P. Trapman, A mathematical model reveals the influence of population het-
477 erogeneity on herd immunity to sars-cov-2, *Science* **369**, 846 (2020).

- 478 [52] M. G. M. Gomes, R. M. Corder, J. G. King, K. E. Langwig, C. Souto-Maior, J. Carneiro, G. Goncalves,
479 C. Penha-Goncalves, M. U. Ferreira, and R. Aguas, Individual variation in susceptibility or exposure
480 to sars-cov-2 lowers the herd immunity threshold, medRxiv 10.1101/2020.04.27.20081893 (2020).
- 481 [53] E. Valdano, M. R. Fiorentin, C. Poletto, and V. Colizza, Epidemic threshold in continuous-time evolving
482 networks, Phys. Rev. Lett. **120**, 068302 (2018).
- 483 [54] A. Koher, H. H. K. Lentz, J. P. Gleeson, and P. Hövel, Contact-based model for epidemic spreading on
484 temporal networks, Phys. Rev. X **9**, 031017 (2019).
- 485 [55] A. Endo *et al.*, Implication of backward contact tracing in the presence of overdispersed transmission
486 in covid-19 outbreaks, Wellcome open research **5** (2020).
- 487 [56] S. Kojaku, L. Hébert-Dufresne, E. Mones, S. Lehmann, and Y.-Y. Ahn, The effectiveness of backward
488 contact tracing in networks, Nature Physics **17**, 652 (2021).
- 489 [57] B. M. Althouse, E. A. Wenger, J. C. Miller, S. V. Scarpino, A. Allard, L. Hébert-Dufresne, and H. Hu,
490 Superspreading events in the transmission dynamics of sars-cov-2: Opportunities for interventions and
491 control, PLoS biology **18**, e3000897 (2020).
- 492 [58] Y. Zhang, Y. Li, L. Wang, M. Li, and X. Zhou, Evaluating transmission heterogeneity and super-
493 spreading event of covid-19 in a metropolis of china, International Journal of Environmental Research
494 and Public Health **17**, 3705 (2020).
- 495 [59] D. Miller, M. A. Martin, N. Harel, T. Kustin, O. Tirosh, M. Meir, N. Sorek, S. Gefen-Halevi, S. Amit,
496 O. Vorontsov, *et al.*, Full genome viral sequences inform patterns of sars-cov-2 spread into and within
497 israel, medRxiv (2020).
- 498 [60] Q. Bi, Y. Wu, S. Mei, C. Ye, X. Zou, Z. Zhang, X. Liu, L. Wei, S. A. Truelove, T. Zhang, *et al.*,
499 Epidemiology and transmission of covid-19 in 391 cases and 1286 of their close contacts in shenzhen,
500 china: a retrospective cohort study, The Lancet Infectious Diseases (2020).
- 501 [61] M. S. Lau, B. Grenfell, M. Thomas, M. Bryan, K. Nelson, and B. Lopman, Characterizing superspread-
502 ing events and age-specific infectiousness of sars-cov-2 transmission in georgia, usa, Proceedings of the
503 National Academy of Sciences (2020).
- 504 [62] J. B. Kirkegaard and K. Sneppen, Variability of individual infectiousness derived from aggregate statis-
505 tics of covid-19, medRxiv (2021).
- 506 [63] B. F. Nielsen, L. Simonsen, and K. Sneppen, COVID-19 superspreading suggests mitigation by social
507 network modulation, Physical Review Letters (to be published).
- 508 [64] M. Andre, K. Ijaz, J. D. Tillinghast, V. E. Krebs, L. A. Diem, B. Metchock, T. Crisp, and P. D.
509 McElroy, Transmission network analysis to complement routine tuberculosis contact investigations,
510 American journal of public health **97**, 470 (2007).

(REVIEW) THE COVID-19 PANDEMIC: KEY CONSIDERATIONS FOR THE EPIDEMIC AND ITS CONTROL

Authors: Søren Ørskov², Bjarke Frost Nielsen¹, Sofie Føns², Kim Sneppen¹ and Lone Simonsen².

¹ The Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark.

² Department of Science and Environment, Roskilde University, Roskilde, Denmark.

My contribution: Contributed to conceptualization, creating figures and writing of the manuscript.

Publication status: Published in *APMIS* (2021).

Hyperlink(s): <https://doi.org/10.1111/apm.13141>

APMIS Special Issue titled: Pandemics: "Pandemics: past, now and in the future"

Editors: Henrik Calum and Thomas Bjarnsholt

Title

The COVID-19 pandemic: Key considerations for the epidemic and its control

Søren Ørskov¹, Bjarke Frost Nielsen², Sofie Føns¹, Kim Sneppen², Lone Simonsen¹

¹ Department of science and environment, Roskilde University, Denmark

² Niels Bohr Institute (NBI), University of Copenhagen, Denmark

Abstract (398 words)

The response to the ongoing COVID-19 pandemic has been characterized by draconian measures and far too many important unknowns, such as the true mortality risk, the role of children as transmitters and the development and duration of immunity in the population. More than a year into the pandemic much has been learned and insights into this novel type of pandemic and options for control are shaping up.

Using a historical lens, we review what we know and still do not know about the ongoing COVID-19 pandemic. A pandemic caused by a member of the coronavirus family is a new situation following more than a century of influenza A pandemics. However, recent pandemic threats such as outbreaks of the related and novel deadly coronavirus SARS in 2003 and of MERS since 2012 had put coronaviruses on WHO's blueprint list of priority diseases. Like pandemic influenza, SARS-CoV-2 is highly transmissible ($R_0 \sim 2.5$). Furthermore, it can fly under the radar due to a broad clinical spectrum where asymptomatic and pre-symptomatic infected persons also transmit the virus – including children. COVID-19 is far more deadly than seasonal influenza; initial data from China suggested a case fatality rate of 2.3% – which would have been on par with the deadly 1918 Spanish influenza. But, while the Spanish influenza killed young, otherwise healthy adults, it is the elderly who are at extreme risk of dying of COVID-19. We review available seroepidemiological evidence of infection rates and compute infection fatality rates (IFR) for Denmark (0.5%), Spain (0.85%) and Iceland (0.3%). We also deduce that population age structure is key. SARS-CoV-2 is characterized by superspreading, so that ~10% of infected

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the [Version of Record](#). Please cite this article as [doi: 10.1111/APM.13141](https://doi.org/10.1111/APM.13141)

This article is protected by copyright. All rights reserved

individuals yield 80% of new infections. This phenomenon turns out to be an Achilles heel of the virus that may explain our ability to effectively mitigate outbreaks so far.

How will this pandemic come to an end? Herd immunity has not been achieved in Europe due to intense mitigation by non-pharmaceutical interventions; for example, only ~8% of Danes were infected across the 1st and 2nd wave. Luckily, we now have several safe and effective vaccines. Global vaccine control of the pandemic depends in great measure on our ability to keep up with current and future immune escape variants of the virus. We should thus be prepared for a race between vaccine updates and mutations of the virus. A permanent reopening of society highly depends on winning that race.

1. Coronaviruses: An era of new pandemic threats.

SARS-CoV-2, the virus that causes COVID-19, has led to the first confirmed coronavirus pandemic, and to many it has come as a surprise. We have experienced regular influenza pandemics for centuries [1], but there have been signs for some time that something new was on the horizon. A first warning came with the deadly outbreak of SARS-CoV in Asia in 2003, in which 10% of known cases died; the outbreak was controlled, and the virus eliminated rapidly despite the high transmissibility. Then MERS-CoV emerged in the Middle East in 2012, a virus with a far higher mortality rate but a poorer ability to spread among humans; it remains a pandemic threat to this day[2]. Previously, coronaviruses were thought to cause only mild illness in humans as the four existing human coronaviruses merely cause a common cold, but after these outbreaks, coronaviruses were put on WHO's blueprint list of priority diseases[3]. Predicting the severity and virus family of the next pandemic is difficult, but one thing is certain: Pandemics will occur intermittently in the future, as they have done historically[1].

Pandemic influenza has been characterized by an emerging novel virus that has adapted to spread effectively among humans. It has historically been accompanied by a shift in mortality to younger ages[4, 5]. But the COVID-19 deaths are largely affecting the elderly, with a mean of ~80 years in Denmark. Likewise, only 2.7% of Danish COVID-19 deaths have occurred in people younger than 60 years of age as of February 15, 2021[6]. This is quite different from historic influenza pandemics[7]: the Spanish Flu (1918), the Asian Flu (1957), the Hong Kong Flu (1968) and the Swine Flu (2009). In both the 1918 and 2009 pandemics, the mean age at death was 25-30 years, and 95% of deaths occurred in people younger than 65 years of age because of a greater degree of immunity in the older generations. The pandemics of 1957 and 1968 were somewhere in between these extremes in terms of age distribution[8]. A historical timeline of pandemics is seen in Figure 1.

Apart from this striking difference in the age-distribution of mortality, SARS-CoV-2 seems to spread in clusters – temporally as well as geographically. This might in part be due to the concept of superspreading which was also a characteristic of SARS-CoV and MERS-CoV.

In the following we will focus on the lessons from the COVID-19 pandemic and extract key insights that may point a way forward to end this world crisis.

Accepted Article

2. What makes SARS-CoV-2 so dangerous?

SARS-CoV-2 is a highly contagious respiratory pathogen that can spread directly through droplets and indirectly through fomite. In addition, there are several reports of superspreader events where aerosolized spread is the most likely explanation (Table 2). However, the relative importance of fomite and aerosols remains unknown[9]. The basic reproduction number, R_0 (the average number of contacts infected by each infected person), is around 2.5[10] in the absence of control. This is on par with the Spanish Flu[11, 12], meaning that a large fraction of the population needs to be immune in order to stop the epidemic from growing. This fraction (F) is classically estimated using the following formula[13]:

$$F = 1 - 1/R_0$$

In the case of SARS-CoV-2, this means that around 60% of the population must be immune in order to reach herd immunity according to this formula. This in turn means that a very large number of people would be infected if we let the epidemic run its natural course.

As SARS-CoV-2 is a newly emerged pathogen meaning there is no specific pre-existing immunity, it is assumed that almost everyone is susceptible to infection

In the beginning of the pandemic, there were reports of a high case fatality rate of around 2.3% and 19% getting severe disease requiring oxygen therapy and/or ICU admission. Some speculated correctly that we were just seeing the tip of the iceberg and thus overestimating these figures while others disagreed[14-17]. We are now certain that these proportions are way too high. Estimating the true proportions of infected people that are hospitalized, admitted to the ICU or die is best done with serology data. We have shown examples of such calculations based on serology, hospital and mortality data from Denmark, Spain and Iceland in table 1[18-24]. See our Danish report with SSI for details on assumptions and calculations [25].

It is interesting that the ICU rate is higher and the IFR lower in Iceland than in Denmark and Spain. Perhaps the Icelandic IFR is simply lower because of the younger age pattern of cases, suggesting elderly were better shielded from infection[18, 26]. ICU and hospitalization rates are difficult to compare across countries as those depend on local admission criteria and ICU definitions.

It is expected that the measured IFR would vary greatly depending on the demography of each country and other factors. For example, when we applied the Spanish age specific IFRs to the far younger demography of Ethiopia, we found an all-age IFR of only 0.10%, compared to 0.85% measured in Spain. To truly know the IFR in low-income settings, we would need national serology studies and complete COVID-19 mortality statistics. But our Ethiopia example gives a sense that the measured IFR may vary 10-fold between countries with an aging and a young population. Large meta-analyses have found similar effects[27, 28].

The WHO published a meta-analysis estimating the global IFR to be ~0.23%[29]. Another meta-analysis based on all available serology studies estimated a mean IFR of 0.68% (0.53-0.82%) [30]. These large differences show the importance of referring to a specific population or age stratum when stating an IFR.

Although the COVID-19 IFR is many times lower than for SARS and MERS, the quick and efficient spread of the SARS-CoV-2 virus has already given rise to many more infections and deaths. An alternative measure to the official death count is excess in all-cause mortality – above what is expected for a specific time of the year. The European EUROMOMO surveillance system allows for timely tracking of excess mortality in European countries and offers both historical and contemporary incidence in mortality (<https://www.euromomo.eu/>). Excess mortality follows the pandemic wave patterns in Europe over the last year. Excess mortality is clearly highest in the older age groups, but a slightly significant excess mortality is also seen in the age group of 15-44 years. No excess mortality is seen in the group of 0-14 years.

Finally, it is becoming increasingly clear that the disease burden is not adequately described by acute illness and mortality alone. An unknown proportion of recovered patients experience longer lasting and, in some instances, debilitating symptoms such as fatigue, dyspnea, chest pain, joint pain, anosmia and dysgeusia[31]. Only with time, and from ongoing study of large, representative populations of seropositive individuals, we will understand the duration of these sequelae and get a better idea of the true proportion of all infected individuals that experiences them.

3. Does SARS-CoV-2 have a weak spot?

Throughout the COVID-19 pandemic, news stories about superspreading events – in which a single person infected a large number of people within a short timeframe – have been cropping up regularly. By now, there is significant evidence from outbreaks and RNA sequence analyses that these are not just isolated events. Rather, a marked heterogeneity in transmission is part of the signature of SARS-CoV-2 [32-34]. Many outbreaks involving such superspreading have already been documented, and a database of more than 2000 cases has been compiled[35], we have included a few clear examples[36-41] where one individual infects several others.

Superspreading is known to have played a significant role in some previous coronavirus outbreaks, such as SARS and MERS[35, 42, 43] and is one of the epidemiological footprints that differentiate them from pandemic influenza[44].

The mechanism behind superspreading is not yet fully determined. It is not clear whether superspreading events can primarily be ascribed to large inter-individual variability in viral shedding over the duration of an infectious period, or if it is perhaps a highly temporal phenomenon, with short-lived spikes in shedding. It is clear that certain behaviors and procedures which facilitate aerosolization can at least contribute. These can range from everyday occurrences such as singing, to medical procedures such as intubation and tracheoscopy. Some studies found large variations in viral shedding and viral load between infected individuals[45, 46], but it is not clear that these were not just representing various stages of infectivity even though some cases point to specific persons being biological superspreaders. Most compelling, in one study from China, a single person caused a superspreading event, then went on to also infect everyone at home, suggesting that it was a particular superspreading person, rather than a singular event[41]. However, in several superspreading events, behavior seems to play a role – examples of high-risk activities are whistling and singing. This suggests that superspreading is a property of action also. Needless to say, the presence of a large (typically indoor) crowd is also a risk factor.

With a basic reproductive number of around 2.5[10], such a marked heterogeneity in transmission entails that the majority of infected individuals hardly transmit the disease at all. In Figure 2a, a

simulated infection chain is shown, clearly showing how the spread of the disease is entirely dependent on superspreaders[47].

This transmission heterogeneity can be summarized by the overdispersion parameter k (a number that - for small values of k - approximates the fraction of people that are responsible for 80% of the transmissions) [42], with lower k denoting a more heterogeneous transmission – i.e., one prone to superspreading events. For SARS-CoV, this was estimated to be 0.16, corresponding to a high level of transmission heterogeneity, while estimates for SARS-CoV-2 have been even more dramatic – around 0.10 [32, 33, 48]. This indicates that for SARS-CoV-2, the 10% most infectious individuals are responsible for approximately 80% of the transmission. Pandemic influenza, on the other hand, is much more homogeneously spread, with an estimated k value close to 1[44]. As we discuss below, this has significant consequences, and so we argue that the k value deserves widespread recognition, similar to the reproduction number R_0 . From a mathematical standpoint, this amounts to saying that it is not just the mean of the infectiousness distribution which matters, but also its variance.

Mathematical models have been used to study the impact that superspreading has on the effectiveness of mitigation strategies, demonstrating that efficiency of such strategies primarily rely on reducing social mixing in society[47], including for example a ban of large gatherings. Capturing these phenomena requires modeling on the level of individuals and this is not possible within traditional compartmental epidemiological models which assume completely random mixing. In popular terms, these models assume that each individual goes to a new job each minute and returns to a new home every evening.

The main finding is that the ability of superspreaders to transmit the disease to anywhere near their full potential can be effectively curbed by even a moderate reduction in the number of contacts that any given person has during an infectious period. For a more homogeneously transmitted disease, this would not be the case. In that case it would be necessary to reduce the number of distinct social contacts very close to the reproductive number R_0 to achieve significant mitigation.

As illustrated in figure 2a, superspreading has a tendency to lead to bursty infection chains which then have an increased chance of dying out, as the chain effectively terminates if none of the recently infected persons are themselves superspreaders[49].

Thus, superspreading is in full agreement with the bursty, geographically clustered outbreaks seen during this pandemic[50].

Figure 2b shows the result of reducing contacts outside households and work in an agent-based model. For a virus prone to superspreading the impact is substantial, while it has little impact for a more homogeneously transmitted virus. Thus, superspreading represents an Achilles heel of SARS-CoV-2, making the epidemic vulnerable to even moderate reductions in contacts. This, in turn, explains the high effectiveness of lockdown strategies.

It may be tempting to think that superspreading is merely a product of some people having many contacts – i.e., social heterogeneity. This social aspect of superspreading is probably partly true as socially active people are more likely to infect more. Interestingly, socially hyperactive people also have higher risk of becoming infected, meaning that highly active people are also super-receivers. A modeling study found that social heterogeneity lowers the herd immunity threshold, even in the absence of mitigation[51]. Purely biological superspreading that does not correlate with the superspreaders' own probability of becoming infected does not change the herd immunity threshold.

We saw earlier how superspreading drastically improves the effect of mitigation strategies which rely on reducing contacts. It is known that social heterogeneity leads to clustering of cases and so increases the effectiveness of another form of mitigation, namely test-trace-isolate strategies[51]. Since cases also have a tendency to occur in clusters in this case, superspreading too should make contact tracing easier and more effective. This is especially true if backward contact tracing is performed, since any given infection is quite likely to stem from a superspreader[52].

In conclusion, superspreading seems to represent an Achilles heel of SARS-CoV-2, which opens up possibilities for particularly effective mitigation, far more than what could ever be achieved for pandemic influenza. We argue that models used to explore the pandemic trajectory should take heterogeneity into account when evaluating possible mitigation strategies (and not just view it as “statistical noise”).

4. Unanswered questions

What role do children play in the COVID-19 pandemic?

In prior influenza pandemics, children played a major role as transmitters. It was therefore surprising that so few children figured amongst known cases in the early phase of the COVID-19 pandemic. This could be explained by children typically having only mild symptoms, but it was suspected early on that children were less susceptible and infectious than adults[53, 54]. Could it be true that children are not important transmitters in this pandemic? The best way to answer this question is by testing for SARS-CoV-2-antibodies in local outbreak settings or in randomized population samples; however, there are ethical and legal concerns when drawing blood from healthy children.

For adolescents (14-20 years old) new evidence has since clarified that this age group indeed plays an important role in the pandemic. High school outbreaks have been reported all over the world. The latest Danish evaluation of population seroprevalence found the second highest seroprevalence in the 12-19 years old age group (6.6% vs. 3.9% in the general population)[22]. A meta-analysis[55] found similar seroprevalences in adolescents and adults in population wide screening studies of several different countries. Secondary household attack rates were as high or higher for adolescents compared to adults.

For younger children, the same analysis found a lower seroprevalence in this age group than in adults[55]. However, to the best of our knowledge, none of the used antibody tests have been validated on pediatric populations. A German study found no difference between viral load in children and adults suggesting that children might be as infectious as adults[56].

A meta-analysis of contact tracing studies suggests a lower probability of secondary infections in children than adults, but the study was not conclusive[55]. When excluding studies with a high risk of bias (e.g., testing only symptomatic contacts – i.e., fewer children), this lower probability became non-significant[57]. A meta-analysis of contact tracing studies suggests a lower probability of secondary infections in children than adults, but the study was not conclusive[55]. When excluding studies with a high risk of bias (e.g., testing only symptomatic contacts – i.e., fewer children), this lower probability became non-significant[57].

Finally, household contact studies show a lower probability of a child being the index case of a household[58]. However, this could be due to a bias in ascertaining the index person – typically a symptomatic adult – masking the possibility that it was an asymptomatic child who brought the disease into the household in the first place.

Since the reopening of countries following the initial lockdowns, several notable outbreaks have been reported in younger pediatric populations. Examples include a youth overnight camp in Georgia for age 6-19 years in the US, where mass PCR testing revealed an attack rate of at least 44% among campers[59]. Additionally, 41 of 825 schools in Berlin had to close two weeks after reopening due to school outbreaks[60]. In the US, serious concerns were raised over re-opening schools after the summer[61].

Studies in which all pupils, teachers and their home contact are all tested – preferably using antibody tests – regardless of symptoms are the most informative and less biased. A study of this kind was performed at a high school in Oise, France, and underscored the high susceptibility and transmissibility in adolescents[62].

In conclusion, while susceptibility and infectiousness of children were downplayed for a long time, it has become increasingly clear (from the above-mentioned serology studies) that adolescents play an important role in this pandemic. The question remains open for younger children, an age group rarely tested. We do not, however, have evidence to suggest they can be disregarded. Furthermore, with the rise of the new British B.1.1.7 variant there is evidence from Israel that this variant leads to high attack rates even among young children[63]. Knowing the infectiousness of young children is naturally of key importance in informing policy decisions about keeping young children in schools and institutions. It is however very clear that children and adolescents have very mild infections – the reason for this remains a mystery, but a tempting explanation is a better innate immune response in children[64]. An exception to this is the multisystem inflammatory syndrome in children (MIS-C) after infection with SARS-CoV-2 which in some aspects resembles Kawasaki disease. Patients present with fever and severe illness involving two or more organ systems. The suggested pathogenesis involves post-infectious immune dysregulation. The syndrome is rare, and when it occurs it has a mortality rate of around 1,5%. The possibility of sequelae in children after SARS-CoV-2 is another important point, but

the data so far are inconclusive, and more research is needed to truly understand the impact of COVID-19 in children[65, 66]. This is important in weighing the risks and benefits of vaccinating children against COVID-19[67, 68].

SARS-CoV-2 immunity

In the early stages of the COVID-19 pandemic, there was intense debate over the immune response to SARS-CoV-2. Some researchers argued for long-lasting, effective immunity – even suggesting that we create immunity passports. Others, however, doubted that antibody responses would be lasting and remain highly prevalent in recovered individuals. Early on some were even concerned that the antibodies might not even be neutralizing[69, 70].

We now know that most people do in fact develop a lasting antibody response, lasting at least several months – and several studies have found antibodies to be neutralizing[24, 71-75]. There is evidence to suggest that cellular immunity is robust as well – and it might prove important if antibody titers decline[76]. Interestingly, between 20-50% of unexposed individuals (that is, from blood drawn before the pandemic virus existed) display significant SARS-CoV-2 specific T-cell response, possibly originating from immunity to the common and related cold coronaviruses [77]. The implications of this are still uncertain, but it would be interesting to examine the effect of this on SARS-CoV-2 susceptibility. More research is needed, and it would be particularly interesting to examine differences in pre-existing immune responses between different age groups including children and elderly.

Re-challenge studies in macaques also points towards a protective immune response[57, 78]. There is thus a theoretical basis for immunity.

An interesting case of real-life immunity was reported in a fishery vessel outbreak with a PCR-confirmed attack rate of 85.2% (104 of 122 individuals). Three previously recovered individuals with neutralizing antibodies were on board and none of them experienced any symptoms nor tested positive in the PCR-test. This real-life situation thus provides evidence of the protective effect of neutralizing antibodies ($p=0.002$, Fisher's exact test)[79]. Another notable real-life example was seen at an overnight summer school retreat in Wisconsin in the summer of 2020 reported by the CDC[80]. There was a great outbreak with an attack rate of 76% (116 cases) among the 152 attendees. 24 of the participants had positive serologic results before going to the

camp – all of these got negative RT-PCR results. The report provides no statistical test for this apparent immunity, but we have performed a Fisher's exact test showing $p < 0.001$. Thus, there is both theoretical and real-life evidence of immunity.

On the other hand, there have been reports of a few credible reinfection events in Hong Kong, Belgium and the Netherlands[81, 82]. Highly concerning is the growing evidence from Manaus, Northern Brazil, where herd immunity following the 1st wave was later overcome by a new variant dominating the 2nd wave[83-85].

What is the best way to control the epidemic until vaccine immunity is achieved?

While the world awaits widespread distribution of effective vaccines, it is critical to find a sustainable and acceptable way of living while suppressing the epidemic until we have vaccine-induced immunity, especially amongst the elderly and others at high risk. In our opinion, this is best achieved by measures that reduce excessive contacts in the public space, to avoid superspreading events[47].

While most countries adapted draconian measures, Sweden did not use lockdowns during the first wave and remained a semi-open society with open borders. Early on, Sweden had a high death toll which can be explained by a late implementation of their control measures, a full two weeks after the other Nordic countries went into lockdown. In Sweden, these measures were focused on voluntary changes in mobility and work behavior and, importantly, a further restriction of gatherings from a maximum of 500 to a maximum of 50 persons, as well as intensified efforts to secure elderly in nursing homes, while schools for children under 16 years remained open. With this relatively light control strategy, they achieved epidemic control around May 1st, so that the effective reproductive number was below 1 over the summer; until autumn where partial lifting of this ban, along with seasonal change, resulted in a substantial second wave (Figure 3). We wonder if the situation in Sweden during May-September showed us the potency of restrictions on large gatherings, isolated from the effect of other factors imposed in a full lockdown.

In our view, the Swedish success is that of getting to R_e below 1 – albeit too late – while maintaining a fairly free and open society. Furthermore, had this been achieved 2 weeks sooner, then Sweden would not have suffered the large death toll in the spring.

Some have argued that allowing R_e to somewhat exceed 1 could be desirable because it allows herd immunity to slowly build up in the population (Great Barrington Declaration)[86]. In our view however, this comes at an unacceptably high cost in terms of disease burden and deaths. We computed that cost for Denmark, by multiplying the IFR and the hospitalization rate, assuming the final epidemic size would be 60% of the danish population (Table 3). Using our estimates based on the latest two seroprevalence studies and the latest blood donor data from Denmark (week 4 of 2021) this gives us:

We found that natural herd immunity in Denmark would lead to ~20,000 deaths and ~90,000 hospitalizations. In developing countries, the cost of following such a strategy would presumably be far less dramatic, due to having low proportions of people above 60 years of age. In Denmark this age group accounts for 97.3% of COVID-19 deaths as of February 15, 2021[6]. One might suggest isolating the elderly and chronically ill and allowing herd immunity to develop in the rest of the population. In a sense, the numbers above actually already account for that because isolation of elderly and chronically ill has already been a part of the Danish strategy from the start. From the seroprevalence data, it is also clear that this has actually been quite successful. In the third round of the national seroprevalence study, the seroprevalence was estimated at 1.9% (0.9-3.4%) in those above 65 years of age while it was 7.3% (5.3-9.9%) in those between 20 and 29 years of age. In a situation with higher infection rates in society, it seems more difficult to avoid infections in nursing homes, hospitals and in the elderly part of the general population. In that case, the estimates of mortality and hospitalizations above are too low.

In addition, to avoid hospital overburdening, the reproductive number would have to be kept close to 1 (meaning a similar degree of restrictions to those needed to keep $R_e < 1$) until significant effects of herd immunity kick in – and this is a very slow process that is nowhere near happening in any western countries despite high death tolls.

On the contrary, with strict border control and quarantining of incoming travelers, a strategy of testing, contact tracing, local outbreak control combined with social distancing and hygiene measures has allowed to suppress the epidemic resulting in very low death tolls in islands such as Iceland, Faeroe Islands, South Korea, Taiwan and New Zealand despite having quite open societies during most of the pandemic[87]. Acting quickly to get R_e below 1 while disease prevalence is still low is, in our view, the best way to keep an open society in the long term.

However, the situation has recently been complicated by new, faster spreading variants such as lineage B.1.1.7, commonly known as the UK variant. This variant requires even tougher restrictions than what has been necessary until now, due to increased (around 50%) higher transmissibility[88].

5. How will this end?

Historically, influenza pandemics ended when sufficient immunity had built up in the population, even in the recent 2009 pandemic when the vaccine only became available after several waves[89]. We see four mutually non-exclusive ways of ending the crisis:

- A highly effective and widely available treatment of COVID-19
- Herd immunity achieved by natural infection of at least 60% of the population
- Herd immunity achieved by mass vaccinations
- Widespread availability of inexpensive rapid tests for repeated mass testing

Several treatments are in use, but none have been proven to drastically improve the prognosis of the disease. Remdesivir seemed initially to improve mortality in a specific subgroup of hospitalized patients[90], but a later meta-analysis by the WHO found no reduction in mortality for Remdesivir nor three other studied drugs (hydroxychloroquine, lopinavir and interferon beta-1a). Furthermore, a recent Cochrane review concludes that there is currently no evidence-based treatment for COVID-19 [91]. Combining this knowledge with the current vaccine advances, a game changer of a treatment does not seem to be the most plausible way out of the crisis in any near future.

As discussed in section 4, aiming for natural herd immunity is undesirable due to the substantial cost in terms of disease burden and lives lost. This is further complicated by the fact that new and more contagious variants like B.1.1.7 have emerged and are replacing the original variant in many countries, thus requiring an even higher percentage of the population to recover from infection in order to achieve herd immunity. Furthermore, allowing widespread infections while building up herd immunity increases the risk of escape mutations that can cause reinfections and give rise to future epidemic waves even though herd immunity has been established. This is a quite probable explanation of recent events in Manaus, Brazil, where a second, deadlier wave of COVID-19 has hit the Amazonas capital after an estimated attack rate of 76% in the first wave had apparently conferred herd immunity[84, 85]. Genetic sequencing points to the immune escape lineage P1 playing a major role in the second wave. In December, 51% of the sequenced SARS-CoV-2 genomes in the Amazonas belonged to the P1 strain and in January this figure had risen to 91% [83].

Preliminary data from the Novavax COVID-19 vaccine trial in South Africa – where an escape variant, B.1.351, is highly prevalent – points towards 60% (19.9-80.1%) protection against symptomatic, confirmed COVID-19 in HIV-negative, vaccinated individuals. Of concern is that the 1/3 of participants who were seropositive at entry (thought to have been due the original SARS-CoV-2 strain in the first wave) had no protection relative to the placebo group. This (along with the Manaus data) points to an unfortunate preliminary conclusion – the naturally acquired immunity does little to nothing to protect against reinfection with escape variants. Luckily, at least the Novavax vaccine seems to offer some protection. It will be interesting to see if this holds true for the other vaccine candidates. Based on in vitro studies on 8 human sera and sera from non-human primates, Moderna has found preliminary evidence suggesting that their mRNA-based COVID-19 vaccine (mRNA-1273) might not induce as high neutralizing antibody titers against the B.1.351 lineage relative to prior strains. The titers are still expected high enough to confer immunity, but out of caution Moderna has already sent an emerging variant booster vaccine into trial (mRNA-1273.351) against the B.1.351 variant[92]. By rapidly updating vaccines, we will have a more sustainable weapon against the new variants than allowing recurrent waves of new escape variants to confer herd immunity. Even though the vaccines do not completely protect against mild infections, all of the approved vaccines in the EU confer very high protection against severe disease[93].

A combination of natural and vaccine-based immunity is also possible, however, and one could argue that broadly imposed restrictions do no longer have ethical merit once those vulnerable to severe outcomes of infection have been vaccinated. However, hospitalization rates are not as age dependent as the fatality rate, so care must be taken that hospitals are not overwhelmed by quick lifting of measures. Also – immune senescence might leave many elderly vulnerable even after vaccination. Gradual reduction of restrictions while maintaining R_e around or below 1 based on hospital admissions might be the best way for a balanced return to a normal society. As more and more risk groups are vaccinated, we should expect a lowering of the risk of hospitalization meaning that an increase in infection rates will not necessarily significantly increase hospitalizations. Mathematical modelling will be crucial in informing the timing of reopening attempts – who and how many must be vaccinated before a COVID-19 wave in an open society is unable to overburden hospitals?

Regardless of how herd immunity is achieved, SARS-CoV-2 is likely to become endemic, and may cause occasional large resurgences, either due to waning of antibodies or due to the appearance of immune escape variants[94]. These phenomena mean that we might need to repeatedly vaccinate a large part of the population – e.g. each winter as we do for the seasonal flu.

It is thus clear that COVID-19 should not just be viewed as a temporary pandemic phenomenon, and that sustainable strategies are required. On a positive note, rapid tests have now become widely available, and these can significantly increase the speed of outbreak detection in vulnerable settings and of contact tracing in general. If rapid tests become cheaper and available for home use, they could realistically be used for recurrent mass testing of the entire population in order to curb the spread – such as it was successfully done in Slovakia in October 2020[95].

Literature list:

1. Miller, M.A., et al., *The Signature Features of Influenza Pandemics — Implications for Policy*. New England Journal of Medicine, 2009. **360**(25): p. 2595-2598.
2. Petrosillo, N., et al., *COVID-19, SARS and MERS: are they closely related?* Clin Microbiol Infect, 2020. **26**(6): p. 729-734.
3. <https://www.who.int/medicines/ebola-treatment/WHO-list-of-top-emerging-diseases/en/>.
4. Simonsen, L., et al., *Pandemic versus epidemic influenza mortality: a pattern of changing age distribution*. J Infect Dis, 1998. **178**(1): p. 53-60.
5. Murray, C.J.L., et al., *Estimation of potential global pandemic influenza mortality on the basis of vital registry data from the 1918–20 pandemic: a quantitative analysis*. The Lancet, 2006. **368**(9554): p. 2211-2218.
6. <https://experience.arcgis.com/experience/aa41b29149f24e20a4007a0c4e13db1d>.
7. Petersen, E., et al., *Comparing SARS-CoV-2 with SARS-CoV and influenza pandemics*. The Lancet Infectious Diseases, 2020. **20**(9): p. e238-e244.
8. Viboud, C., et al., *Preliminary Estimates of Mortality and Years of Life Lost Associated with the 2009 A/H1N1 Pandemic in the US and Comparison with Past Influenza Seasons*. PLoS Curr, 2010. **2**: p. Rrn1153.
9. <https://www.who.int/news-room/commentaries/detail/transmission-of-sars-cov-2-implications-for-infection-prevention-precautions>.
10. https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200306-sitrep-46-covid-19.pdf?sfvrsn=96b04adf_4.
11. Andreasen, V., C. Viboud, and L. Simonsen, *Epidemiologic characterization of the 1918 influenza pandemic summer wave in Copenhagen: implications for pandemic control strategies*. The Journal of infectious diseases, 2008. **197**(2): p. 270-278.
12. Mills, C.E., J.M. Robins, and M. Lipsitch, *Transmissibility of 1918 pandemic influenza*. Nature, 2004. **432**(7019): p. 904-906.
13. Fine, P., K. Eames, and D.L. Heymann, *“Herd Immunity”: A Rough Guide*. Clinical Infectious Diseases, 2011. **52**(7): p. 911-916.
14. Wu, Z. and J.M. McGoogan, *Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72 314 Cases From the Chinese Center for Disease Control and Prevention*. JAMA, 2020. **323**(13): p. 1239-1242.
15. <https://www.bmj.com/content/368/bmj.m606/rr-5>.

16. <https://www.nytimes.com/2020/03/04/health/china-lessons-aylward.html>.
17. <https://www.who.int/docs/default-source/coronaviruse/clinical-management-of-novel-cov.pdf>.
18. <https://www.covid.is/data-old>.
19. <https://www.isciii.es/Noticias/Noticias/Paginas/Noticias/PrimerosDatosEstudioENE COVID19.aspx?fbclid=IwAR3LuOScys3SZc37J5ldXB95k2LU5SRZ6Ujji583S1NQQPWU NHToNOu8EIA>. Available from:
20. <https://www.mscbs.gob.es/profesionales/saludPublica/ccayes/alertasActual/nCov/situacionActual.htm>.
21. <https://www.ssi.dk/-/media/arkiv/dk/aktuelt/nyheder/2020/notat---covid-19-prvalensundersgelsen.pdf?la=da>.
22. https://files.ssi.dk/praevalensundersoegelse_runde3.
23. <https://bloddonor.dk/coronavirus/>.
24. Gudbjartsson, D.F., et al., *Humoral Immune Response to SARS-CoV-2 in Iceland*. New England Journal of Medicine, 2020.
25. Ørskov, S., et al., *Fokusrapport: Mørketal. 2020, SSI*: <https://www.ssi.dk/-/media/arkiv/dk/aktuelt/sygdomsudbrud/covid19/fokusrapport---uge-35---mrketallet---final.pdf?la=da>.
26. Gudbjartsson, D.F., et al., *Spread of SARS-CoV-2 in the Icelandic Population*. New England Journal of Medicine, 2020. **382**(24): p. 2302-2315.
27. Ghisolfi, S., et al., *Predicted COVID-19 fatality rates based on age, sex, comorbidities, and health system capacity*. medRxiv, 2020: p. 2020.06.05.20123489.
28. Levin, A.T., et al., *ASSESSING THE AGE SPECIFICITY OF INFECTION FATALITY RATES FOR COVID-19: SYSTEMATIC REVIEW, META-ANALYSIS, AND PUBLIC POLICY IMPLICATIONS*. medRxiv, 2020: p. 2020.07.23.20160895.
29. Ioannidis, J., *The infection fatality rate of COVID-19 inferred from seroprevalence data*. 2020.
30. Meyerowitz-Katz, G. and L. Merone, *A systematic review and meta-analysis of published research data on COVID-19 infection-fatality rates*. medRxiv, 2020: p. 2020.05.03.20089854.
31. Carfi, A., et al., *Persistent Symptoms in Patients After Acute COVID-19*. JAMA, 2020. **324**(6): p. 603-605.

32. Bi, Q., et al., *Epidemiology and transmission of COVID-19 in 391 cases and 1286 of their close contacts in Shenzhen, China: a retrospective cohort study*. The Lancet Infectious Diseases, 2020. **20**(8): p. 911-919.
33. Endo, A., et al., *Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China*. Wellcome open research, 2020. **5**: p. 67-67.
34. Miller, D., et al., *Full genome viral sequences inform patterns of SARS-CoV-2 spread into and within Israel*. medRxiv, 2020: p. 2020.05.21.20104521.
35. <https://medium.com/@codecodekoen/covid-19-superspreading-events-database-4c0a7aa2342b>.
36. Hamner L, Dubbel P, Capron I, et al. High SARS-CoV-2 Attack Rate Following Exposure at a Choir Practice — Skagit County, Washington, March 2020. MMWR Morb Mortal Wkly Rep 2020;69:606–610. DOI: <http://dx.doi.org/10.15585/mmwr.mm6919e6>.
37. <https://calgary.ctvnews.ca/i-would-do-anything-for-a-do-over-calgary-church-hopes-others-learn-from-their-tragic-covid-19-experience-1.4933461>.
38. <https://nationalpost.com/news/how-an-edmonton-curling-tournament-became-a-hotspot-for-the-covid-19-outbreak-in-canada>.
39. Ghinai I, Woods S, Ritger KA, et al. Community Transmission of SARS-CoV-2 at Two Family Gatherings — Chicago, Illinois, February–March 2020. MMWR Morb Mortal Wkly Rep 2020;69:446–450. DOI: <http://dx.doi.org/10.15585/mmwr.mm6915e1>.
40. Lu, J., et al., *COVID-19 Outbreak Associated with Air Conditioning in Restaurant, Guangzhou, China, 2020*. Emerging Infectious Disease journal, 2020. **26**(7): p. 1628.
41. Shen, Y., et al., *Community Outbreak Investigation of SARS-CoV-2 Transmission Among Bus Riders in Eastern China*. JAMA Internal Medicine, 2020.
42. Lloyd-Smith, J.O., et al., *Superspreading and the effect of individual variation on disease emergence*. Nature, 2005. **438**(7066): p. 355-359.
43. Wong, G., et al., *MERS, SARS, and Ebola: The Role of Super-Spreaders in Infectious Disease*. Cell Host & Microbe, 2015. **18**(4): p. 398-401.
44. Fraser, C., et al., *Influenza transmission in households during the 1918 pandemic*. American journal of epidemiology, 2011. **174**(5): p. 505-514.
45. Kleiboeker, S., et al., *SARS-CoV-2 viral load assessment in respiratory samples*. Journal of Clinical Virology, 2020. **129**: p. 104439.
46. Pujadas, E., et al., *SARS-CoV-2 viral load predicts COVID-19 mortality*. The Lancet Respiratory Medicine, 2020. **8**(9): p. e70.
47. Sneppen, K., R.J. Taylor, and L. Simonsen, *Impact of Superspreaders on dissemination and mitigation of COVID-19*. medRxiv, 2020: p. 2020.05.17.20104745.

48. Kirkegaard, J.B. and K. Sneppen, *Variability of Individual Infectiousness Derived from Aggregate Statistics of COVID-19*. medRxiv, 2021: p. 2021.01.15.21249870.
49. Nielsen, B. F., Simonsen, L., Sneppen, K., *COVID-19 superspreading suggests mitigation by social network modulation. Physical Review Letters (to be published), 2021.*
50. Eilersen, A. and K. Sneppen, *COVID-19 superspreading in cities versus the countryside*. medRxiv, 2020: p. 2020.09.04.20188359.
51. Nielsen, B.F., et al., *Social network heterogeneity is essential for contact tracing*. medRxiv, 2020: p. 2020.06.05.20123141.
52. Endo A, Centre for the Mathematical Modelling of Infectious Diseases COVID-19 Working Group, Leclerc QJ et al. *Implication of backward contact tracing in the presence of overdispersed transmission in COVID-19 outbreaks [version 1; peer review: 2 approved]. Wellcome Open Res 2020, 5:239 (<https://doi.org/10.12688/wellcomeopenres.16344.1>).*
53. Li, X., et al., *The role of children in transmission of SARS-CoV-2: A rapid review*. Journal of global health, 2020. **10**(1): p. 011101-011101.
54. Munro, A.P.S. and S.N. Faust, *Children are not COVID-19 super spreaders: time to go back to school*. Archives of Disease in Childhood, 2020. **105**(7): p. 618-619.
55. Viner, R.M., et al., *Susceptibility to and transmission of COVID-19 amongst children and adolescents compared with adults: a systematic review and meta-analysis*. medRxiv, 2020: p. 2020.05.20.20108126.
56. Jones, T.C., et al., *An analysis of SARS-CoV-2 viral load by patient age*. medRxiv, 2020: p. 2020.06.08.20125484.
57. Chandrashekar, A., et al., *SARS-CoV-2 infection protects against rechallenge in rhesus macaques*. Science, 2020. **369**(6505): p. 812-817.
58. Ludvigsson, J.F., *Children are unlikely to be the main drivers of the COVID-19 pandemic – A systematic review*. Acta Paediatrica, 2020. **109**(8): p. 1525-1530.
59. Szablewski, C., et al., *SARS-CoV-2 Transmission and Infection Among Attendees of an Overnight Camp — Georgia, June 2020*. MMWR. Morbidity and Mortality Weekly Report, 2020. **69**.
60. <https://www.theguardian.com/world/2020/aug/21/coronavirus-iurope-dozens-schools-report-infections-berlin-germany-spain>.
61. <https://www.theguardian.com/world/2020/aug/14/school-reopenings-covid-19-coronavirus-us>.
62. Fontanet, A., et al., *Cluster of COVID-19 in northern France: A retrospective closed cohort study*. medRxiv, 2020: p. 2020.04.18.20071134.

63. Day, M., *Covid-19: More young children are being infected in Israel and Italy, emerging data suggest*. *BMJ*, 2021. **372**: p. n383.
64. Carsetti, R., et al., *The immune system of children: the key to understanding SARS-CoV-2 susceptibility?* *Lancet Child Adolesc Health*, 2020. **4**(6): p. 414-416.
65. Buonsenso, D., et al., *Preliminary Evidence on Long COVID in children*. medRxiv, 2021: p. 2021.01.23.21250375.
66. Denina, M., et al., *Sequelae of COVID-19 in Hospitalized Children: A 4-Months Follow-Up*. *The Pediatric Infectious Disease Journal*, 2020. **39**(12): p. e458-e459.
67. Nakra, N.A., et al., *Multi-System Inflammatory Syndrome in Children (MIS-C) Following SARS-CoV-2 Infection: Review of Clinical Presentation, Hypothetical Pathogenesis, and Proposed Management*. *Children*, 2020. **7**(7).
68. Radia, T., et al., *Multi-system inflammatory syndrome in children & adolescents (MIS-C): A systematic review of clinical features and presentation*. *Paediatric Respiratory Reviews*, 2020.
69. Phelan, A.L., *COVID-19 immunity passports and vaccination certificates: scientific, equitable, and legal challenges*. *The Lancet*, 2020. **395**(10237): p. 1595-1598.
70. <https://www.the-scientist.com/news-opinion/what-do-antibody-tests-for-sars-cov-2-tell-us-about-immunity--67425>.
71. Kreer, C., et al., *Longitudinal Isolation of Potent Near-Germline SARS-CoV-2-Neutralizing Antibodies from COVID-19 Patients*. *Cell*, 2020. **182**(4): p. 843-854.e12.
72. Pollán, M., et al., *Prevalence of SARS-CoV-2 in Spain (ENE-COVID): a nationwide, population-based seroepidemiological study*. *The Lancet*, 2020. **396**(10250): p. 535-544.
73. Ripperger, T.J., et al., *Detection, prevalence, and duration of humoral responses to SARS-CoV-2 under conditions of limited population exposure*. medRxiv, 2020: p. 2020.08.14.20174490.
74. Rodda, L.B., et al., *Functional SARS-CoV-2-specific immune memory persists after mild COVID-19*. medRxiv, 2020: p. 2020.08.11.20171843.
75. Wang, X., et al., *Neutralizing Antibody Responses to Severe Acute Respiratory Syndrome Coronavirus 2 in Coronavirus Disease 2019 Inpatients and Convalescent Patients*. *Clinical Infectious Diseases*, 2020.
76. Sekine, T., et al., *Robust T cell immunity in convalescent individuals with asymptomatic or mild COVID-19*. *Cell*, 2020.
77. Sette, A. and S. Crotty, *Pre-existing immunity to SARS-CoV-2: the knowns and unknowns*. *Nat Rev Immunol*, 2020. **20**(8): p. 457-458.

78. Bao, L., et al., *Reinfection could not occur in SARS-CoV-2 infected rhesus macaques*. bioRxiv, 2020: p. 2020.03.13.990226.
79. Addetia, A., et al., *Neutralizing antibodies correlate with protection from SARS-CoV-2 in humans during a fishery vessel outbreak with high attack rate*. medRxiv, 2020: p. 2020.08.13.20173161.
80. Pray IW, Gibbons-Burgener SN, Rosenberg AZ, et al. *COVID-19 Outbreak at an Overnight Summer School Retreat — Wisconsin, July–August 2020*. *MMWR Morb Mortal Wkly Rep* 2020;69:1600–1604. DOI: <http://dx.doi.org/10.15585/mmwr.mm6943a4>.
81. <https://www.businessinsider.com/2-new-coronavirus-reinfection-cases-belgium-netherlands-hong-kong-2020-8?r=US&IR=T>.
82. To, K.K.-W., et al., *COVID-19 re-infection by a phylogenetically distinct SARS-coronavirus-2 strain confirmed by whole genome sequencing*. *Clinical Infectious Diseases*, 2020.
83. Naveca et al.: *Phylogenetic relationship of SARS-CoV-2 sequences from Amazonas with emerging Brazilian variants harboring mutations E484K and N501Y in the Spike protein*, *Virological*, 2021.
84. Buss, L.F., et al., *Three-quarters attack rate of SARS-CoV-2 in the Brazilian Amazon during a largely unmitigated epidemic*. *Science*, 2021. **371**(6526): p. 288-292.
85. Sabino, E.C., et al., *Resurgence of COVID-19 in Manaus, Brazil, despite high seroprevalence*. *Lancet*, 2021.
86. <https://gbdeclaration.org/>.
87. <https://ourworldindata.org/>.
88. Davies, N.G., et al., *Estimated transmissibility and severity of novel SARS-CoV-2 Variant of Concern 202012/01 in England*. medRxiv, 2020: p. 2020.12.24.20248822.
89. Van Kerkhove, M.D., et al., *Estimating age-specific cumulative incidence for the 2009 influenza pandemic: a meta-analysis of A(H1N1)pdm09 serological studies from 19 countries*. *Influenza Other Respir Viruses*, 2013. **7**(5): p. 872-86.
90. Beigel, J.H., et al., *Remdesivir for the Treatment of Covid-19 — Preliminary Report*. *New England Journal of Medicine*, 2020.
91. Juul, S., et al., *Interventions for treatment of COVID-19: second edition of a living systematic review with meta-analyses and trial sequential analyses (The LIVING Project)*. medRxiv, 2020: p. 2020.11.22.20236448.
92. <https://investors.modernatx.com/news-releases/news-release-details/moderna-covid-19-vaccine-retains-neutralizing-activity-against/>.
93. Creech, C.B., S.C. Walker, and R.J. Samuels, *SARS-CoV-2 Vaccines*. *JAMA*, 2021.

94. Lavine, J.S., O.N. Bjornstad, and R. Antia, *Immunological characteristics govern the transition of COVID-19 to endemicity*. *Science*, 2021. **371**(6530): p. 741-745.

95. Pavelka, M., et al., *The effectiveness of population-wide, rapid antigen test based screening in reducing SARS-CoV-2 infection prevalence in Slovakia*. medRxiv, 2020: p. 2020.12.02.20240648.

96. https://experience.arcgis.com/experience/09f821667ce64bf7be6f9f87457ed9aa/page/page_0/.

Timeline of pandemics in the 20th and 21st century
COVID-19 (SARS-CoV-2) is the first coronavirus pandemic on record

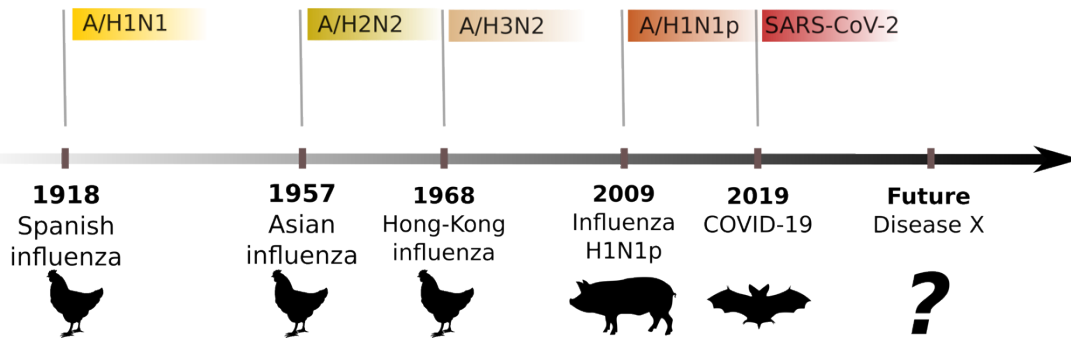


Figure 1 Timeline of respiratory viral pandemics in the 20th and 21st century. After a century of influenza A pandemics, a pandemic coronavirus emerged. The 1918, 1957 and 1968 pandemics are thought to have arisen from birds in Asia, whereas the 2009 originated in Mexican pigs. The origin of SARS-CoV-2 is thought to be Chinese bats. The colored labels indicate the pathogen responsible for the disease in question.

	Testing period	Seroprevalence estimate	Hospitalization rate	ICU rate	Infection fatality rate (IFR)
Danish seroprevalence study, round 2	17/8-4/9	2.2% (1.8-2.6%)	2.2% (1.9-2.7%)	--	0.49% (0.41-0.60%)
Danish seroprevalence study, round 3	14/12-8/1	3.9% (3.3-4.6%)	3.0% (2.6-3.6%)	--	0.55% (0.46-0.65%)
Danish blood donors, week 4 of 2021	25/1-29/1	8.1% (6.9-8.9%)	2.4% (2.2-2.8%)	--	0.46% (0.42-0.54%)
Spanish data, ENECOVID	27/4-11/5	6.1%*	2.59%	0.24%	0.85%
Data from Iceland	Post first wave seroprevalence	0.9% (0.8-0.9%)	3.6%	0.9%	0.3%

Table 1 shows

estimates of the proportion of all infected individuals who are hospitalized, admitted to the ICU and die. We base our estimates of the number of infected individuals by inferring from seroprevalence studies[19, 21-24]

*We have adjusted the crude Spanish estimate of 5.0% for estimated sensitivity (82.1%) and specificity (100%) of the used IgG POCT.

Location	Event type and comments	Date (duration)	Estimated number of secondary infections from one superspreader	Participants	Attack rate
Skagit County, USA	Choir practice with social distancing transmission*	March 10 (2.5 hours)	52	61	87%
Calgary, Australia	Service and party in a church with social distancing*	Mid-March (a few hours)	23	41	59%
Guangzhou, China	Restaurant, asymptomatic superspreader **	January 24 (one lunch period)	9	91	11%
Edmonton, Canada	Bonspiel (curling event)	March 11-14 (4 days)	23**	72***	33%
Chicago, USA	A dinner, a funeral and a birthday party	Feb-March (three distinct events)	10	-	-

Zhejiang province, China	Bus ride and worship event (WE)*	Bus ride: 100 mins WE: 150 mins	Bus 1: 0 Bus 2: 23 WE, others: 7 WE, total: 30	Bus 1: 60 Bus 2: 68 WE, others: 172 WE, total: 300	Bus 1: 0 % Bus 2: 35% WE, others: 4% WE, total: 10%
--------------------------	----------------------------------	------------------------------------	---	---	--

Table 2 Shows examples[36-41] of evident COVID-19 superspreader events, meaning that they occurred in a limited time period so that it most likely represents multiple secondary infection from a single superspreader.

A long list of 1400 outbreaks is available in the following database:

<https://docs.google.com/spreadsheets/d/1c9jwMyT1lw2P0d6SDTno6nHLGMtpheO9xJyGHgdBoco/edit>

*Highly probable case of aerosolized transmission

**High probability of at least some tertiary infections

***“Roughly 72 attendees”

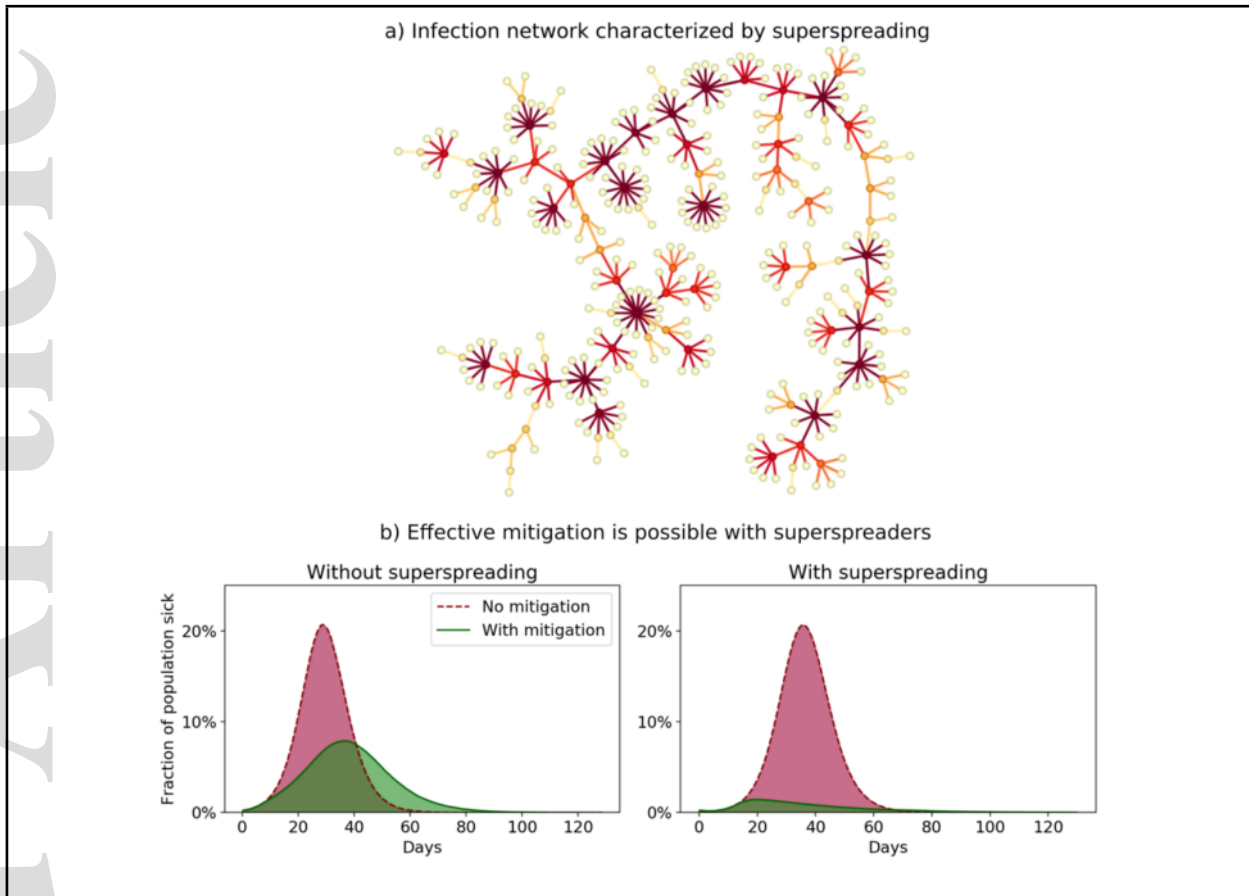


Figure 2. Simulations of an agent-based model with network and superspreaders (see full model and assumptions in [47]). a) A single infection tree – the result of a model simulation of superspreading. The epidemic spreads due to a small proportion of individuals who are highly infectious, while the majority do not transmit the disease.

b) Effect of mitigating in the public domain to reduce opportunities for superspreading. If a sizable proportion of infections are caused by superspreaders, the simulations show that just reducing contacts in the public space (that is, outside households and workplaces/schools), has a large mitigation effect (right subpanel); but without superspreaders in the model, not much is gained (left subpanel). Data for panel b from [47].

In these simulations, superspreaders are individuals with a higher personal reproductive number, thus having the potential to transmit the disease to many in an unmitigated scenario. Drastically reducing the number of *different* persons that one meets (by e.g., banning large gatherings) thus has an outsized effect in a disease characterized by superspreading, providing an opportunity for improved mitigation. The theoretical background for this effect is explored in ref. [49].

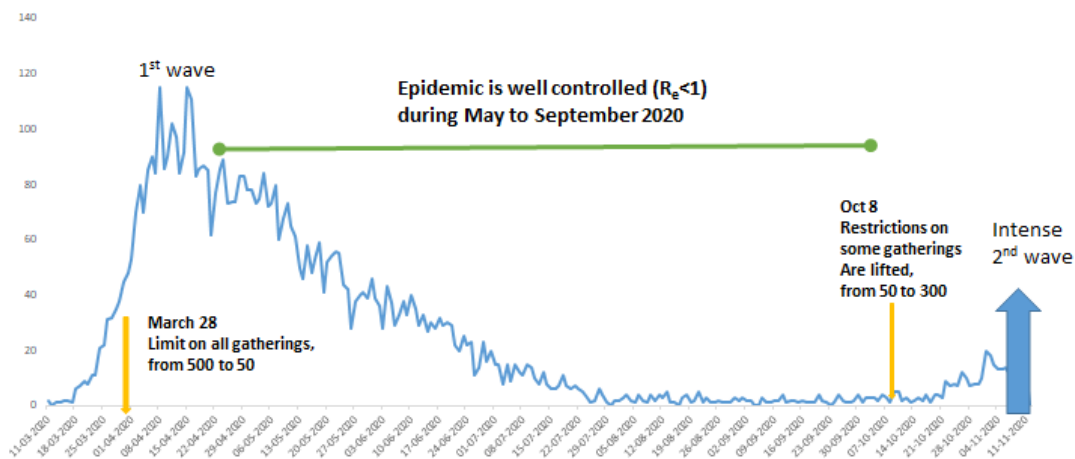


Figure 3 A sustainable control strategy in Sweden? On March 28th, Sweden introduced a ban on events >50 persons and the daily numbers of deaths started to decline a few weeks after[96]. On October 8 some gatherings were again allowed up to 300. Many other factors were in effect in Sweden, including working from home, less traveling, more effective shielding of the elderly, closed universities and the seasonal changes in temperature and humidity. But borders remained open, as did schools for children up to 16 years of age in this time period.

	COVID-19 Hospitalizations	% of population hospitalized	COVID-19 Deaths	% of population dead
Estimates based on seroprevalence study, round 2	76,500 (64731-93500)	1.3% (1.1-1.6%)	17,100 (14469-20900)	0.29% (0.25-0.36%)
Estimates based on seroprevalence study, round 3	105,538 (89478-124727)	1.8% (1.5-2.1%)	19,154 (16239-22636)	0.33% (0.28-0.39%)
Estimates based on Danish blood donor serology study; week 4, 2021	82,993 (75533-97426)	1.4% (1.3-1.7%)	16,059 (14616-18852)	0.28% (0.25-0.32%)

Table 3 Shows the hypothetical cost of controlled, natural herd immunity in Denmark in terms of deaths and hospitalizations. The resulting figures are far greater than the current cumulative burden of ~2,300 deaths and ~12,000 hospitalizations in our country (as of Feb 16, 2021).

ADDITIONAL PUBLICATIONS

In addition to the publications included in the preceding chapters, I published one additional paper in *Physical Review E* during my PhD programme [71]. This manuscript is **not** included as part of the works submitted for evaluation, but merely included for completeness.

NEWTON-CARTAN SUBMANIFOLDS AND FLUID MEMBRANES

Authors: Jay Armas^{2,3}, Jelle Hartong⁴, Emil Have⁴, Bjarke Frost Nielsen¹, and Niels A. Obers^{1,5}.

¹ The Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark.

² Institute for Theoretical Physics, University of Amsterdam, Amsterdam, the Netherlands.

³ Dutch Institute for Emergent Phenomena, Amsterdam, the Netherlands.

⁴ School of Mathematics and Maxwell Institute for Mathematical Sciences, University of Edinburgh, Edinburgh, United Kingdom.

⁵ Nordita, KTH Royal Institute of Technology, and Stockholm University, Stockholm, Sweden

My contribution: Contributed to conceptualization, theory development, creating figures and writing of the manuscript.

Publication status: Published in *Physical Review E* (2020).

Hyperlink(s): <https://doi.org/10.1103/PhysRevE.101.062803>

Newton-Cartan submanifolds and fluid membranes

Jay Armas^{1,*}, Jelle Hartong^{2,†}, Emil Have^{2,‡}, Bjarke F. Nielsen^{3,§} and Niels A. Obers^{4,3,||}

¹*Institute for Theoretical Physics, University of Amsterdam, 1090 GL Amsterdam, the Netherlands, and Dutch Institute for Emergent Phenomena, 1090 GL Amsterdam, the Netherlands*

²*School of Mathematics and Maxwell Institute for Mathematical Sciences, University of Edinburgh, Peter Guthrie Tait Road, Edinburgh EH9 3FD, United Kingdom*

³*Niels Bohr Institute, University of Copenhagen Blegdamsvej 17, DK-2100 Copenhagen Ø, Denmark*

⁴*Nordita, KTH Royal Institute of Technology, and Stockholm University, Roslagstullsbacken 23, SE-106 91 Stockholm, Sweden*



(Received 9 January 2020; accepted 15 May 2020; published 18 June 2020)

We develop the geometric description of submanifolds in Newton-Cartan spacetime. This provides the necessary starting point for a covariant spacetime formulation of Galilean-invariant hydrodynamics on curved surfaces. We argue that this is the natural geometrical framework to study fluid membranes in thermal equilibrium and their dynamics out of equilibrium. A simple model of fluid membranes that only depends on the surface tension is presented and, extracting the resulting stresses, we show that perturbations away from equilibrium yield the standard result for the dispersion of elastic waves. We also find a generalization of the Canham-Helfrich bending energy for lipid vesicles that takes into account the requirements of thermal equilibrium.

DOI: [10.1103/PhysRevE.101.062803](https://doi.org/10.1103/PhysRevE.101.062803)

I. INTRODUCTION

The dynamics of surfaces and interfaces plays a prominent role in various instances of physical phenomena, ranging from fluid membranes in biological systems [1,2], to the interplay between liquid crystal geometry and hydrodynamics [3], to surface or edge physics in condensed matter systems [4]. Fluid membranes comprising lipid bilayers are essential in the physics of biological systems, and the characterization of their geometric properties has been an active field of research for decades, as well as being key in understanding experimental outcomes (see, e.g., Refs. [5–9] for reviews). Hydrodynamics on curved surfaces has also recently received considerable attention, not only due to its relevance in embryonic processes [10] or cell migration [11] where activity also plays a role, but also due to its relevance in understanding topological properties of wave dynamics such as Kelvin-Yanai waves on the Earth's equator [12], flocking on a sphere [13], or turbulence in active nematics [14–16].

While the geometry and dynamics of surfaces in (pseudo)-Riemannian geometry has been deeply studied in both physics and mathematics, a systematic treatment using covariant and geometrical structures has so far not been developed for Galilean-invariant systems. In view of the relevance of such systems in many branches of physics, and immediate applications in biophysical systems detailed below, the main goal of this paper is to develop the theory of submanifolds in Newton-Cartan spacetime. This can be considered as the Galilean

analog of the (pseudo)-Riemannian case for which the geometry and its embeddings have local Euclidean (Poincaré) symmetry as opposed to Galilean symmetries. The formalism we develop allows for a covariant spacetime formulation of Galilean-invariant hydrodynamics on curved surfaces.

As such it is thus the natural framework to study fluid membranes in thermal equilibrium along with their dynamics away from equilibrium. This includes in particular biophysical membranes such as lipid bilayers, which are membranes composed of lipid molecules that enclose the cytoplasm. The lipid molecules move as a fluid along the membrane surface, which itself behaves elastically when bent. It is well known that at mesoscopic scales, lipid bilayers can be approximated by thin surfaces whose equilibrium configurations are accurately described by geometrical degrees of freedom and a small set of material coefficients that encode the more microscopic biochemical details (see, e.g., Ref. [9]). The shapes of lipid bilayers, such as discoids characterizing the morphology of red blood cells, are found by extremizing the Canham-Helfrich (CH) free energy [5,6], which depends only on geometric properties. The stresses associated to such bilayers have received considerable attention [9,17] as well as deformations of the CH free energy away from equilibrium in order to identify stable deformations [18].

However, despite the CH free energy being taken to represent a system in thermodynamic equilibrium [19] (as well as its analog in nematic liquid crystals, the Frank energy [20]), it disregards the basic lesson of equilibrium thermal field theory: that temperature and mass chemical potential (conjugate to particle number) also have a geometric interpretation. This results in the CH free energy giving rise to inaccurate stresses characterizing the membrane, explicit by the fact that they do not describe the stresses intrinsic to a fluid, and neither do they yield elastic wave dispersion relations when deforming away from equilibrium. In this paper, we argue that the development

*j.armas@uva.nl

†j.hartong@ed.ac.uk

‡emil.have@ed.ac.uk

§bjarkenielson@nbi.ku.dk

||obers@nbi.ku.dk

of a spacetime covariant formulation of Galilean-invariant hydrodynamics using Newton-Cartan geometry is a more useful approach to understanding fluid dynamics on curved surfaces and the physics of equilibrium fluid membranes.

Newton-Cartan (NC) geometry was pioneered by Cartan in order to geometrize Newton's theory of gravity [21,22].¹ As a non-dynamical geometry its importance stems from the fact that it is the natural background geometry that nonrelativistic field theories couple to [25,26]² and thus provides a geometric and covariant formulation of many aspects of nonrelativistic physics including broad classes of long-wavelength effective theories such as hydrodynamics. In particular, in the past few years NC geometry and variants have been applied to the formulation of Galilean-invariant fluid dynamics [33,34], Lifshitz fluid dynamics [35,36] as well as hydrodynamics without boost symmetry [37–40],³ which encapsulate the former as cases with extra symmetries. Furthermore, in the context of condensed matter systems, it was realized that NC geometry is the natural setting for developing an effective theory of the fractional quantum Hall effect [41–44]. This body of work, together with previous work on Galilean superfluid droplets [45] and connections between black holes and CH functionals [46,47], suggests that NC geometry can also be useful in describing hydrodynamics on curved surfaces.

The development of submanifold calculus in (pseudo-)Riemannian or Euclidean geometry, written in multiple volumes (e.g., Ref. [48]) and furthered in different contexts [49–53], is an essential prerequisite for describing surfaces and hence for formulating and extremizing the CH free energy. Therefore, the majority of the work presented in this paper, in particular Secs. II and III and Appendix A, consists of the development of submanifold calculus in Newton-Cartan geometry, the identification of geometrical properties describing surfaces, and the formulation of appropriate geometric functionals whose extrema are NC surfaces. Thus, the main part of the work presented here is foundational. However, in Sec. IV we apply this machinery to different fluid membrane systems in order to show its usefulness and provide a generalized CH model that takes into account the requirements of thermodynamic equilibrium. The work developed here will be the basis for a more detailed study of effective theories of fluid membranes, which takes into account a larger set of responses including viscosity, providing a more solid foundation for the physics of fluid membranes [54].

¹See also Ref. [23] for a modern perspective and earlier references, and the recent work [24] for an action principle for Newtonian gravity.

²In particular, the most general coupling requires a torsionful generalization of NC geometry, called torsional Newton-Cartan (TNC) geometry which was first observed as the boundary geometry in the context of Lifshitz holography [27–29]. TNC geometry also appears as the ambient space-time for nonrelativistic strings; see, e.g., Refs. [30–32].

³The boost-noninvariant hydrodynamics of these papers is formulated in the regime where momentum is conserved, but may be generalized to include further breaking of translation symmetry, in which case it applies to flocking and active matter.

A. Organization of the paper

A more detailed outline of the paper, including a brief summary of the main results is as follows.

In Sec. II, after reviewing the geometric structure of a Newton-Cartan spacetime, we first define what a submanifold structure is in such spacetimes. In particular, we develop the necessary geometric tools to define an induced NC structure on the submanifold. We highlight in particular how the objects transform under local Galilean boosts, which is a key property for nonrelativistic geometries. We then show, using the affine connection that is known for NC structures, how to construct a covariant derivative along the surface directions, and give an expression for the corresponding surface torsion tensor. With this in hand, we discuss the exterior curvature and show how the (Riemannian) Weingarten identity gets modified in this case.

Section III develops the variational calculus for NC submanifolds, which is essential technology in order to find equations of motion from effective actions. We consider first general variations of the relevant quantities describing the embedding. Subsequently we obtain expressions for embedding map variations as well as Lagrangian variations, which are diffeomorphisms in the ambient NC spacetime that keep the embedding maps fixed. From the corresponding variations of the induced NC structures and the normal vectors we find in particular how the extrinsic curvature transforms under such variations. We subsequently use this technology to consider the dynamics of submanifolds that arises from extremization of an action. The resulting equations of motions for NC submanifolds are thus obtained from the general response to varying the induced NC metric structure on the manifold and the extrinsic curvature. These split up in a set of intrinsic equations, which are conservation equations of the world-volume stress tensor and mass current accompanied by a set of extrinsic equations. We also analyze the boundary terms that appear as a result of varying the general action functional and obtain the resulting boundary conditions.

Then in Sec. IV we apply the action formalism presented in the previous section to describe equilibrium fluid membranes and lipid vesicles as well as their fluctuations. We will show that employing NC geometry for such surfaces is not only natural but also provides a more complete description. First, it introduces (absolute) time and therefore fluctuations of the system can include temporal dynamics in a covariant form. Moreover, the symmetries of the problem are made manifest via the geometry of the submanifold and ambient spacetime. Even more important is the aspect that NC geometry allows to properly introduce thermal field theory of equilibrium fluid membranes. To illustrate all this we first consider equilibrium fluid membranes, i.e., stationary fluid configurations on an arbitrary surface and the simplest example with a free energy depending on surface tension only, for which we compute the resulting stresses. We then show that perturbations away from equilibrium yield the standard result for the dispersion of elastic waves. We also briefly consider the case of a droplet, by adding internal or external pressure to the previous case. Then we revisit the celebrated Canham-Helfrich model, which describes equilibrium configurations of biophysical membranes. We show how this model can be described using

Newton-Cartan geometry and generalize it by allowing its (material) parameters to depend on temperature and chemical potential. Finally, we review the classic lipid vesicles using this framework.

We end in Sec. V with a brief discussion and description of further avenues of investigation.

A number of appendices are included containing further details. Since it is known that torsional NC spacetimes can be obtained from Lorentzian spacetime using null reduction, we show in Appendix A a complimentary perspective on NC submanifolds, by null reducing submanifolds of Lorentzian spacetimes. Appendix B describes different classes of NC spacetimes, depending on properties of the torsion. In Appendix C we find the relation between the NC connections of the ambient spacetime and the submanifold (described in Sec. II B 5). Finally, in Appendix D we show how the Gauss–Bonnet theorem reduces the number of independent terms in an effective action for $(2 + 1)$ -dimensional membranes that appear as closed codimension one surfaces embedded in flat $(3 + 1)$ -dimensional Newton-Cartan geometry.

II. THE GEOMETRY OF NEWTON-CARTAN SUBMANIFOLDS

This section is devoted to a proper geometrical treatment of surfaces (or embedded submanifolds) in NC geometry with the goal of subsequently applying it to the description of membrane elasticity and fluidity in later sections. To that aim, we begin by introducing the reader to the essential details of NC geometry. The basic structures that define a given NC geometry are then understood as background fields for the dynamical surfaces or objects, in direct analogy with embedding of surfaces in a (pseudo-)Riemannian geometry with background metric $g_{\mu\nu}$. This paves the way for defining the geometric structures that characterize nonrelativistic surfaces.⁴ In Appendix A we provide an alternative method for obtaining the theory of NC surfaces directly from the theory of surfaces in Lorentzian geometry.

A. Newton-Cartan geometry

Let \mathcal{M}_{d+1} be a $(d + 1)$ -dimensional manifold endowed with a Newton-Cartan structure, which consists of the fields $(\tau_\mu, h_{\mu\nu}, m_\mu)$. Here the Greek indices denote spacetime indices such that $\mu, \nu, \dots = 0, \dots, d$. The tensor $h_{\mu\nu}$ is symmetric with rank d and has signature $(0, 1, 1, \dots)$, while the nowhere vanishing 1-form τ_μ is such that $-\tau_\mu\tau_\nu + h_{\mu\nu}$ has full rank. The field m_μ is the connection of an Abelian gauge symmetry that from the point of view of a Galilean field theory on a NC spacetime can be thought of as the symmetry underlying particle number conservation. Since the latter is

a compact Abelian symmetry we refer to m_μ as the $U(1)$ gauge connection. It is useful to define an inverse NC structure $(v^\mu, h^{\mu\nu})$, where v^μ spans the kernel of $h_{\mu\nu}$ and τ_μ spans the kernel of $h^{\mu\nu}$. The 1-form τ_μ is sometimes called the *clock 1-form*, while the vector v^μ is known as the *Newton-Cartan velocity*. These structures satisfy the completeness relation and normalization condition:

$$\delta_v^\mu = -v^\mu\tau_\nu + h^{\mu\rho}h_{\rho\nu}, \quad \text{so that} \quad v^\mu\tau_\mu = -1. \quad (2.1)$$

It is occasionally useful to introduce vielbeins $e_{\underline{a}}^\mu, e_{\underline{a}}^\mu$ with $\underline{a}, \underline{b}, \dots = 1, \dots, d$ (that is, spatial tangent space indices are underlined lowercase Latin letters) such that

$$h_{\mu\nu} = \delta_{\underline{a}\underline{b}}e_{\underline{a}}^\mu e_{\underline{b}}^\nu, \quad h^{\mu\nu} = \delta^{\underline{a}\underline{b}}e_{\underline{a}}^\mu e_{\underline{b}}^\nu, \quad (2.2)$$

which furthermore satisfy the orthogonality relations

$$v^\mu e_{\underline{a}}^\mu = 0, \quad \tau_\mu e_{\underline{a}}^\mu = 0, \quad e_{\underline{a}}^\mu e_{\underline{b}}^\mu = \delta_{\underline{a}\underline{b}}. \quad (2.3)$$

The Newton-Cartan structure on \mathcal{M}_{d+1} in terms of the fields $(\tau_\mu, h_{\mu\nu}, m_\mu)$ transforms under diffeomorphisms (coordinate transformations), $U(1)$ (mass) gauge transformations (akin to gauge transformations in Maxwell theory), local rotations and local Galilean boosts (also known as Milne boosts) in the following way:

$$\begin{aligned} \delta\tau_\mu &= \xi_\xi\tau_\mu, & \delta e_{\underline{a}}^\mu &= \xi_\xi e_{\underline{a}}^\mu + \lambda_{\underline{a}}^b e_{\underline{b}}^\mu + \lambda_{\underline{a}}^a \tau_\mu, \\ \delta m_\mu &= \xi_\xi m_\mu + \lambda_{\underline{a}} e_{\underline{a}}^\mu + \partial_\mu \sigma, & & \\ \delta v^\mu &= \xi_\xi v^\mu + \lambda^a e_{\underline{a}}^\mu, & \delta e_{\underline{a}}^\mu &= \xi_\xi e_{\underline{a}}^\mu + \lambda_{\underline{a}}^b e_{\underline{b}}^\mu. \end{aligned} \quad (2.4)$$

Here ξ^μ is the generator of diffeomorphisms, σ is the parameter of mass gauge transformations, and λ^a is the parameter of local Galilean boosts. Finally, $\lambda_{\underline{a}}^b = -\lambda_{\underline{b}}^a$ corresponds to local $\mathfrak{so}(d)$ transformations. When describing physical systems in NC geometry by means of a Lagrangian or action functional, one requires invariance under the gauge transformations (2.4). In the restricted setting of a flat NC background (i.e., a spacetime with absolute time whose constant time slices are described by Euclidean geometry), which is the most relevant case in the context of biophysical membranes, invariance under (2.4) implies invariance under global Galilean symmetries centrally extended to include mass conservation. The centrally extended Galilei group is known as the Bargmann group. This implies that the geometry can be viewed as originating from “gauging” the Bargmann algebra as detailed in Ref. [23].

1. Galilean boost-invariant structures

One may readily check that given (2.4), the NC fields $h^{\mu\nu}$ and $h_{\mu\nu}$, which are constructed out of the vielbeins as in (2.2), transform as

$$\delta h^{\mu\nu} = \xi_\xi h^{\mu\nu}, \quad \delta h_{\mu\nu} = \xi_\xi h_{\mu\nu} + 2\lambda_{(\mu}\tau_{\nu)}, \quad (2.5)$$

where $\lambda_\mu = e_{\underline{a}}^\mu \lambda_{\underline{a}}$, immediately implying that $\lambda_\mu v^\mu = 0$. We conclude from this that $h^{\mu\nu}\partial_\mu\partial_\nu$ is an invariant of the geometry, a cometric, while $h_{\mu\nu}dx^\mu dx^\nu$ is not an invariant because it transforms under the Galilean boosts. On the other hand $\tau_\mu dx^\mu$ is invariant. This means that NC geometry has a

⁴Intuition originating from the description of surfaces in (pseudo-)Riemannian geometry suggests that geometric structures characterizing surfaces in NC geometry would naively be constructed from pullbacks of NC ambient spacetime fields. It will turn out that this is only true for submanifolds of NC geometry provided we take the pullbacks of quantities that are invariant under the local Galilean boost transformations of the ambient NC geometry.

degenerate metric structure given by $\tau_\mu \tau_\nu$ and $h^{\mu\nu}$ and that $h_{\mu\nu}$ should not be viewed as a metric.⁵

Notice that while $h_{\mu\nu}$ transforms under Galilean boosts it does not transform under $U(1)$ gauge transformations. It is possible to define objects that have the opposite property, namely, that they are Galilean boost invariant but not $U(1)$ invariant. We will often work with these fields, and so we discuss their construction here. We can trade $U(1)$ gauge invariance for boost invariance by introducing the set of fields

$$\bar{h}_{\mu\nu} = h_{\mu\nu} - 2\tau_{(\mu} m_{\nu)}, \quad \hat{v}^\mu = v^\mu - h^{\mu\nu} m_\nu, \quad (2.6)$$

which transform as⁶

$$\delta \bar{h}_{\mu\nu} = \xi_\xi \bar{h}_{\mu\nu} - 2\tau_{(\mu} \partial_{\nu)} \sigma, \quad \delta \hat{v}^\mu = \xi_\xi \hat{v}^\mu - h^{\mu\nu} \partial_\nu \sigma, \quad (2.7)$$

and hence are manifestly Galilean boost invariant. Additionally, it is also possible to construct a boost-invariant scalar, which is the boost-invariant counterpart of the Newtonian potential [56], namely,

$$\tilde{\Phi} = -v^\mu m_\mu + \frac{1}{2} h^{\mu\nu} m_\mu m_\nu. \quad (2.8)$$

The Newtonian potential itself is just the time component of m_μ . These quantities will be useful when discussing effective actions for fluid membranes in later sections.

2. Covariant differentiation and affine connection

NC geometry provides a way of formulating nonrelativistic physics in curved backgrounds and substrates which has recently become an active research direction in soft matter [12–16]. Additionally, even in the traditional case of lipid membranes sitting in Euclidean space, it is useful to have explicit coordinate independence as it can simplify many problems of interest. Therefore, it is important to introduce a covariant derivative adapted to curved backgrounds. However, in contrast to (pseudo-)Riemannian geometry without torsion, there is no unique metric-compatible connection in Newton-Cartan geometry. Rather, the analog of metric compatibility in NC geometry is

$$\nabla_\mu \tau_\nu = 0, \quad \nabla_\mu h^{\nu\rho} = 0, \quad (2.9)$$

where ∇ is the covariant derivative with respect to the affine connection $\Gamma_{\mu\nu}^\rho$. It is possible to choose the affine connection

⁵We can fix diffeomorphisms such that $\tau_i = 0$ where we split the spacetime coordinates $x^\mu = (t, x^i)$. In this restricted gauge the metric on slices of constant time t is given by $h_{ij} dx^i dx^j$ which is invariant under the diffeomorphisms that do not affect time. In this sense the constant time slices are described by standard Riemannian geometry. However, when we include time into the formalism we have to abandon the notion of a metric and instead work with the NC triplet $(\tau_\mu, h_{\mu\nu}, m_\mu)$. In this setting, in order to evaluate areas or volumes of given surfaces one can use the integration measure $e = \sqrt{-\det(-\tau_\mu \tau_\nu + h_{\mu\nu})}$, which is both Galilean boost and $U(1)$ invariant.

⁶Note that this is possible because the $U(1)$ connection m_μ also transforms under Galilean boosts. In this sense it is different from the Maxwell potential. The difference comes from the fact that the mass generator forms a central extension of the Galilei algebra whereas the charge $U(1)$ generator of Maxwell's theory forms a direct sum with in that case the Poincaré algebra. See Refs. [23,55] for more details.

as [57,58]⁷

$$\Gamma_{\mu\nu}^\rho = -\hat{v}^\rho \partial_\mu \tau_\nu + \frac{1}{2} h^{\rho\sigma} (\partial_\mu \bar{h}_{\nu\sigma} + \partial_\nu \bar{h}_{\mu\sigma} - \partial_\sigma \bar{h}_{\mu\nu}). \quad (2.10)$$

Given the connection Γ , covariant differentiation acts on an arbitrary vector X^μ in a similar manner as in (pseudo-)Riemannian geometry:

$$\nabla_\mu X^\nu = \partial_\mu X^\nu + \Gamma_{\mu\rho}^\nu X^\rho. \quad (2.11)$$

Notably, and in contradistinction to the Levi-Civita connection of (pseudo-)Riemannian geometry, the connection $\Gamma_{\mu\nu}^\lambda$ is generally torsionful. This is due to the condition $\nabla_\mu \tau_\nu = 0$. In particular, the affine connection has an antisymmetric part given by

$$2\Gamma_{[\mu\nu]}^\lambda = -2\hat{v}^\lambda \partial_{[\mu} \tau_{\nu]} = -\hat{v}^\lambda \tau_{\mu\nu}, \quad (2.12)$$

where we defined the torsion 2-form

$$\tau_{\mu\nu} = 2\partial_{[\mu} \tau_{\nu]}. \quad (2.13)$$

For all physical systems studied in this paper, the torsion vanishes. However, when performing variational calculus (of the NC fields) it is required to keep variations of τ_μ arbitrary.⁸

As written in (2.10) in terms of boost-invariant quantities, the affine connection does not transform under Galilean boosts. However, under the $U(1)$ gauge transformations (2.7), it transforms as

$$\delta_\sigma \Gamma_{\mu\nu}^\rho = \frac{1}{2} h^{\rho\lambda} (\tau_{\mu\nu} \partial_\lambda \sigma + \tau_{\lambda\nu} \partial_\mu \sigma + \tau_{\lambda\mu} \partial_\nu \sigma). \quad (2.14)$$

In the absence of torsion, $\tau_{\mu\nu} = 0$, the connection is invariant under such transformations.

3. Absolute time and flat space

Depending on the conditions imposed on the clock 1-form τ_μ , there are different classes of NC geometries [28,58]. We refer the curious reader to Appendix B, which contains a classification of the different classes NC geometries, while in this section we focus on the most relevant case for the purposes of this work. If τ_μ is exact, that is, $\tau_\mu = \partial_\mu T$ for some scalar T , the torsion (2.13) vanishes and we are dealing with Newtonian absolute time. This is the simplest kind of Newton-Cartan geometry and the relevant one for the applications we consider in this work, namely, lipid vesicles or fluid membranes. For example, for membrane geometries, which for each instant in time are embedded in three-dimensional Euclidean space, the ambient NC spacetime in Cartesian coordinates can be parametrized as

$$\begin{aligned} \tau_\mu &= \delta_\mu^0, & h_{\mu\nu} &= \delta_\mu^i \delta_\nu^i, & v^\mu &= -\delta_0^\mu, & h^{\mu\nu} &= \delta_i^\mu \delta_i^\nu, \\ m_\mu &= 0. \end{aligned} \quad (2.15)$$

⁷As shown in Refs. [57,58], the most general affine connection satisfying (2.9) takes the form $\Gamma_{\mu\nu}^\rho = \Gamma_{\mu\nu}^\rho + W_{\mu\nu}^\rho$ where $W_{\mu\nu}^\rho$ is the pseudocontortion tensor, obeying $\tau_\rho W_{\mu\nu}^\rho = 0$ and $W_{\mu\lambda}^\nu h^{\lambda\rho} + W_{\mu\lambda}^\rho h^{\nu\lambda} = 0$. The choice (2.10) corresponds to $W_{\mu\nu}^\rho = 0$. This choice is also the natural choice from the perspective of the Noether procedure [55].

⁸The condition that τ_μ be unconstrained is not necessary when we perform variations of embedding scalars in a fixed ambient space geometry.

In the context of nonrelativistic physics in spatially curved backgrounds, the clock 1-form will still have the form $\tau_\mu = \delta_\mu^0$ but the tensor $h_{\mu\nu}$ can be nontrivial in the sense that it is not gauge equivalent to flat space. Thus for all practical applications, the first term in the affine connection (2.10) vanishes and the connection is purely spatial. However, while for physically relevant spacetimes we will always require that τ_μ must be of the form $\tau_\mu = \partial_\mu T$, when we are dealing with τ_μ as a background source in some action functional for matter fields, we need to require that it is unconstrained in order to be able to vary it freely.

B. Submanifolds in Newton-Cartan geometry

In this section we formulate the theory of nonrelativistic NC timelike⁹ surfaces (or submanifolds) embedded in arbitrary NC geometries. Following the literature that deals with the relativistic counterpart [53], we focus on the description of a single surface placed in an ambient NC spacetime and not on a foliation of such surfaces. In practice, this means that all geometric quantities, such as tangent and normal vectors, describing the surface are only well defined on the surface and not away from it. In this section we introduce the necessary geometrical structures for dealing with a single surface in a NC spacetime.

1. Embedding map, tangent, and normal vectors

A $(p+1)$ -dimensional Newton-Cartan submanifold Σ_{p+1} of a $(d+1)$ -dimensional Newton-Cartan manifold \mathcal{M}_{d+1} is specified by the embedding map

$$X^\mu : \Sigma \rightarrow \mathcal{M}, \mu = 0, \dots, d, \quad (2.16)$$

which maps the coordinates σ^a on Σ_{p+1} to $X^\mu(\sigma^a)$ on \mathcal{M} (lowercase Latin letters, $a, b, \dots = 0, \dots, p$, denote submanifold spacetime indices). Concretely, the embedding map specifies the location of the surface as $x^\mu = X^\mu(\sigma^a)$ where x^μ are coordinates in \mathcal{M} . The manifold \mathcal{M} into which the embedding scalars map is usually referred to as the target spacetime. The manifold described by the spacetime coordinates x^μ is the ambient spacetime. For simplicity, we will refer to both as ambient spacetime.

Given the embedding map, the tangent vectors to the surface are explicitly defined via $u_a^\mu = \partial_a X^\mu$. In turn, the normal 1-forms $n_\mu^I dx^\mu$ (where I runs over the $d-p$ transverse directions) are implicitly defined via the relations

$$n_\mu^I u_a^\mu = 0, \quad h^{\mu\nu} n_\mu^I n_\nu^J = \delta^{IJ}, \quad I = 1, \dots, d-p. \quad (2.17)$$

This normalization implies that in the normal directions we can use δ_{IJ} and δ^{IJ} to raise and lower transverse indices, meaning that we can write $Y_I Y^I = Y^I Y^I$ for some arbitrary vector Y^I . However, Eq. (2.17) does not fix the normal 1-forms uniquely. In fact, the 1-forms n_μ^I transform under local $\text{SO}(d-p)$ rotations such that

$$n_\mu^I \rightarrow \mathcal{M}^I_J n_\mu^J, \quad (2.18)$$

where \mathcal{M}^I_J is an element of $\text{SO}(d-p)$. The transformation (2.18) leaves (2.17) invariant and hence expresses the freedom of choosing the normal 1-forms.¹⁰

We can furthermore introduce “inverse objects” u_μ^a and n_I^μ to the tangent vectors and normal 1-forms via the completeness relation

$$\delta_\nu^\mu = u_\mu^a u_\nu^a + n_I^\mu n_\nu^I, \quad (2.19)$$

which in turn satisfy the relations

$$u_\mu^a n_I^\mu = 0, \quad u_a^\mu u_\mu^b = \delta_a^b, \quad n_I^\mu n_\mu^J = \delta_I^J. \quad (2.20)$$

The tangent vectors, normal 1-forms and their inverses can be used to project any tensor tangentially or orthogonally to the surface. For instance, we may project some tensor $X^\mu{}_{\nu\rho}{}^\lambda$ and denote the result as

$$X^a{}_{Ib}{}^J = u_\mu^a n_I^\nu u_b^\rho n_\rho^J X^\mu{}_{\nu\rho}{}^\lambda. \quad (2.21)$$

It is also useful to define the tangential spacetime projector

$$P_\nu^\mu = u_\mu^a u_\nu^a = \delta_\nu^\mu - n_I^\mu n_\nu^I, \quad (2.22)$$

which can be shown to be idempotent and of rank $p+1$. The object (2.22) can be used to project arbitrary tensors onto tangential directions along the surface and satisfies $P_\nu^\mu n_\mu^I = 0$.

2. Timelike submanifolds and boost invariance

Our goal is formulate a theory of nonrelativistic submanifolds Σ_{p+1} characterized by a Newton-Cartan structure that is inherited from the NC structure of the ambient spacetime. We introduce the submanifold clock 1-form as the pullback of the clock 1-form of the ambient spacetime such that

$$\tau_a = u_a^\mu \tau_\mu. \quad (2.23)$$

As mentioned earlier, we focus on timelike submanifolds, by which we mean that the normal vectors n_I^μ satisfy

$$\tau_I = n_I^\mu \tau_\mu = 0, \quad (2.24)$$

and so τ_a is nowhere vanishing on Σ_{p+1} (see Fig. 1 for an illustration of this condition). Then, taking

$$n^{\mu I} = h^{\mu\nu} n_\nu^I, \quad (2.25)$$

we make (2.24) manifest. We note that these considerations imply that

$$h^{IJ} = h^{\mu\nu} n_\mu^I n_\nu^J = \delta^{IJ}, \quad (2.26)$$

$$h^{aI} = h^{\mu\nu} u_\mu^a n_\nu^I = u_\mu^a n^{\mu I} = 0, \quad (2.27)$$

$$h_{IJ} = h_{\mu\nu} n_I^\mu n_J^\nu = h_{\mu\nu} h^{\nu\rho} n_I^\mu n_{\rho J} = (\delta_\mu^\rho + v^\rho \tau_\mu) n_I^\mu n_{\rho J} = \delta_{IJ}, \quad (2.28)$$

$$h_{aI} = h_{\mu\nu} u_\mu^a n_\nu^I = h_{\mu\nu} u_\mu^a h^{\nu\rho} n_{\rho I} = v_I \tau_a, \quad (2.29)$$

where $v_I = n_{I\mu} v^\mu$, which we will denote as the normal velocity.

⁹The submanifolds we consider are timelike in the sense that the normal vectors are required to be spacelike [see (2.24)]. The submanifolds will inherit a NC structure of their own.

¹⁰More formally, since the orientation of the normal 1-forms can be chosen freely as inward or outward pointing, \mathcal{M}^I_J is a matrix in $O(d-p)$.

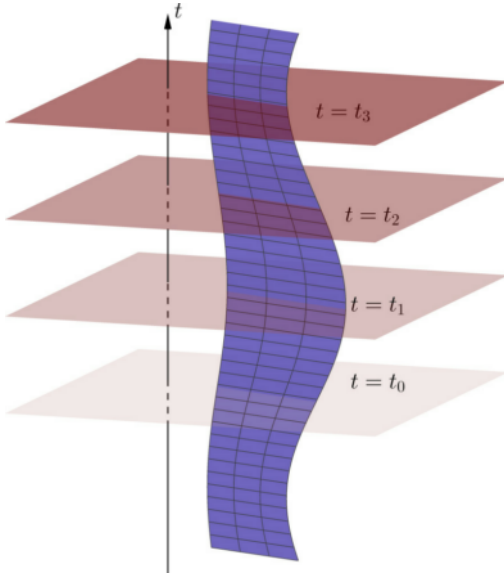


FIG. 1. Graphical depiction of the embedding of timelike Newton-Cartan submanifolds. The vertical direction represents the time t , and the spatial directions are in the plane orthogonal to the t axis. The spatial hypersurfaces of constant time are denoted by their corresponding value of t . Note in particular that the condition (2.24) implies that the submanifold does not “bend” away from the time direction in the ambient spacetime.

The description of submanifolds in NC geometry must be invariant under Galilean boosts, as these just express a choice of frame. This implies that the defining structure of NC submanifolds, namely, (2.17) and (2.20), must be invariant under local Galilean boost transformations. We start by noting that the embedding map does not transform under boosts, that is

$$\delta_G X^\mu = 0 \Rightarrow \delta_G u_a^\mu = 0, \quad (2.30)$$

and hence the tangent vectors to the surface are boost invariant.¹¹ Specializing to timelike submanifolds, using (2.25), the variations of (2.17) and (2.20), together with (2.30), require

$$\begin{aligned} u_a^\mu \delta_G n_\mu^I &= -n_\mu^I \delta_G u_a^\mu, & n^{\mu J} \delta_G n_\mu^I &= 0 \Rightarrow \delta_G n_\mu^I = 0, \\ u_\mu^a h^{\mu\nu} \delta_G n_{\mu I} &= -n_\mu^I \delta_G u_\mu^a, & & \\ u_a^\mu \delta_G u_\mu^b &= -u_\mu^b \delta_G u_a^\mu \Rightarrow \delta_G u_\mu^a = 0, \end{aligned} \quad (2.31)$$

while $\delta_G n_\mu^I = 0$ follows trivially from (2.25). Thus, Eq. (2.30) ensures that the defining structure of timelike NC submanifolds is boost invariant.¹²

¹¹Note that the embedding map specifies the location of the surface such that $x^\mu = X^\mu(\sigma^a)$. The spacetime coordinates x^μ do not transform under local Galilean boosts and hence neither does the embedding map $X^\mu(\sigma)$.

¹²In particular, (2.31) implies that $\delta_G v^I = n_\mu^I \delta_G v^\mu = n_\mu^I h^{\mu\nu} \lambda_\nu = \lambda^I$. This is consistent with (2.31) since $n_\mu^I = n_{***a}^\mu e^{***a} - v^I \tau_\mu$, so that $n_{***a}^\mu = e^{***a} n_\mu^I$. Given that $\delta_G e^{***a} = \delta_G \tau_\mu = 0$ and $\delta_G e_\mu^{***a} = \lambda^{***a} \tau_\mu$, we find that $\delta_G n_\mu^I = n_{***a}^\mu \lambda^{***a} \tau_\mu - \lambda^I \tau_\mu$ and since $\lambda_\mu = n_\mu^I \lambda_I + u_\mu^a \lambda_a$, we get $n_{***a}^\mu \lambda^{***a} = n_{***a}^\mu e^{***a} \lambda_\mu = \lambda^I$, thus confirming (2.31).

3. Induced Newton-Cartan structures

Besides the defining conditions (2.17) and (2.20), NC submanifolds have other inherent geometric structures, such as induced tensors, that can be introduced via appropriate contractions of ambient tensors with any of the objects u_a^μ and u_μ^a . We wish to identify the induced NC structures on the submanifold that have the same properties as the NC structures of the ambient spacetime. For instance, these induced structures should transform as in (2.4) and (2.5) but now involving only tangential directions to the submanifold.

The basic building blocks are the clock 1-form τ_a in Eq. (2.23) and the normal velocity v^I in Eq. (2.29) along with the pullbacks of the remaining ambient space fields

$$\begin{aligned} h_{ab} &= u_a^\mu u_b^\nu h_{\mu\nu}, & v^a &= u_\mu^a v^\mu, & h^{ab} &= u_\mu^a u_\nu^b h^{\mu\nu}, \\ m_a &= u_a^\mu m_\mu. \end{aligned} \quad (2.32)$$

It is possible to see that these structures mimic many of the properties of the ambient NC structure. For instance, we have $\tau_a h^{ab} = 0$ and $v^a \tau_a = -1$ by virtue of (2.24) and $\tau_\mu h^{\mu\nu} = 0$ as well as $v^\mu \tau_\mu = -1$. Additionally, they give rise to the completeness relation $h^{ac} h_{cb} = \delta_b^a + v^a \tau_b$, which in turn implies the relation $h^{\mu\nu} u_\mu^a = h^{ab} u_b^\nu$. However, using (2.29), we find that

$$v^a h_{ab} = u_\mu^a v^\mu u_b^\sigma u_\rho^\sigma h_{\rho\sigma} = -v^I h_{Ib} = -v^I v^J \tau_b, \quad (2.33)$$

which is nonzero, contrary to the corresponding ambient NC result $v^\mu h_{\mu\nu} = 0$. Hence, the individual structures in (2.32) do not form a NC geometry on the submanifold. Using (2.33) we instead define

$$\check{h}_{ab} = h_{ab} - v^I v_I \tau_a \tau_b, \quad (2.34)$$

which leads to a completeness relation and satisfies the required orthogonality condition, that is

$$h^{ac} \check{h}_{cb} = \delta_b^a + \tau_b v^a, \quad v^a \check{h}_{ab} = 0. \quad (2.35)$$

For \check{h}_{ab} to be considered a NC structure on the submanifold, one must also ensure that it transforms under Galilean boosts as its ambient space counterpart $h_{\mu\nu}$ [cf. (2.5)]. Using (2.4), (2.5), (2.31), and¹³ $v^a \lambda_a = -v^I \lambda_I$, it can be shown that

$$\begin{aligned} \delta_G v^a &= h^{ab} \check{\lambda}_b, & \delta_G(v^a h_{ab}) &= -2\tau_b \lambda^I v_I, & \delta_G h_{ab} &= 2\tau_{(a} \lambda_{b)}, \\ \delta_G \check{h}_{ab} &= 2\tau_{(a} \lambda_{b)} - 2\tau_a \tau_b v^I \lambda_I = 2\tau_{(a} \lambda_{b)} + 2\tau_a \tau_b v^c \lambda_c \\ &= 2\tau_{(a} \check{\lambda}_{b)}, \end{aligned} \quad (2.36)$$

where we have defined

$$\check{\lambda}_a = \lambda_a + v^c \lambda_c \tau_a = \check{h}_{ab} h^{bc} \lambda_c, \quad (2.37)$$

which satisfies $v^a \check{\lambda}_a = 0$, analogously to the ambient orthogonality condition $v^\mu \lambda_\mu = 0$. Thus \check{h}_{ab} transforms under submanifold Galilean boosts in the same manner as $h_{\mu\nu}$ transforms under ambient Galilean boosts.

NC submanifolds admit boost-invariant structures similar to the ambient structures (2.6) and (2.8). Given that the set of tangent and normal vectors is boost invariant [see Eq. (2.31)],

¹³This follows from the statement that $v^\mu \lambda_\mu = 0$.

two of these structures are obtained by contractions of the corresponding ambient quantities, namely,

$$\bar{h}_{ab} = u_a^\mu u_b^\nu \bar{h}_{\mu\nu} = \check{h}_{ab} - 2\tau_{(a}\check{m}_{b)}, \quad \hat{v}^a = u_a^\mu \hat{v}^\mu = v^a - h^{ab}\check{m}_b, \quad (2.38)$$

where we have defined the submanifold $U(1)$ connection

$$\check{m}_a = m_a - \frac{1}{2}v^I v_I \tau_a, \quad (2.39)$$

which transforms under boosts as $\delta_G \check{m}_a = \check{\lambda}_a$, analogous to the boost transformation of the ambient connection m_μ . Given that in the ambient space we have the identity $\hat{v}^\nu \bar{h}_{\nu\mu} = 2\check{\Phi}\tau_\mu$ where $\check{\Phi}$ is defined in (2.8) we require an analog condition of the form $\hat{v}^a \bar{h}_{ab} = 2\check{\Phi}\tau_b$ for some scalar $\check{\Phi}$. Explicit manipulation shows that

$$\begin{aligned} \hat{v}^a \bar{h}_{ab} &= u_a^\mu \hat{v}^\mu u_b^\nu \bar{h}_{\nu\rho} = \hat{v}^\nu \bar{h}_{\nu\rho} u_b^\rho - n_\mu^I h^{\nu\sigma} n_\sigma^I \hat{v}^\mu u_b^\rho \bar{h}_{\nu\rho} \\ &= 2(\check{\Phi} - 1/2\hat{v}^I \hat{v}^I)\tau_b, \end{aligned} \quad (2.40)$$

which leads us to identify

$$\check{\Phi} = \check{\Phi} - \frac{1}{2}\hat{v}^I \hat{v}^I = -v^a \check{m}_a + \frac{1}{2}h^{ab}\check{m}_a \check{m}_b, \quad (2.41)$$

thus taking the same form as its ambient counterpart (2.8) but now in terms of \check{m}_a .

In summary, we define the induced Newton-Cartan structure on the submanifold Σ_{p+1} to consist of the fields $(\tau_a, \check{h}_{ab}, \check{m}_a)$ and (v^a, h^{ab}) along with the boost-invariant combinations \hat{v}^a, \bar{h}_{ab} , and $\check{\Phi}$, satisfying the relations

$$\delta_b^a = h^{ac}\check{h}_{cb} - \tau_b v^a, \quad \tau_a h^{ab} = 0, \quad v^a \check{h}_{ab} = 0, \quad (2.42)$$

as well as

$$\hat{v}^a \bar{h}_{ab} = 2\check{\Phi}\tau_b. \quad (2.43)$$

These are related to the ambient Newton-Cartan structures in the following way:

$$\tau_a = u_a^\mu \tau_\mu, \quad \check{h}_{ab} = u_a^\mu u_b^\nu \check{h}_{\mu\nu} - v^I v_I \tau_a \tau_b = h_{ab} - v^I v_I \tau_a \tau_b, \quad (2.44)$$

$$\begin{aligned} \check{m}_a &= u_a^\mu m_\mu - \frac{1}{2}v^I v_I \tau_a = m_a - \frac{1}{2}v^I v_I \tau_a, \quad v^a = u_a^\mu v^\mu, \\ h^{ab} &= u_a^\mu u_b^\nu h^{\mu\nu}, \end{aligned} \quad (2.45)$$

$$\begin{aligned} \hat{v}^a &= v^a - h^{ab}\check{m}_b = u_a^\mu \hat{v}^\mu, \quad \bar{h}_{ab} = \check{h}_{ab} - 2\tau_{(a}\check{m}_{b)} \\ &= u_a^\mu u_b^\nu \bar{h}_{\mu\nu}, \end{aligned} \quad (2.46)$$

$$\check{\Phi} = -v^a \check{m}_a + \frac{1}{2}h^{ab}\check{m}_a \check{m}_b = \check{\Phi} - \frac{1}{2}\hat{v}^I \hat{v}^I. \quad (2.47)$$

These structures transform according to

$$\begin{aligned} \delta\tau_a &= \xi_\zeta \tau_a, \quad \delta\check{h}_{ab} = \xi_\zeta \check{h}_{ab} + 2\check{\lambda}_{(a}\tau_{b)}, \\ \delta\check{m}_a &= \xi_\zeta \check{m}_a + \check{\lambda}_a + \partial_a \sigma, \end{aligned} \quad (2.48)$$

$$\delta v^a = \xi_\zeta v^a + h^{ab}\check{\lambda}_b, \quad \delta h^{ab} = \xi_\zeta h^{ab}, \quad (2.49)$$

$$\begin{aligned} \delta\hat{v}^a &= \xi_\zeta \hat{v}^a - h^{ab}\partial_b \sigma, \quad \delta\bar{h}_{ab} = \xi_\zeta \bar{h}_{ab} - 2\tau_{(a}\partial_{b)}\sigma, \\ \delta\check{\Phi} &= \xi_\zeta \check{\Phi} - \hat{v}^a \partial_a \sigma, \end{aligned} \quad (2.50)$$

under submanifold diffeomorphisms ζ^a , Galilean boosts $\check{\lambda}_a$ (satisfying $v^a \check{\lambda}_a = 0$), and $U(1)$ gauge transformations σ .

4. The role of the transverse velocity v^I

In order to elucidate the role of v^I , we consider for concreteness a codimension one submanifold Σ moving with (constant) linear velocity $v_\Sigma^\mu = (0, 0, 0, \mathbf{v})$ in the z direction of a four-dimensional flat ambient Newton-Cartan spacetime, which was introduced in (2.15) and where i runs only over spatial directions. Defining Σ via the embedding equation

$$F(x, y, z - \mathbf{v}t) = 0, \quad (2.51)$$

we can write the normal 1-form as

$$n = NdF = N\partial_x F + N\partial_y F + N\partial_u F dz - \mathbf{v}\partial_u F dt, \quad (2.52)$$

where we have defined $u = z - \mathbf{v}t$ and where N is fixed by the normalization condition (2.17). This means that

$$v^\mu n_\mu = -n_0 = \mathbf{v}N\partial_u F, \quad v_\Sigma^\mu n_\mu = \mathbf{v}n_z = \mathbf{v}N\partial_u F, \quad (2.53)$$

leading us to conclude that $v^\mu n_\mu = v_\Sigma^\mu n_\mu$. Thus, the normal projection of the NC velocity is the same as the normal projection of the linear velocity vector v_Σ^μ of the submanifold Σ .

To illustrate this in the simplest possible setting, we consider an infinitely extended moving flat membrane embedded in $(3+1)$ -dimensional flat NC space, described by

$$u = z - \mathbf{v}t = 0, \quad (2.54)$$

leading to the normal 1-form

$$n_\mu = -\mathbf{v}\delta_\mu^0 + \delta_\mu^3 \Rightarrow v^\mu n_\mu = \mathbf{v}. \quad (2.55)$$

Therefore, for a flat brane, where the normal vector is the same everywhere, we see that the normal projection of the NC velocity vector is just the magnitude of the linear velocity of the plane.

5. Covariant derivatives, extrinsic curvature, and external rotation

Since we are dealing with the description of a single surface, and not of a foliation, covariant differentiation of submanifold structures only has meaning along tangential directions to the surface. Analogously to Lorentzian surfaces (see, e.g., Ref. [53]), we define a covariant derivative along surface directions that is compatible both with the surface Newton-Cartan structure, $D_a \tau_b = 0 = D_a h^{bc}$, and the ambient Newton-Cartan structure, $D_a \tau_\mu = 0 = D_a h^{\mu\nu}$, that acts on an arbitrary mixed tensor $T^{b\mu}$ as

$$D_a T^{b\mu} = \partial_a T^{b\mu} + \gamma_{ac}^b T^{c\mu} + u_a^\rho \Gamma_{\rho\lambda}^\mu T^{b\lambda}, \quad (2.56)$$

where we have introduced the surface affine connection according to

$$\gamma_{ab}^c = -\hat{v}^c \partial_a \tau_b + \frac{1}{2}h^{cd}(\partial_a \bar{h}_{bd} + \partial_b \bar{h}_{ad} - \partial_d \bar{h}_{ab}), \quad (2.57)$$

in analogy with the the spacetime affine connection (2.10). Note in particular that D_a does not act on transverse indices. The relation between γ_{ab}^c and $\Gamma_{\rho\lambda}^\mu$ is obtained in Appendix C and is shown to be

$$\gamma_{ab}^c = \Gamma_{ab}^c + u_c^\mu \partial_a u_b^\mu = u_c^\mu u_a^\nu \nabla_\nu u_b^\mu, \quad (2.58)$$

where the corresponding surface torsion tensor is

$$2\gamma_{[ab]}^c = -\hat{v}^c \tau_{ab} = -\hat{v}^c u_a^\mu u_b^\nu \tau_{\mu\nu}, \quad (2.59)$$

and where the last equality follows from the fact that exterior derivatives commute with pullbacks.¹⁴

It is also convenient to introduce a covariant derivative \mathfrak{D}_a that acts on all indices, i.e., μ, a, I [53], and whose action on the normal 1-forms and tangent vectors allows for the Weingarten decomposition¹⁵

$$\begin{aligned}\mathfrak{D}_a n_\sigma^I &= \partial_a n_\sigma^I - \Gamma_{\mu\sigma}^\lambda u_a^\mu n_\lambda^I - \omega_a^I J n_\sigma^I = -u_\sigma^b K_{ab}^I + \frac{1}{2} u_\sigma^b \hat{v}^I \tau_{ab}, \\ \mathfrak{D}_a u_b^\mu &= D_a u_b^\mu = n_I^\mu K_{ab}^I - \frac{1}{2} n_I^\mu \hat{v}^I \tau_{ab},\end{aligned}\quad (2.60)$$

where we have defined the extrinsic curvature to the submanifold according to

$$\begin{aligned}K_{ab}^I &= n_\mu^I D_a u_b^\mu + \frac{1}{2} \hat{v}^I \tau_{ab} = n_\mu^I (\partial_a u_b^\mu + u_a^\nu u_b^\rho \Gamma_{(\nu\rho)}^\mu) \\ &= -u_a^\mu u_b^\nu \nabla_{(\mu} n_{\nu)}^I.\end{aligned}\quad (2.61)$$

The extrinsic curvature tensor, when defined in this manner, is symmetric and invariant under Galilean boosts but transforms under $U(1)$ gauge transformations according to

$$\delta_\sigma K_{ab}^I = \frac{1}{2} \tau_{Ia} \partial_b \sigma + \frac{1}{2} \tau_{Ib} \partial_a \sigma, \quad (2.62)$$

where we used (2.14). In (2.60) we also introduced the external rotation tensor, which can be interpreted as a $SO(d-p)$ connection, defined as

$$\omega_a^I J = n_I^\mu D_a n_{\mu}^I, \quad (2.63)$$

which is antisymmetric in I, J indices and transforms under $U(1)$ gauge transformations as

$$\delta_\sigma \omega_a^I J = -\frac{1}{2} (\tau_{aJ} \partial^I \sigma + \tau^I J \partial_a \sigma + \tau^I_a \partial_J \sigma). \quad (2.64)$$

If the submanifold is codimension one, the external rotation vanishes by definition.

Both the extrinsic curvature tensor and the external rotation tensor introduced here are direct analogues of their Lorentzian counterparts [53]. To see that $\omega_a^I J$ transforms as a connection we examine what happens if we perform a local $SO(d-p)$ rotation of the normal vectors as in (2.18). If we focus on an infinitesimal rotation $\mathcal{M}^I J = \delta^I J + \lambda^I J$ where $\lambda^I J = -\lambda^J I$, the extrinsic curvature tensor and external rotation tensor transform as

$$\delta_\lambda K_{ab}^I = \lambda^I J K_{ab}^J, \quad \delta_\lambda \omega_a^I J = \partial_a \lambda^I J + \lambda^I_K \omega_a^K J + \lambda_J^K \omega_a^I K. \quad (2.65)$$

In addition, under a change of sign of the normal vectors $n_I^\mu \rightarrow -n_I^\mu$, the extrinsic curvature changes sign.

6. Integrability conditions

Certain combinations of geometric structures of Lorentzian submanifolds are related to specific contractions of the Riemann tensor of the ambient space. These are known as integrability conditions. In this section we derive the analogous conditions in the context of NC submanifolds, which are

¹⁴Alternatively, this conclusion can be reached via the relation $\partial_a u_b^\mu = \partial_a \partial_b X^\mu = \partial_b \partial_a X^\mu = \partial_b u_a^\mu$.

¹⁵The action of \mathfrak{D}_a on some vector T^I takes the form $\mathfrak{D}_a T^I = D_a T^I - \omega_a^I J T^J$.

known as the Codazzi-Mainardi, Gauss-Codazzi, and Ricci-Voss equations. In order to do so, we note that in the presence of torsion, the Ricci identity takes the form

$$[\nabla_\mu, \nabla_\nu] X_\sigma = R_{\mu\nu\sigma}{}^\rho X_\rho - 2\Gamma_{[\mu\nu]}^\rho \nabla_\rho X_\sigma, \quad (2.66)$$

where the Riemann tensor $R_{\mu\nu\sigma}{}^\rho$ of the ambient space is given by

$$R_{\mu\nu\sigma}{}^\rho = -\partial_\mu \Gamma_{\nu\sigma}^\rho + \partial_\nu \Gamma_{\mu\sigma}^\rho - \Gamma_{\mu\lambda}^\rho \Gamma_{\nu\sigma}^\lambda + \Gamma_{\nu\lambda}^\rho \Gamma_{\mu\sigma}^\lambda. \quad (2.67)$$

The integrability conditions to be derived below take a nice form if we work with an object that is closely related to the extrinsic curvature, namely,

$$\tilde{K}_{ab}^I = n_\mu^I D_a u_b^\mu = K_{ab}^I - \frac{1}{2} \hat{v}^I \tau_{ab}, \quad (2.68)$$

which has a nonvanishing antisymmetric part $2\tilde{K}_{[ab]}^I = -\hat{v}^I \tau_{ab}$.

We begin by deriving the Codazzi-Mainardi equation (see, e.g., Refs. [48,53]) by considering the quantity $D_a \tilde{K}_{bc}^I - D_b \tilde{K}_{ac}^I$. We find

$$D_a \tilde{K}_{bc}^I = \tilde{K}_{bc}^I n_I^\rho (\nabla_\rho u_c^\mu) n_\mu^I - \omega_b^I J \tilde{K}_{ac}^J - u_c^\mu u_a^\nu u_b^\sigma \nabla_\rho \nabla_\sigma n_\mu^I, \quad (2.69)$$

where we used (2.63). From here, using (2.66) and the covariant derivative \mathfrak{D}_a introduced in (2.60) we derive the Codazzi-Mainardi equation

$$\mathfrak{D}_a \tilde{K}_{bc}^I - \mathfrak{D}_b \tilde{K}_{ac}^I = -R_{abc}^I + \hat{v}^d \tau_{ab} \tilde{K}_{dc}^I. \quad (2.70)$$

In order to derive the Gauss-Codazzi equation, we let ω_c be any submanifold 1-form that is the pullback of ω_μ whose normal components vanish, i.e., $\omega_\mu = u_\mu^c \omega_c$. The Ricci identity for the submanifold reads

$$[D_a, D_b] \omega_c = \mathcal{R}_{abc}{}^d \omega_d + \hat{v}^d \tau_{ab} D_d \omega_c, \quad (2.71)$$

where $\mathcal{R}_{abc}{}^d$ is the Riemann tensor of the submanifold and takes the same form as (2.67) but with the connection $\Gamma_{\nu\sigma}^\rho$ replaced by γ_{ab}^c of (2.57). Using $u_\mu^a D_b u_c^\mu = 0$ [which follows from (2.58)] and $n_I^\mu D_b u_\mu^I = h^{de} \tilde{K}_{beI}$, explicit manipulation leads to

$$\begin{aligned}\mathcal{R}_{abc}{}^d \omega_d + \hat{v}^d \tau_{ab} D_d \omega_c &= h^{ed} \tilde{K}_{ac}^I \tilde{K}_{beI} \omega_d - h^{ed} \tilde{K}_{bc}^I \tilde{K}_{aeI} \omega_d \\ &\quad + R_{abc}{}^d \omega_d + \tau_{ab} (-\hat{v}^I n_I^\rho u_c^\mu \nabla_\rho \omega_\mu \\ &\quad + \hat{v}^v u_c^\mu \nabla_\nu \omega_\mu),\end{aligned}\quad (2.72)$$

where we used (2.66). In this expression, the terms proportional to τ_{ab} on both sides cancel and since it must be true for any one form ω_c , the Gauss-Codazzi equation becomes

$$\mathcal{R}_{abc}{}^d = \tilde{K}_{ac}^I \tilde{K}_b{}^d{}_I - \tilde{K}_{bc}^I \tilde{K}_a{}^d{}_I + R_{abc}{}^d, \quad (2.73)$$

where $\tilde{K}_b{}^d{}_I = h^{dc} \tilde{K}_{bcI}$.

Although we will not use it in this paper, we will briefly discuss the Ricci-Voss equation for completeness. This equation becomes useful for surfaces of codimension higher than one, where we can define the outer curvature in terms of the external rotation tensor (2.63) as

$$\Omega^I{}_{Jab} = 2\partial_{[a} \omega_{b]}^I J - 2\omega_{[a}^I K \omega_{|b]}^K J. \quad (2.74)$$

In terms of this tensor, the Ricci-Voss equation for Newton-Cartan geometry can be shown to read

$$\Omega^I{}_{Jab} = R_{abJ}{}^I - 2h^{cd}\tilde{K}_{[a|c}{}^I\tilde{K}_{|b]dJ}. \quad (2.75)$$

This completes the description of the geometric structures of NC submanifolds.

III. VARIATIONS AND DYNAMICS OF NEWTON-CARTAN SUBMANIFOLDS

In the previous section we defined timelike NC submanifolds and their characteristic geometric properties. In this section, closely following the Lorentzian case [53], we develop the variational calculus for NC submanifolds for the geometric structures of interest. These results are necessary to later introduce geometric action functionals capable of describing different types of soft matter systems, including the case of bending energies for lipid vesicles.

A. Variations of Newton-Cartan objects on the submanifold

In the following, we consider two types of variations, namely, embedding map variations, which are displacements of the submanifold, and Lagrangian variations which consist of the class of diffeomorphisms that displace the ambient space but keep the embedding map fixed (see, e.g., Refs. [49,50] and also Refs. [46,53]). As in the Lorentzian case [53], the sum of these two types of variations yield the transformation properties of the submanifold structures under full ambient space diffeomorphisms. When considering action functionals that give dynamics to submanifolds, they are equivalent, up to normal rotations.¹⁶

1. Embedding map variations

Before specializing to any of the two types of variations, it is useful to consider general variations of the normal vectors. In particular, we decompose the variation of the normal vectors as

$$\begin{aligned} \delta n_\mu^I &= u_\mu^a u_a^\nu \delta n_\nu^I + n_\mu^J n_J^\nu \delta n_\nu^I \\ &= -u_\mu^a n_\nu^I \delta u_a^\nu + \frac{1}{2} n_{\mu J} (n^{\nu J} \delta n_\nu^I + n^{\nu I} \delta n_\nu^J) \\ &\quad + \lambda^I{}_J n_\mu^J, \end{aligned} \quad (3.1)$$

where

$$\lambda^I{}_J = \frac{1}{2} (n_J^\nu \delta n_\nu^I - n^{\nu I} \delta n_{J\nu}), \quad (3.2)$$

is a local $\mathfrak{so}(d-p)$ transformation of the normal vectors. By varying the second relation in (2.17), we find the relation $n^{\nu J} \delta n_\nu^I + n^{\nu I} \delta n_\nu^J = -n_\mu^I n_\nu^J \delta h^{\mu\nu}$ and hence

$$\delta n_\mu^I = -u_\mu^a n_\nu^I \delta u_a^\nu - \frac{1}{2} n_{\mu J} n_\nu^J \delta h^{\nu\rho} + \lambda^I{}_J n_\mu^J. \quad (3.3)$$

By varying the completeness relation (2.19) one may express variations of $\delta h^{\nu\rho}$ in terms of variations of τ_ν and $h_{\nu\rho}$ such that $\delta h^{\mu\nu} = 2v^{(\mu} h^{\nu)\lambda} \delta \tau_\lambda - h^{\mu\rho} h^{\nu\sigma} \delta h_{\rho\sigma}$. This leads to

$$\begin{aligned} \delta n_\mu^I &= -v^{(I} n^{J)\nu} n_{\mu J} \delta \tau_\nu + \frac{1}{2} n^{\rho J} n_{\mu J} n^{\nu I} \delta h_{\rho\nu} - n_\nu^I u_\mu^a \delta u_a^\nu \\ &\quad + \lambda^I{}_J n_\mu^J, \end{aligned} \quad (3.4)$$

¹⁶In the context of continuum mechanics, these two viewpoints are known as the Lagrangian and Eulerian descriptions.

which describes arbitrary infinitesimal variations of the normal vectors.

We now specialise to infinitesimal variations of the embedding map which we denote by

$$\delta X^\mu(\sigma) = -\xi^\mu(\sigma), \quad (3.5)$$

where $\xi^\mu(\sigma)$ is understood as being an infinitesimal first order variation. Under this variation, the ambient tensor structures evaluated at the surface [i.e., $\tau_\mu(X)$, $\bar{h}_{\mu\nu}(X)$] vary as

$$\delta_X \tau_\mu(X) = -\xi^\nu \partial_\nu \tau_\mu, \quad \delta_X \bar{h}_{\mu\nu}(X) = -\xi^\rho \partial_\rho \bar{h}_{\mu\nu}, \quad (3.6)$$

which follows from $\delta_X \tau_\mu(X) = \tau_\mu(X - \xi) - \tau_\mu(X) = -\xi^\nu \partial_\nu \tau_\mu + \mathcal{O}(\xi^2)$. In turn, the tangent vectors transform as

$$\delta_X u_a^\mu = \partial_a \delta X^\mu = -\partial_a \xi^\mu, \quad (3.7)$$

while variations of the induced metric structures take the form

$$\delta_X \tau_a = -u_a^\mu \xi_\xi \tau_\mu, \quad \delta_X \bar{h}_{ab} = -u_a^\mu u_b^\nu \xi_\xi \bar{h}_{\mu\nu}. \quad (3.8)$$

In other words, for these structures, performing embedding map variations is equivalent to performing a diffeomorphism in the space of embedding maps that keep u_a^μ fixed, i.e., they are diffeomorphisms that are independent of σ^a . Using (3.4), we can write the variation of the normal vector as

$$\begin{aligned} \delta_X n_\mu^I &= -n_{\mu J} n_\rho^I n^{J\nu} \nabla_\nu \xi^\rho - n_{\mu J} \hat{v}^{(I} n^{J)\nu} \tau_{\nu\rho} \xi^\rho \\ &\quad + n_\rho^I \partial_\mu \xi^\rho + \tilde{\lambda}^{IJ} n_{\mu J}, \end{aligned} \quad (3.9)$$

where the third term ensures that the orthogonality relation $u_a^\mu n_\mu^I = 0$ is obeyed after the variation while the last term is a local transverse rotation of the form $\tilde{\lambda}^{IJ} = \lambda^{IJ} + n_\rho^I n^{\rho J} \partial_\nu \xi^\rho$.

For the purposes of this work, as mentioned in Sec. II A 3, we will be focusing on ambient NC geometries with absolute time, i.e., zero torsion. This extra assumption greatly simplifies many expressions after variation. We stress, however, that it is in general not possible to assume zero torsion before variation, as variation and setting torsion to zero do not always commute.¹⁷

However, specifically in the case of embedding map or Lagrangian variations, the variation of $\tau_{\mu\nu}$ is guaranteed to vanish when the torsion itself vanishes. This means that we can set torsion to zero in the Lagrangian if all we are interested in are the equations of motion for X^μ . For example, $\delta_X \tau_{\mu\nu}(X) = -\xi^\rho \partial_\rho \tau_{\mu\nu}$, which vanishes when $d\tau = 0$. Under the assumption of vanishing torsion, variations of the extrinsic curvature (2.61) take the form

$$\begin{aligned} \delta_X K_{ab}{}^I &= (\delta_X n_\mu^I) \partial_a u_b^\mu - n_\mu^I \partial_a \partial_b \xi^\mu + (\delta_X n_\mu^I) u_a^\rho \Gamma_{\rho\lambda}^\mu u_b^\lambda \\ &\quad - n_\mu^I (\partial_a \xi^\rho) \Gamma_{\rho\lambda}^\mu u_b^\lambda - n_\mu^I u_a^\rho \xi^\kappa \partial_\kappa (\Gamma_{\rho\lambda}^\mu) u_b^\lambda \\ &\quad - n_\mu^I u_a^\rho \Gamma_{\rho\lambda}^\mu \partial_b \xi^\lambda \\ &= -n_\mu^I D_a D_b \xi^\mu + \xi^\rho R_{\rho ab}{}^I + n_\rho^I n^{J\nu} \Gamma_{\nu\sigma}^\rho \xi^\sigma K_{abJ}, \end{aligned} \quad (3.10)$$

¹⁷For instance, when considering equations of motion for surfaces via extremization of a Lagrangian as in the next section, a term of the form $X^{\mu\nu} \tau_{\mu\nu}$ in the Lagrangian can give a nonzero contribution to the equation of motion of τ as neither $X^{\mu\nu}$ nor $\delta \tau_{\mu\nu}$ need to vanish on ambient spaces with zero torsion.

where we have used (3.9) as well as $\delta_X \Gamma_{\rho\lambda}^\mu(X) = -\xi^\kappa \partial_\kappa \Gamma_{\rho\lambda}^\mu$. The last term in (3.10) denotes a local $\mathfrak{so}(d-p)$ transformation, and we have explicitly ignored further rotations by setting $\lambda^{IJ} = 0$ in (3.4). It is also straightforward to consider variations of the external rotation tensor (2.63), but since we do not explicitly consider this structure in the dynamics of submanifolds, we will not dwell on this.

2. Lagrangian variations

In the previous section we have described how to perform variations of the embedding map. In this section we focus on a particular class of diffeomorphisms $x^\mu \rightarrow x^\mu - \xi^\mu$ that act only on fields with support in the entire ambient spacetime, that is, they only act on the NC triplet $(\tau_\mu(x), h_{\mu\nu}(x), m_\mu(x))$. In general, diffeomorphisms also displace the embedding map according to $\delta_\xi X^\mu = -\xi^\mu$ where δ_ξ denotes an infinitesimal diffeomorphism variation. However, here we consider the case of Lagrangian variations for which $\delta_\xi X^\mu = 0$ (see, e.g., Refs. [46,49,50,53]). In turn, this implies that the tangent vectors do not vary:¹⁸

$$\delta_\xi u_a^\mu = 0. \quad (3.11)$$

In the remainder of this section, we will explicitly work out Lagrangian variations of submanifold structures and compare them with embedding map variations, thereby extracting the transformation properties under full ambient diffeomorphisms. In particular, using (3.11) and the fact that $\delta_\xi \tau_\mu = \mathcal{L}_\xi \tau_\mu$ and $\delta_\xi \bar{h}_{\mu\nu} = \mathcal{L}_\xi \bar{h}_{\mu\nu}$ we find

$$\delta_\xi \tau_a = u_a^\mu \mathcal{L}_\xi \tau_\mu, \quad \delta_\xi \bar{h}_{ab} = u_a^\mu u_b^\nu \mathcal{L}_\xi \bar{h}_{\mu\nu}. \quad (3.12)$$

Comparing this with (3.8), it follows that for pullbacks of Newton-Cartan objects we have the relations

$$(\delta_\xi + \delta_X) \tau_a = (\delta_\xi + \delta_X) \bar{h}_{ab} = 0, \quad (3.13)$$

and thus these objects transform as scalars under ambient diffeomorphisms. For later purposes, we rewrite these results as

$$\delta_\xi \tau_a = \tau_\rho D_a \xi^\rho, \quad (3.14)$$

$$\begin{aligned} \delta_\xi \bar{h}_{ab} &= \bar{h}_{\rho b} D_a \xi^\rho + \bar{h}_{\rho a} D_b \xi^\rho - 2\tau_a \tau_b \xi^\sigma \partial_\sigma \bar{\Phi} \\ &\quad - 2\xi^\sigma \tau_\sigma \tau_{(a} \partial_{b)} \bar{\Phi} - 2\tau_{(a} \bar{\mathcal{K}}_{b)\sigma} \xi^\sigma, \end{aligned} \quad (3.15)$$

where we have used the relation (valid in the absence of torsion)

$$\nabla_\sigma \bar{h}_{\mu\nu} = -2\tau_\mu \tau_\nu \partial_\sigma \bar{\Phi} - 2\tau_\sigma \tau_{(\mu} \partial_{\nu)} \bar{\Phi} - 2\tau_{(\mu} \bar{\mathcal{K}}_{\nu)\sigma}, \quad (3.16)$$

as well as $\hat{v}^\lambda \bar{h}_{\lambda\mu} = 2\tau_\mu \bar{\Phi}$ and where $\bar{\mathcal{K}}_{\mu\nu} = -\mathcal{L}_{\hat{v}} \bar{h}_{\mu\nu}/2$.

Considering the normal 1-forms, using (3.4) we find that

$$\begin{aligned} \delta_\xi n_\mu^I &= -v^{(I} n^{J)\nu} n_{\mu J} \tau_\rho \nabla_\nu \xi^\rho + n^{\lambda J} n_{\mu J} n^{\nu I} h_{\rho(\lambda} \nabla_{\nu)} \xi^\rho \\ &= n_{\mu J} n_\rho^{(I} n^{J)\nu} \nabla_\nu \xi^\rho, \end{aligned} \quad (3.17)$$

where we have used (3.16) as well as the identity $n_\mu^I h_{\rho\lambda} = h_{\rho I} = \tau_\rho v_I + n_{\rho I}$ and assumed vanishing torsion. Comparing this to the embedding map variation (3.9), we find that

$$(\delta_\xi + \delta_X) n_\mu^I = \tilde{\lambda}^I J n_\mu^I + n_\rho^I \partial_\mu \xi^\rho, \quad (3.18)$$

where $\tilde{\lambda}^{IJ} = -n_\rho^{(I} n^{J)\nu} \partial_\nu \xi^\rho$ is a local $\mathfrak{so}(d-p)$ transformation and we have set $\lambda^{IJ} = 0$ in (3.4). This implies that, up to a $\text{SO}(d-p)$ rotation, the normal 1-forms n_μ^I transform as 1-forms under ambient diffeomorphisms. This is the expected result (and analogous to the Lorentzian case [53]) as the 1-forms carry a spacetime index μ . Repeating this procedure for the extrinsic curvature, we find that

$$\delta_\xi K_{ab}^I = K_{ab}^\mu \delta_\xi n_\mu^I + n_\mu^I u_a^\rho u_b^\lambda \delta_\xi \Gamma_{\rho\lambda}^\mu. \quad (3.19)$$

Since $\Gamma_{\rho\lambda}^\mu$ is an affine connection, it transforms in the following way under diffeomorphisms

$$\begin{aligned} \delta_\xi \Gamma_{\lambda\nu}^\mu &= \xi^\rho \partial_\rho \Gamma_{\lambda\nu}^\mu - \Gamma_{\lambda\nu}^\rho \partial_\rho \xi^\mu + \Gamma_{\rho\nu}^\mu \partial_\lambda \xi^\rho + \Gamma_{\lambda\rho}^\mu \partial_\nu \xi^\rho + \partial_\lambda \partial_\nu \xi^\mu \\ &= \nabla_\lambda \nabla_\nu \xi^\mu - \xi^\rho R_{\rho\lambda\nu}^\mu, \end{aligned} \quad (3.20)$$

where in the second equality we assumed vanishing torsion. This implies that

$$\begin{aligned} \delta_\xi K_{ab}^I &= n_\mu^I D_a D_b \xi^\mu - \frac{1}{2} n_\mu^I K_{ab}^\sigma \nabla_\sigma \xi^\mu + \frac{1}{2} K_{abJ} n_\rho^J n^{I\nu} \nabla_\nu \xi^\rho \\ &\quad - n_\mu^I u_a^\lambda u_b^\nu \xi^\rho R_{\rho\lambda\nu}^\mu \\ &= n_\mu^I D_a D_b \xi^\mu - \xi^\rho R_{\rho ab}^I - K_{abJ} n_\rho^J n^{I\nu} \nabla_\nu \xi^\rho. \end{aligned} \quad (3.21)$$

Comparing this to (3.10), we obtain

$$(\delta_X + \delta_\xi) K_{ab}^I = \tilde{\lambda}^{IJ} K_{abJ}, \quad (3.22)$$

which, as in the Lorentzian case [53], states that the extrinsic curvature transforms as a scalar under ambient diffeomorphisms up to a transverse rotation.

B. Action principle and equations of motion

Equipped with the variational technology of the previous section, we consider the dynamics of submanifolds that arise via the extremization of an action. In the context of soft matter systems this action can be interpreted as a free energy functional that depends on geometrical degrees of freedom. Examples of such systems are fluid membranes and lipid vesicles, described by Canham-Helfrich-type free energies. The equations of motion that arise from extremization naturally split into tangential energy and mass-momentum conservation equations in addition to the shape equation (which describes the mechanical balance of forces in the normal directions), as well as constraints (Ward identities) arising from $\text{SO}(d-p)$ rotational invariance and boundary conditions.

¹⁸If we were working with foliations of surfaces instead of a single surface, we could define a set of vector fields $u_a^\mu(x)$ where x is any point in the ambient spacetime. We could then require that the Lie brackets between these vector fields vanish so that their integral curves can be thought of as locally describing a set of curvilinear coordinates for the submanifold. In other words, the restriction of these vector fields to the submanifold obeys the condition that the u_a^μ are tangent vectors, i.e., $u_a^\mu(x)|_{x=X} = \partial_a X^\mu$. When we perform ambient diffeomorphisms within the context of a foliation, we must ensure that this condition is respected. This means that $[\xi^\rho(x) \partial_\rho u_a^\mu(x) - u_a^\rho(x) \partial_\rho \xi^\mu(x)]|_{x=X} = \mathcal{L}_\xi u_a^\mu = \delta_\xi u_a^\mu = 0$. Lagrangian diffeomorphisms are thus generated by $\xi^\mu(x)$ such that (3.11) is obeyed. See, e.g., Ref. [52].

1. Equations of motion and rotational invariance

Following Ref. [53], we consider an action S on a $(p+1)$ -dimensional NC submanifold that is a functional of the metric data τ_a, \bar{h}_{ab} (this set contains all the fields τ_a, h_{ab}, m_a and is an equivalent choice of NC objects) as well as the extrinsic curvature, that is $S = S[\tau_a, \bar{h}_{ab}, K_{ab}^I]$. The variation of this action takes the general form

$$\delta S = \int_{\Sigma} d^{p+1} \sigma e (\mathcal{T}^a \delta \tau_a + \frac{1}{2} \mathcal{T}^{ab} \delta \bar{h}_{ab} + \mathcal{D}^{ab}{}_I \delta K_{ab}^I). \quad (3.23)$$

Here e is the integration measure given by $e = \sqrt{-\det(-\tau_a \tau_b + \bar{h}_{ab})}$ and invariant under local Galilean boosts and $U(1)$ gauge transformations. The response \mathcal{T}^a is the energy current,¹⁹ while the response \mathcal{T}^{ab} is the Cauchy stress-mass tensor [59]. Finally, $\mathcal{D}^{ab}{}_I$ is the bending moment, encoding elastic responses, and typically takes the form of an elasticity tensor contracted with the extrinsic curvature (strain) [46,53]. Both \mathcal{T}^{ab} and $\mathcal{D}^{ab}{}_I$ are symmetric as they

inherit the symmetry properties of \bar{h}_{ab} and K_{ab}^I . The temporal projection of the Cauchy stress-mass tensor, $\tau_b \mathcal{T}^{ab}$, is the mass current.

We require the action (3.23) to be invariant under $U(1)$ gauge transformations for which $\delta_{\sigma} \bar{h}_{ab} = -2\tau_{(a} \partial_{b)} \sigma$ and invariant under $SO(d-p)$ rotations for which the extrinsic curvature transforms according to (2.65). Ignoring boundary terms, to be dealt with in Sec. III B 2, this leads to mass conservation and a constraint on the bending moment, respectively:

$$D_b (\mathcal{T}^{ab} \tau_a) = 0, \quad \mathcal{D}^{abI} K_{ab}^J = 0. \quad (3.24)$$

In particular, the latter condition takes exactly the same form as in the Lorentzian context [46,53] and can also be obtained by performing a Lagrangian variation of (3.23) as we shall see. In order to obtain the equations of motion arising from (3.23), we can perform a Lagrangian variation as originally considered in Refs. [49,50] and developed further in Ref. [53].²⁰ Under a Lagrangian variation, using Sec. III A 2, the action (3.23) varies according to

$$\begin{aligned} \delta_{\xi} S = & \int_{\Sigma} d^{p+1} \sigma e \xi^{\rho} [-\tau_{\rho} D_a \mathcal{T}^a - D_a (\bar{h}_{\rho b} \mathcal{T}^{ab}) - \mathcal{T}^{ab} \{\tau_a \tau_b \partial_{\rho} \tilde{\Phi} + \tau_{\rho} \tau_a \partial_b \tilde{\Phi} + \tau_a \bar{K}_{b\rho}\} + D_a D_b (\mathcal{D}^{ab}{}_I n_{\rho}^I) - \mathcal{D}^{ab}{}_I R_{\rho ab}^I] \\ & + \int_{\Sigma} d^{p+1} \sigma e D_a [\mathcal{T}^a \tau_{\rho} \xi^{\rho} + \mathcal{T}^{ab} \bar{h}_{\rho b} \xi^{\rho} + \mathcal{D}^{ab}{}_I n_{\rho}^I D_b \xi^{\rho} - D_b (\mathcal{D}^{ab}{}_I n_{\rho}^I) \xi^{\rho}] + \int_{\Sigma} d^{p+1} \sigma e \mathcal{D}^{ab}{}_I K_{abJ} n_{\rho}^I n^{J\rho} \nabla_{\sigma} \xi^{\rho}. \end{aligned} \quad (3.25)$$

In this equation, the second integral gives rise to a boundary term which we consider in Sec. (III B 2). The last integral vanishes due to the requirement of rotational invariance (3.24). However, even if (3.24) was not imposed, given that the last term involves a normal derivative of ξ_{μ} , it cannot be integrated out and hence must vanish independently giving again rise to the second condition in (3.24), as in the Lorentzian case [53].

The first integral in (3.25) must vanish for an arbitrary vector field ξ^{μ} and hence it gives rise to the equation of motion

$$\begin{aligned} -\tau_{\rho} D_a \mathcal{T}^a - \bar{h}_{\rho b} D_a \mathcal{T}^{ab} - \mathcal{T}^{ab} \bar{h}_{\rho\sigma} K_{ab}^{\sigma} + 2\tau_{\rho} \mathcal{T}^{ab} \tau_a \partial_b \tilde{\Phi} \\ + \tau_{\rho} \bar{K}_{ab} \mathcal{T}^{ab} + D_a D_b (\mathcal{D}^{ab}{}_I n_{\rho}^I) - \mathcal{D}^{ab}{}_I R_{\rho ab}^I = 0, \end{aligned} \quad (3.26)$$

where we have used (3.16). In Appendix A we provide the relation (A28) between \bar{K}_{ab} , which is the pullback of $\bar{K}_{\mu\nu}$, and $\bar{K}_{ab}^{\Sigma} = -\mathcal{L}_{\hat{v}}^{\Sigma} \bar{h}_{ab}/2$ which is the actual surface-equivalent of $\bar{K}_{\mu\nu}$. Here $\mathcal{L}_{\hat{v}}^{\Sigma}$ denotes the surface Lie derivative along \hat{v}^a . Using this relation, as well as (2.41), which relates the Newtonian potential on the submanifold $\tilde{\Phi}$ to its ambient spacetime counterpart $\tilde{\Phi}$, the equation of motion (3.26) can

be written as

$$\begin{aligned} \tau_{\rho} D_a \mathcal{T}^a + \bar{h}_{\rho b} D_a \mathcal{T}^{ab} + \mathcal{T}^{ab} \bar{h}_{\rho\sigma} K_{ab}^{\sigma} - 2\tau_{\rho} \mathcal{T}^{ab} \tau_a \partial_b \tilde{\Phi} \\ - \tau_{\rho} \bar{K}_{ab}^{\Sigma} \mathcal{T}^{ab} - \tau_{\rho} \hat{v}^I K_{ab}^I \mathcal{T}^{ab} - D_a D_b (\mathcal{D}^{ab}{}_I n_{\rho}^I) \\ + \mathcal{D}^{ab}{}_I R_{\rho ab}^I = 0. \end{aligned} \quad (3.27)$$

The equation of motion (3.27) can be projected tangentially or orthogonally to Σ , yielding two independent equations. The tangential projection, known as the intrinsic equation of motion, is given by

$$\begin{aligned} \tau_c [D_a (\mathcal{T}^a - 2\tilde{\Phi} \mathcal{T}^{ab} \tau_b) - \mathcal{T}^{ab} \bar{K}_{ab}^{\Sigma}] + \bar{h}_{bc} D_a \mathcal{T}^{ab} \\ + 2D_a (K_{bc}^I \mathcal{D}^{ab}{}_I) - \mathcal{D}^{ab}{}_I D_c K_{ab}^I = 0, \end{aligned} \quad (3.28)$$

where we have used the Codazzi-Mainardi equation (2.70), assuming vanishing torsion, in order to eliminate contractions with the Riemann tensor. Equation (3.28) can be further projected along h^{cd} and \hat{v}^c , which again yields two independent equations. These projections can be simplified by defining $\mathcal{T}_m^{ad} = \mathcal{T}^{ad} + 2\mathcal{D}^{b(a}{}_I h^{d)c} K_{bc}^I$ and $\mathcal{T}_m^a = \mathcal{T}^a - 2\hat{v}^c K_{bc}^I \mathcal{D}^{ab}{}_I$. In particular, the spatial projection using h^{cd} gives rise to mass and momentum conservation

$$D_a \mathcal{T}_m^{ad} + 2D_a (\mathcal{D}^{b[a}{}_I h^{d]c} K_{bc}^I) - h^{cd} \mathcal{D}^{ab}{}_I D_c K_{ab}^I = 0, \quad (3.29)$$

where we have used invariance under $U(1)$ gauge transformations [the first condition in (3.24)]. In turn, the projection along \hat{v}^c leads to energy conservation

$$D_a \mathcal{T}_m^a - \mathcal{T}_m^{ab} \bar{K}_{ab}^{\Sigma} - 2\mathcal{T}_m^{ab} \tau_b \partial_a \tilde{\Phi} + \mathcal{D}^{ab}{}_I \hat{v}^c D_c K_{ab}^I = 0, \quad (3.30)$$

where we have used the identity $D_a \hat{v}^c = -h^{cd} (\bar{K}_{ad}^{\Sigma} + \tau_a \partial_d \tilde{\Phi})$ as well as the first condition in (3.24).

¹⁹As mentioned throughout this paper, we have focused on the case of vanishing torsion $\tau_{\mu\nu} = 0$, meaning that $\tau_a = \partial_a T$, where T is some scalar. Therefore, varying τ_a is actually varying T in (3.23), which in turn implies that we are not able to extract \mathcal{T}^a from the action but only its divergence. This is sufficient for the purposes of this work.

²⁰Alternatively, we may perform embedding map variations.

The intrinsic equations (3.29) and (3.30) result from diffeomorphism invariance along the tangential directions $\xi^a = u_\mu^a \xi^\mu$ or, equivalently, from tangential reparametrization invariance $\delta X^\mu = u_\mu^a \delta X^a$. Since the action only depends on the NC objects τ_a , \bar{h}_{ab} , and K_{ab}^I , the intrinsic equations are nothing but Bianchi identities that result from the diffeomorphism invariance of the action and hence are identically satisfied.

Finally, the transverse projection of (3.27) is usually referred to as the *shape equation*, and it is given by

$$\mathcal{T}^{ab} K_{ab}^I = \mathfrak{D}_a \mathfrak{D}_b \mathcal{D}^{abI} - \mathcal{D}^{ab}{}_J K_{ac}^I K_{bd}^J h^{cd} - \mathcal{D}^{ab}{}_J R_{Iab}^J, \quad (3.31)$$

where we have used the covariant derivative \mathfrak{D}_a introduced in (2.60). Equation (3.31) is valid in the absence of torsion and takes the exact same form as its Lorentzian counterpart [46,53], and it is a nontrivial dynamical equation that determines the set of embedding functions $n_\mu^I X^\mu$. This equation, which is one of the main results of the paper, appears extensively in the context of lipid vesicles (see, e.g., Ref. [9]) but without time components.

2. Boundary conditions

In the previous section we considered the equations of motion arising from (3.23) on Σ . In this section we consider the possibility of such submanifolds having a boundary. In such cases, the second integral in (3.25) is nontrivial and gives rise to a nontrivial boundary term that must vanish, namely,

$$\int_{\partial\Sigma} d^p y e_\partial \eta_a [(\mathcal{T}^a \tau_\rho + \mathcal{T}^{ab} \bar{h}_{\rho b} - D_b \mathcal{D}^{ab}{}_\rho - \mathcal{D}^{ab}{}_I D_b n_\rho^I) \xi^\rho + \mathcal{D}^{ab}{}_I D_b \xi^I] = 0, \quad (3.32)$$

where η_a is a normal covector to the boundary while e_∂ is the integration measure on $\partial\Sigma$ (parameterized by y). With the help of the boundary completeness relation $\Pi_b^c = \delta_b^c - \eta_b \eta^c$ where $\eta^c = h^{cd} \eta_d$, the boundary term can be rewritten as

$$\int_{\partial\Sigma} d^p y e_\partial \eta_a \eta_b \mathcal{D}^{ab}{}_I \eta^c \partial_c \xi^I + \int_{\partial\Sigma} d^p y e_\partial \eta_a \{ [\mathcal{T}^a \tau_\rho + \mathcal{T}^{ab} \bar{h}_{\rho b} - D_b (\mathcal{D}^{ab}{}_I n_\rho^I) - \mathcal{D}^{ab}{}_I D_b n_\rho^I] \xi^\rho + \Pi_b^c \mathcal{D}^{ab}{}_I \partial_c \xi^I \} = 0. \quad (3.33)$$

As in the case of the bulk equations of motion on Σ , normal derivatives to the boundary of the form $\eta^c \partial_c \xi^I$ cannot be integrated out. Hence the above equation splits into two independent conditions:

$$\eta_a \eta_b \mathcal{D}^{ab}{}_I \Big|_{\partial\Sigma} = 0, \quad (3.34)$$

$$\left\{ \eta_a [\mathcal{T}^a \tau_\rho + \mathcal{T}^{ab} \bar{h}_{\rho b} - D_b (\mathcal{D}^{ab}{}_I n_\rho^I) - \mathcal{D}^{ab}{}_I D_b n_\rho^I] - n_\rho^I \Pi_c^d D_d (\eta_a \mathcal{D}^{ab}{}_I \Pi_b^c) \right\} \Big|_{\partial\Sigma} = 0. \quad (3.35)$$

The first boundary condition in (3.34) is a consequence of $\text{SO}(d-p)$ invariance of the action and can also be derived by keeping track of boundary terms when using (2.65) in (3.23). The second of these conditions can be projected tangentially

and transversely to Σ , yielding, respectively,

$$\eta_a [\mathcal{T}^a \tau_c + \mathcal{T}^{ab} \bar{h}_{bc} + 2\mathcal{D}^{ab}{}_I K_{bc}^I]_{\partial\Sigma} = 0, \\ [\mathcal{D}^{ab}{}_J \Pi_b^c D_c \eta_a - 2\mathfrak{D}_b (\eta_a \mathcal{D}^{ab}{}_J)]_{\partial\Sigma} = 0, \quad (3.36)$$

where we have used the first boundary condition (3.34) as well as $\eta_a \tau_b \mathcal{T}^{ab} \Big|_{\partial\Sigma} = 0$, which is a consequence of the $U(1)$ invariance of (3.23). These boundary conditions can be further projected along h^{cd} and \hat{v}^c , leading to

$$[\eta_a \mathcal{T}_m^{ad} + 2\eta_a \mathcal{D}^{b[a}{}_I h^{d]c} K_{bc}^I]_{\partial\Sigma} = 0, \eta_a \mathcal{T}_m^a \Big|_{\partial\Sigma} = 0, \quad (3.37)$$

where \mathcal{T}_m^{ad} and \mathcal{T}_m^a were introduced in (3.29) and (3.30), respectively. This completes the analysis of the equations of motion and its boundary conditions. In the specific examples below, however, we will not consider the presence of boundaries.

IV. APPLICATIONS TO SOFT MATTER SYSTEMS

In this section we apply the action formalism in order to describe equilibrium fluid membranes and lipid vesicles as well as their fluctuations. These systems are such that their deformations, at mesoscopic scales, are described by purely geometric degrees of freedom (see, e.g., Ref. [9]) and few material or transport coefficients, such as the bending modulus κ . The development of Newton-Cartan geometry for surfaces in the previous sections brings several advantages to the description of these systems. First, it introduces absolute time and therefore fluctuations of the system can include temporal dynamics in a covariant form. Second, the symmetries of the problem are manifested via the geometry of the submanifold or ambient spacetime.²¹

More importantly, however, is perhaps the fact that NC geometry allows to properly introduce thermal field theory of equilibrium fluid membranes. Material coefficients such as κ are functions of the temperature T (see, e.g., Ref. [19]) but also of the mass density μ . However, the fact that T and μ can be given a geometric interpretation, via the hydrostatic partition function approach, in which case they are associated with the existence of a background isometry (or timelike Killing vector field), is disregarded in all models of lipid vesicles. However this approach is required in order to understand the correct equations that describe fluctuations. We begin with a simple fluid membrane with only surface tension in order to elucidate these fundamental aspects and end with a generalization of the Canham-Helfrich model.

A. Fluid membranes

In this section we consider equilibrium fluid membranes, by which we mean stationary fluid configurations that live

²¹This point is reminiscent of the strategy adopted by Son *et al.* in Refs. [41,42,60] where the authors take advantage of the fact that Newton-Cartan geometry is the natural geometric arena for the effective description of the fractional quantum Hall effect. In this way, by coupling a suitable field theory to Newton-Cartan geometry, information about correlation functions involving mass, energy and momentum currents can be extracted via geometric considerations.

on some arbitrary surface.²² As mentioned above, equilibrium requires the existence of an ambient timelike Killing vector field k^μ such that the fluid configuration is time independent. In general, since we wish to describe fluids that are rotating or boosted along some directions, equilibrium requires the existence of a set of symmetry parameters $K = (k^\mu, \lambda_\mu^K, \Lambda^K)$ such that the transformation on the NC triplet [cf. Eqs. (2.4) and (2.5)] vanishes,

$$\begin{aligned} \mathcal{L}_k \tau_\mu &= 0, & \mathcal{L}_k \bar{h}_{\mu\nu} &= 2\tau_{(\mu} \mathcal{L}_k m_{\nu)} + 2\tau_{(\mu} \partial_{\nu)} \Lambda^K, \\ \mathcal{L}_k m_\mu + \lambda_\mu^K + \partial_\mu \Lambda^K &= 0, \end{aligned} \quad (4.1)$$

and whose pullback $k^a = u_\mu^a k^\mu$ is also a submanifold Killing vector field satisfying the relations

$$\begin{aligned} \mathcal{L}_k \tau_a &= 0, & \mathcal{L}_k \bar{h}_{ab} &= 2\tau_{(a} \mathcal{L}_k \check{m}_{b)} + 2\tau_{(a} \partial_{b)} \Lambda^K, \\ \mathcal{L}_k \check{m}_a + \check{\lambda}_a^K + \partial_a \Lambda^K &= 0. \end{aligned} \quad (4.2)$$

These relations make sure that the space in which the fluid lives does not depend on time.

The simplest example of k^μ in flat NC space (2.15) is the case of a static Killing vector where $k^\mu = \delta_t^\mu$.²³ Since the fluid is in equilibrium, it is straightforward to construct an Euclidean free energy²⁴ from the action S by Wick rotation $t \rightarrow it$, compactification of t with period $1/T_0$ and integration over the time circle, where T_0 is the constant global temperature. This means that the Euclidean free energy \mathcal{F} is given by

$$\mathcal{F}[\tau_a, \bar{h}_{ab}, K_{ab}^I] = T_0 S_{t \rightarrow it}. \quad (4.3)$$

Given the transformations (4.1) and (4.2), the free energy can depend on two scalars, the local temperature T and chemical potential μ (associated with particle number conservation), defined in terms of the symmetry parameters as

$$T = \frac{T_0}{k^a \tau_a}, \quad \frac{\mu}{T} = \frac{\Lambda^K}{T_0} + \frac{1}{2T} \bar{h}_{ab} u^a u^b, \quad u^b = \frac{k^b}{k^a \tau_a}, \quad (4.4)$$

where u^μ is the fluid velocity.²⁵ We will now look at different cases.

²²We follow previous constructions of relativistic [46,61–63] and nonrelativistic fluids [33,34].

²³Specific surfaces where the fluid lives, besides a timelike isometry, may have additional translational or rotational isometries. In such situations the Killing vector k^μ can have components along those spatial directions. The chemical potential μ introduced in (4.4) captures the spatial norm of the Killing vector, which is associated with the presence of linear or angular momenta.

²⁴This is also referred to as hydrostatic partition function $-i \ln \mathcal{Z} = T_0 \mathcal{F}$ [61,62].

²⁵The free energy considered here only depends on geometric quantities such as T and μ , where the Killing vector K^μ and the gauge parameter Λ^K solve (4.2). It is possible to promote the free energy to an effective action that does not require time independence by treating S as also being dependent on an arbitrary vector β^μ and gauge parameter Λ (see Ref. [64]).

1. Surface tension

The simplest example of a fluid membrane is one in which the action depends only on the surface tension $\chi(T, \mu)$. Such an action describes, for instance, soap films. Thus the free energy (4.3) takes the form

$$\mathcal{F} = \int_{\Sigma_s} d^p \sigma e_s \chi(T, \mu), \quad (4.5)$$

where Σ_s and e_s denote the spatial part of Σ and the volume form e , respectively, due to integration over the time direction. We can now use (3.23) to extract the currents at fixed symmetry parameters. It is useful to explicitly evaluate the variations

$$\delta T = -T u^a \delta \tau_a, \quad \delta \mu = \frac{\Lambda^K}{T_0} \delta T + \frac{1}{2} u^a u^b \delta \bar{h}_{ab} + \bar{u}^2 \frac{\delta T}{T}, \quad (4.6)$$

where we have defined $\bar{u}^2 = \bar{h}_{ab} u^a u^b$. This allows us to derive the variation of the surface tension as

$$\delta \chi = s \delta T + n \delta \mu = -\left(Ts + n\mu + \frac{n}{2} \bar{u}^2\right) u^a \delta \tau_a + \frac{n}{2} u^a u^b \delta \bar{h}_{ab}, \quad (4.7)$$

where we have defined the surface entropy density and surface particle number density (mass density) as

$$s = \left(\frac{\partial \chi}{\partial T}\right)_\mu, \quad n = \left(\frac{\partial \chi}{\partial \mu}\right)_T. \quad (4.8)$$

From (4.7) we also directly extract the Gibbs-Duhem relation $d\chi = s dT + n d\mu$. Using (4.7) we also determine the currents

$$\mathcal{T}^a = -\chi \hat{v}^a - \left(\varepsilon + \chi + \frac{n}{2} \bar{u}^2\right) u^a, \quad \mathcal{T}^{ab} = \chi h^{ab} + n u^a u^b, \quad (4.9)$$

where we have defined the internal energy ε via the Euler relation $\varepsilon + \chi = Ts + n\mu$. This defines the constitutive relations of a Galilean fluid living on a submanifold in an ambient NC spacetime. Using the stress-mass tensor in (4.9), the nontrivial shape equation (3.31) in the absence of bending moment becomes

$$\mathcal{T}^{ab} K_{ab}^I = 0 \Rightarrow \chi K^I + n u^a u^b K_{ab}^I = 0. \quad (4.10)$$

Physically relevant fluid membranes are codimension one and so we can omit the transverse index I . The shape equation (4.10) expresses the balance of forces between the surface tension χK (normal stress) and the normal acceleration $n u^a u^b K_{ab}$ of the fluid.²⁶ If we would consider a surface tension with no dependence on the temperature and chemical potential, then $n = 0$ and the shape equation reduces to the equation of a minimal surface. To complete the thermodynamic interpretation of (4.5), we note that varying the free energy with respect to the global temperature T_0 gives rise to the global entropy

$$S = \frac{\partial \mathcal{F}}{\partial T_0} = \int_{\Sigma_s} d^p \sigma e_s \frac{s}{k^a \tau_a} = \int_{\Sigma_s} d^p \sigma e_s s u^a t_a, \quad (4.11)$$

²⁶Using the definition of extrinsic curvature (2.61), we can rewrite $u^a u^b K_{ab}^I = n_\mu^I u^\nu \nabla_\nu u^\mu$. Hence the second term in (4.10) is in fact the normal component of the acceleration of the fluid $u^\nu \nabla_\nu u^\mu$ where $u^\mu = u_a^\mu u^a$. If the fluid is rotating along the surface, this term gives rise to centrifugal acceleration.

where we have defined the timelike vector $t_a = \tau_a / (k^b \tau_b)$, and where su^a is the entropy current.

2. Surface fluctuations: Elastic waves

The shape equation (4.10) describes equilibrium configurations of fluid membranes in the absence of any bending moment. We consider a fluid at rest in the simplest scenario of a surface with two spatial dimensions embedded in a NC spacetime with 3 spatial dimensions such that $\tau_a = \delta_a^t$ where $a = t, 1, 2$. The fluid thus has a velocity $u^a = (1, 0, 0)$. Such a trivial time embedding, $\tau_a = \delta_a^t$, is typically the most physically relevant setting for soft matter applications. In this context, we have that $u^a u^b K_{ab} = 0$ since $K_{tb} = 0$ trivially. Thus, the second term in (4.10) does not contribute in equilibrium, and it is acceptable to simply ignore the fact that the surface tension depends on the temperature and chemical potential. However, if one is interested in fluctuations away from equilibrium, the second term in (4.10) cannot be ignored. Here we consider the simplest case where the surface is flat and hence also trivially embedded in space such that

$$h_{ab} = \delta_a^i \delta_b^i, \quad m_a = 0, \quad n_\mu = \delta_\mu^3. \quad (4.12)$$

This is an equilibrium configuration that trivially solves (4.10) since $K_{ab} = 0$.

We now consider a small fluctuation of the embedding map along the normal direction $X^3 = X^\perp$. Using (3.10) we find

$$\delta_X \mathcal{T}^{ab} K_{ab} + \mathcal{T}^{ab} \delta_X K_{ab} = (\chi h^{ab} + nu^a u^b) \partial_a \partial_b \xi^\perp = 0, \quad (4.13)$$

where we have used that $K_{ab} = 0$ to eliminate the first term and converted $\mathfrak{D}_a \rightarrow \partial_a$ as we are dealing with a flat surface in a flat ambient space. Equation (4.13) is a wave equation, and considering wavelike solutions of the form $\xi^\perp \sim e^{-i\omega t + i(k_1 \sigma_1 + k_2 \sigma_2)}$ one finds the linear dispersion relation

$$\omega = \pm \sqrt{\frac{-\chi}{n}} k, \quad (4.14)$$

where ω is the frequency, k_1, k_2 are wave numbers, and $k^2 = k_1^2 + k_2^2$.²⁷ This is the classical answer for the oscillations of uniform elastic sheets (see, e.g., Ref. [65]).

This result shows the importance of considering NC geometry in the theory of fluid membranes, since omitting the dependence of the surface tension on the temperature and chemical potential would not have allowed for the derivation of (4.14). We note that the result (4.14) is valid for any type of elastic membrane with mass density and does not require any ‘‘flow’’ on the membrane, in particular the initial equilibrium configuration was static $u^a = (1, 0, 0)$.²⁸ In a future publication, we will consider a more general analysis of fluctuations of fluid membranes which will also include the Canham-Helfrich model [54].

²⁷Note that in order to match conventions with the classical literature one should redefine $\chi \rightarrow -\chi$.

²⁸If one was describing an elastic material, the surface tension would also be dependent on the Goldstone modes of broken translations and hence on intrinsic elastic moduli.

3. Droplets

Here we briefly consider the case of a droplet (or soap bubble) in which the fluid membrane encloses some volume with uniform internal pressure P_{int} separating it from an exterior medium with uniform external pressure P_{ext} . In order to describe these situations we augment the action with the bulk pieces

$$S_{\text{bulk}} = \int_{\text{int}(\Sigma)} d^{d+1}x e_b P_{\text{int}} + \int_{\text{ext}(\Sigma)} d^{d+1}x e_b P_{\text{ext}}, \quad (4.15)$$

where e_b is the bulk measure and $\text{int}(\Sigma)$ is the interior of the closed surface Σ ,²⁹ whereas $\text{ext}(\Sigma)$ is the exterior region of the bulk outside the surface. The variation of the density e_b with respect to a bulk (or ambient spacetime) diffeomorphism reads

$$\delta_\xi e_b = \partial_\mu (e_b \xi^\mu), \quad (4.16)$$

which, using Stokes theorem, implies that the variation takes the form

$$\delta_\xi S_{\text{bulk}} = -\Delta p \int_\Sigma d^d \sigma n_\mu \xi^\mu, \quad (4.17)$$

where $\Delta p = P_{\text{ext}} - P_{\text{int}}$ is the constant pressure difference across the surface Σ .³⁰ In a biophysical context, where the pressure difference is attributable to two different chemical solutions separated by a semipermeable membrane, this pressure is the *osmotic pressure* [66].

From (4.17), we deduce that S_{bulk} does not contribute to the intrinsic equations of motion, while it adds the constant term $-\Delta p$ to the shape equation (4.10) such that

$$\mathcal{T}^{ab} K_{ab} = \chi K + nu^a u^b K_{ab} = -\Delta p. \quad (4.18)$$

This is a generalization of the Young-Laplace equation, which includes the possibility of the fluid having nontrivial acceleration, and was first derived in Ref. [45] in the context of null reduction.

B. The Canham-Helfrich model revisited

In this section we consider a more elaborate case of fluid membranes, namely, that of the Canham-Helfrich model [1, 2]. This model describes equilibrium configurations of biophysical membranes (see, e.g., Ref. [6]) comprised of a phospholipid bilayer [67], and captures several shapes of biophysical interest [6], namely, the sphere (corresponding to spherical vesicles such as liposomes), the torus (toroidal vesicles) and the biconcave discoid (the red blood cell or *erythrocyte*). This model includes, besides the presence of a surface tension χ , also the bending modulus κ that incorporates the bending energy of the membrane. We show how to describe this model within Newton-Cartan geometry and generalize it by allowing the material parameters to be functions of T, μ . We also review the family of classical lipid vesicles (spherical,

²⁹By a closed surface we mean a NC submanifold whose constant time slices are closed.

³⁰In order to describe gases or fluids in the interior or exterior, one should consider the dependence of internal or external pressures on bulk temperature and chemical potential as in Ref. [45].

toroidal, discoid) within this framework. We leave a more detailed analysis of this model and its generalizations to a future publication [54].

1. Generalized Canham-Helfrich model

The Canham-Helfrich model contains quadratic terms in the extrinsic curvature and a set of material coefficients. It describes lipid vesicles in thermal equilibrium. As in the previous section, a proper description of such systems requires taking into account the dependence of the material coefficients on the temperature and chemical potential. As a starting point we take the more general free energy

$$\mathcal{F}_{\text{CH}} = \int_{\Sigma_s} d^p \sigma e_s [a_0(T, \mu) + a_1(T, \mu)K + a_2(T, \mu)K^2 + a_3(T, \mu)K \cdot K], \quad (4.19)$$

where $\{a_0, a_1, a_2, a_3\}$ is a set of material coefficients characterizing the phenomenological specifics of the biophysical system under scrutiny. In the expression above, we have defined $K \cdot K = h^{ac}h^{bc}K_{ab}K_{cd}$.

It is well known that the last term in (4.19) can usually be ignored due to the Gauss-Codazzi equation (2.73) in flat ambient space, as it can be related to the Gaussian curvature of the membrane and hence integrated out for two-dimensional surfaces (see Appendix D for details). However, this is possible only if a_3 is treated as a constant. Since a proper geometric and thermodynamic treatment requires promoting a_3 to a nontrivial function of T, μ this implies that new nontrivial contributions to the equations of motion will appear. Additionally, based solely on effective field theory reasoning, it is possible to augment (4.19) with further terms involving the fluid velocity (see [46] for the relativistic case). We will leave a thorough analysis of this for the future [54]. Here we focus on extracting the stresses on the membrane using (3.23).

We find the energy current

$$\mathcal{T}^a = -(a_0 + a_1K + a_2K^2 + a_3K \cdot K)\delta^a - (L_0 + L_1K + L_2K^2 + L_3K \cdot K)u^a, \quad (4.20)$$

where we have defined the thermodynamic parameters

$$L_i = T s_i + n_i \mu + \frac{n_i}{2} \bar{u}^2, \quad s_i = \left(\frac{\partial a_i}{\partial T} \right)_\mu, \quad n_i = \left(\frac{\partial a_i}{\partial \mu} \right)_T. \quad (4.21)$$

Similarly, we extract the Cauchy stress-mass tensor

$$\begin{aligned} \mathcal{T}^{ab} = & h^{ab}(a_0 + a_1K + a_2K^2 + a_3K \cdot K) \\ & - 2h^{ac}h^{bd}K_{cd}(a_1 + 2a_2K) - 4a_3h^{fd}h^{ca}h^{eb}K_{cd}K_{ef} \\ & + (n_0 + n_1K + n_2K^2 + n_3K \cdot K)u^a u^b. \end{aligned} \quad (4.22)$$

As this model contains terms involving the extrinsic curvature, it has a bending moment of the form

$$\mathcal{D}^{ab} = a_1 h^{ab} + \mathcal{Y}^{abcd} K_{cd}, \quad \mathcal{Y}^{abcd} = 2a_2 h^{ab} h^{cd} + 2a_3 h^{a(c} h^{d)b}, \quad (4.23)$$

where \mathcal{Y}^{abcd} is the Young modulus of the membrane and has the usual symmetries of a classical elasticity tensor.³¹ Equations (4.22) and (4.23) demonstrate that if a_3 is a nontrivial function of T, μ , then it will contribute nontrivially to the shape equation (3.31).

Let us be a bit more precise about the role of a_3 . First, we redefine the coefficient a_2 as $a_2 = \tilde{a}_2 - a_3$ so that a_3 now multiplies the integrand of the Gauss-Bonnet term, the Gaussian curvature. All terms proportional to a_3 in the shape equation can be shown to cancel identically using a set of identities such as the Codazzi-Mainardi and Gauss-Codazzi equations [i.e., (2.70) and (2.73) suitably adapted to the case of a codimension one submanifold] as well as the identity (D6) which expresses the fact that the Einstein tensor of the Riemannian geometry on constant time slices vanishes in two dimensions. This means that a_3 will contribute only to the shape equation through its derivatives that we denoted by s_3 and n_3 . There are only two such terms, $n_3 K \cdot K u^a u^b K_{ab}$ and $(h^{ac}h^{bd}K_{cd} - h^{ab}K)D_a D_b a_3$. In particular the latter is interesting since it will make a contribution to the shape equation even in the case of a static fluid.

We now show how the model (4.19) recovers the standard Canham-Helfrich model.

2. The standard Canham-Helfrich model

We focus on three-dimensional flat spacetime (2.15) and surfaces with two spatial dimensions. We also assume that the functions $\{a_0, a_1, a_2, a_3\}$ are constant. In this case, as explained above and detailed in Appendix D, we can set $a_3 = 0$. Additionally, we require the free energy (4.19) to be invariant under a change of the inwards and outwards orientation of normal vectors, that is, invariant under $n^\mu \rightarrow -n^\mu$. This leads to

$$\mathcal{F}_{\text{CH}} = \int_{\Sigma_s} d^2 \sigma e_s [\chi + \kappa(K + c_0)^2], \quad (4.24)$$

where we have redefined the coefficients such that

$$a_0 = \chi + \kappa c_0^2, \quad a_1 = 2\kappa c_0, \quad a_2 = \kappa, \quad (4.25)$$

and where c_0 changes sign under $n^\mu \rightarrow -n^\mu$. This is the direct analog of the Canham-Helfrich model of lipid bilayer membranes [2]. The constant c_0 is the *spontaneous curvature*, which reflects a preference to adopt a specific curvature due to, e.g., different aqueous environments or lipid densities on the two sides of the bilayer [69]. The parameter χ is the surface tension and the parameter κ is the *bending modulus* [6]. In this case, $s_i = n_i = 0$ and the shape equation (3.31) upon using (4.22) and (4.23) becomes

$$\begin{aligned} & -a_0K - a_1K^2 - a_2K^3 + a_1K \cdot K + 2a_2K(K \cdot K) \\ & + 2a_2h^{ab}D_a D_b K - \Delta p = 0, \end{aligned} \quad (4.26)$$

where we have added the contribution from constant interior and exterior pressures as in Sec. IV A 3. We will now review particular solutions to this model.

³¹This was first introduced in an effective theory for relativistic fluids in Ref. [46]. The Young modulus tensor also appears when considering finite size effects in the dynamics of black branes [68].

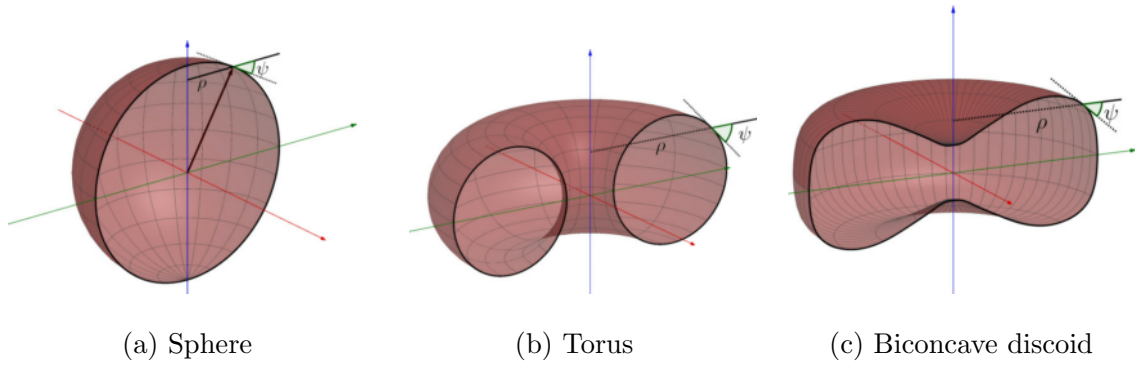


FIG. 2. The three axisymmetric biophysical solutions to the Canham-Helfrich model and how they arise as surfaces of revolution. The coordinate ρ measures the perpendicular distance to the z axis (blue), and ψ is the angle between the tangent of the contour and the ρ axis.

3. Biophysical solutions: Axisymmetric vesicles

Here we discuss three well-known axisymmetric solutions of the Canham-Helfrich model [6] (the spherical vesicle, the toroidal vesicle, and the red blood cell) and how they are described within this approach. These surfaces arise as surfaces of revolution and therefore a particularly convenient way of parametrizing these is to consider a “cross-sectional contour” described by the perpendicular distance ρ to the symmetry axis (which we will take to be the z axis) and the angle ψ , which is the angle between the tangent of the contour and the ρ axis (see Fig. 2 for a graphical depiction). This gives us the relation $\tan \psi(\rho) = \frac{dz}{d\rho}$. The entire surface is then obtained by rotating this contour such that

$$X^\mu = \begin{pmatrix} t \\ \rho \cos \phi \\ \rho \sin \phi \\ z_0 + \int_0^\rho d\tilde{\rho} \tan \psi(\tilde{\rho}) \end{pmatrix}, \quad (4.27)$$

which in turn gives rise to

$$K = -\frac{\sin \psi(\rho)}{\rho} - \cos \psi(\rho) \psi'(\rho),$$

$$K \cdot K = \frac{\sin^2 \psi(\rho)}{\rho^2} + \cos^2 \psi(\rho) [\psi'(\rho)]^2. \quad (4.28)$$

Spherical vesicle:. A sphere of radius R [see Fig. 2(a)] is described by

$$\sin \psi(\rho) = \frac{\rho}{R}, \quad (4.29)$$

which gives rise to the equation

$$0 = \Delta p R^2 + 4c_0 \kappa + 2c_0^2 R \kappa + 2R \lambda. \quad (4.30)$$

$$0 = [\kappa a^3 - 2\kappa a^2 b - 4\kappa a b^2 + 4\kappa a b c_0 - a(\kappa c_0^2 + \chi) + 4\kappa b^2 c_0 - 2b(\kappa c_0^2 + \chi) + \Delta p] \\ + \log \rho [-2\kappa a^3 - 8\kappa a^2 b + 4\kappa a^2 c_0 + 8\kappa a b c_0 - 2a(\kappa c_0^2 + \chi)] + \log^2 \rho (-4a^3 \kappa + 4a^2 \kappa c_0), \quad (4.35)$$

As was also pointed out in Ref. [6], this has two solutions when viewed as an equation for the radius, provided that $\Delta p < 0$ and $-4c_0 \kappa \Delta p + (\kappa c_0^2 + \chi)^2 > 0$. The first condition reflects the fact that the internal pressure must be greater than the external pressure to stabilize the structure.

Torus:. The torus can also be obtained as a surface of revolution [Fig. 2(b)]. This is achieved via

$$\sin \psi(\rho) = \frac{1}{r} \rho + \frac{R}{r}, \quad (4.31)$$

where R is the major axis and r the minor axis. From this, we get the shape equation

$$0 = (-\kappa R^3 + 2\kappa r^2 R) + \rho^2 [r^2 R (-\kappa c_0^2 - \chi) - 4\alpha c_0 r R] \\ + \rho^3 [-2r^2 (\kappa c_0^2 + \chi) + 4\kappa c_0 r + \Delta p r^3]. \quad (4.32)$$

Each coefficient of $\{\rho^0, \rho^2, \rho^3\}$ must vanish independently, giving us three equations

$$R = \sqrt{2}r, \quad \chi = \frac{\kappa c_0 (4 - c_0 r)}{r}, \quad \Delta p = \frac{4\kappa c_0}{r^2}. \quad (4.33)$$

The first of these predicts a universal ratio between the major and minor axes. Theoretically predicted in Ref. [70], this ratio was observed experimentally in Ref. [71] with high precision.

Biconcave discoid:. The biconcave discoid [Fig. 2(c)] is the shape of the red blood cell. This axisymmetric vesicle is described by

$$\sin \psi(\rho) = a\rho(\log \rho + b), \quad (4.34)$$

where a, b are parameters that are related to the characteristics of the discoid.³² The resulting equation of motion is

which again gives three equations. These equations yield

$$a = c_0, \quad \chi = \Delta p = 0. \quad (4.36)$$

Thus, we recover the result that the biconcave shape of the red blood cell relies on isotonicity, i.e., that the pressures on each side of the membrane are equal [66] (see also Ref. [72]).

³²For example, the radius of the discoid, i.e., the maximum value of $\rho = \rho_R$, is implicitly given by $1 = a\rho_R(\log \rho_R + b)$, since $\psi(\rho_R) = \pi/2$ (see also Ref. [66]).

V. DISCUSSION AND OUTLOOK

The majority of the work presented here was of a foundational nature. In order to describe the physical properties of fluid membranes in thermodynamic equilibrium, we developed the submanifold calculus for Newton-Cartan geometry. This parallels how the submanifold calculus of (pseudo-) Riemannian or Euclidean geometry is a prerequisite for formulating and varying the standard Canham-Helfrich bending energy. We identified the geometric structures characterizing timelike submanifolds in NC geometry³³ and obtained the associated integrability conditions. Deriving expressions for the infinitesimal variations and transformation properties of the basic objects allowed us to formulate a generic extremization problem for broad classes of NC surfaces, including fluid membranes whose equilibrium configurations only depend on geometric properties.³⁴

In Sec. IV we applied this toolbox that we developed to the description of fluid membranes in thermodynamic equilibrium. The unique aspect of these applications is that the dependence on temperature and chemical potential of material coefficients, such as surface tension and bending modulus, is critical for the emergence of wave excitations. This relied on the fact that temperature and chemical potential have a geometric interpretation related to the existence of a timelike isometry in the ambient spacetime. Standard examples of free energies such as the Canham-Helfrich bending energy are straightforwardly generalized by taking into account the geometric interpretation of thermodynamic variables. The resulting free energies are still purely geometric but the derived stresses on the membrane are different than standard results found in the literature. In particular, the Gaussian bending modulus can play a role in the shape of lipid vesicles since the Gaussian curvature cannot be integrated out when material coefficients are not constant. The resulting stresses produce elastic waves when perturbing away from equilibrium thus providing the correct dynamics of fluid membranes.

This paves the way for tackling several open questions, which we plan to address in a future publication [54]:

(1) The fact that the Gaussian curvature cannot be integrated out in thermal equilibrium suggests that the family of closed lipid vesicles reviewed in Sec. IV B 3 should be revisited and the effects of the Gaussian bending modulus should be considered [i.e., a_3 in (4.19)], including the effects on deviations away from equilibrium.

(2) The lipid vesicle solutions in Sec. IV B 3 are static solutions, in which $u^a = (1, 0, 0)$. However, in principle such

solutions can sustain rotation along the direction ϕ . The question is thus: is it possible to obtain lipid vesicles with stationary flows?

(3) From an effective field theory point of view, the Canham-Helfrich bending energy (4.19) does not contain all possible responses that take into account thermal equilibrium. For instance a term quadratic in the extrinsic curvature of the form $u^a u^b h^{cd} K_{bc} K_{ad}$ involving the fluid velocity can be added to (4.19) (similarly to its relativistic counterpart [46]). However, there are further couplings that involve derivatives of u_a such as the square of the fluid acceleration $(u^a \mathcal{D}_a u^b)^2$ or the square of the vorticity. Some of these terms are related to the Gaussian curvature and thus, by the Gauss-Codazzi equation (2.73), to combinations of squares of the extrinsic curvature. Therefore, from an effective theory point of view, they cannot be ignored *a priori*.

(4) We have shown in Sec. IV A 2 that taking into account the geometric definitions of temperature and mass chemical potential in equilibrium gives rise to the correct dispersion relation for an elastic membrane when perturbing away from equilibrium. It would now be interesting to consider perturbations away from equilibrium solutions of the Canham-Helfrich model (4.19) using the stresses (4.20)–(4.23). This would shed light on the stability of lipid vesicles.

(5) The construction of effective actions or free energies in the manner described in this work is appropriate to describe equilibrium configurations. However, including different types of dissipation [78], either due to viscous flows or diffusion of embedded proteins is of interest [8]. In order to include dissipation from an effective action point of view one could consider the more elaborate Schwinger-Keldysh framework [79–81] and adapt it to nonrelativistic systems. Alternatively, one may construct the effective theory in a long-wavelength hydrodynamic expansion by classifying potential terms appearing in the currents \mathcal{T}^a and \mathcal{T}^{ab} and obtaining constitutive relations (see, e.g., Refs. [82,83]). We plan on addressing this in the near future.

(6) We focused on extrinsic curvature terms in effective actions (3.23), but it would also be interesting to consider the effect of the external rotation tensor (2.63). In the (pseudo-) Riemannian or Euclidean setting, this corresponds to spinning point particles/membranes [46,53,84,85] and are directly related to the Frenet curvature and Euler elastica (see, e.g., Refs. [86–88] for a recent discussion).

(7) In Secs. II and III we formulated the description of a single surface in Newton-Cartan geometry for which the scalars X^μ can be seen as Goldstone modes of spontaneous broken translations at the location of the surface. It would be interesting to extend this further to the case of a foliation of surfaces, in which case the scalars X^μ form a lattice and can be used to describe viscoelasticity as in Ref. [89].

In this work we considered Newton-Cartan geometry but there are many other types of non-Lorentzian geometries depending on the space-time symmetry group, which can be, e.g., Lifshitz, Schrödinger, or Aristotelian, which have direct applications for the hydrodynamics of strongly correlated electron systems as well as for the hydrodynamics of flocking behavior and active matter [37–39,90]. In these contexts, it is required to develop the mathematical description of submanifolds within these different types of ambient spacetimes.

³³The case of spacelike submanifolds is also interesting to pursue as it can be useful for understanding entanglement entropy in nonrelativistic field theories [73].

³⁴It would be interesting to understand the connection between this work and other recently considered constructions involving extended objects embedded in Newton-Cartan spacetime (or related geometries), such as nonrelativistic strings [30–32,74], nonrelativistic D-branes [75], and Newton-Cartan p -branes [76]. It would also be interesting to connect this work to Ref. [77], where the boundary description of quantum Hall states involves a notion of Newton-Cartan submanifolds.

The description of surfaces within these geometries will be of interest for surface or edge physics in hard condensed matter.

ACKNOWLEDGMENTS

We thank L. Giomi and R. S. Green for useful discussions. J.A. is partly supported by the Netherlands Organization for Scientific Research (NWO). The work of J.H. is supported by the Royal Society University Research Fellowship ‘‘Non-Lorentzian Geometry in Holography’’ (Grant No. UF160197). The work of E.H. is supported by the Royal Society Research Grant for Research Fellows 2017 ‘‘A Universal Theory for Fluid Dynamics’’ (Grant No. RGF\R1\180017). The work of N.O. is supported in part by the project ‘‘Towards a Deeper Understanding of Black Holes with Non-relativistic Holography’’ of the Independent Research Fund Denmark (Grant No. DFF-6108-00340) and by the Villum Foundation Experiment project 00023086.

APPENDIX A: NULL REDUCTION OF RIEMANNIAN SURFACES AND PERFECT FLUIDS

In this Appendix we provide a completely different approach to formulating the theory of surfaces and fluid membranes in Newton-Cartan geometry. This approach consists in starting from relativistic surfaces and fluid membranes and performing a null reduction so as to obtain results in NC geometry. The purpose of this technical Appendix is to provide a nontrivial check of the main results in the core of this paper.

1. Submanifolds from null reduction

It is well known that any Newton-Cartan geometry can be obtained as the null reduction of a Lorentzian manifold in one dimension higher equipped with a null killing vector [28,36,91]. Therefore, if we choose a timelike submanifold in a Lorentzian geometry such that the null Killing vector is tangent to the submanifold, its null reduction provides us with a Newton-Cartan submanifold embedded in a Newton-Cartan ambient spacetime. We illustrate this in the commuting diagram below:

$$\begin{array}{ccc}
 (\widehat{\Sigma}_{p+2}, \hat{\gamma}) & \xleftarrow{\hat{v}_a^\mu} & (\widehat{\mathcal{M}}_{d+2}, \hat{g}) \\
 \text{null red.} \downarrow & & \downarrow \text{null red.} \\
 (\Sigma_{p+1}, \{\tau|_{\Sigma}, \check{h}, \check{m}\}) & \xleftarrow{u_a^\mu} & (\mathcal{M}_{d+1}, \{\tau, h, m\})
 \end{array} \tag{A1}$$

In Sec. II B we described how to go from the NC manifold $(\mathcal{M}_{d+1}, \{\tau, h, m\})$ to the NC submanifold $(\Sigma_{p+1}, \{\tau|_{\Sigma}, \check{h}, \check{m}\})$, while passing from the Lorentzian manifold $(\widehat{\mathcal{M}}_{d+2}, \hat{g})$ to the Newton-Cartan manifold $(\mathcal{M}_{d+1}, \{\tau, h, m\})$ is achieved by null reduction.

In this Appendix, we will traverse the other route: our goal is to go from $(\widehat{\mathcal{M}}_{d+2}, \hat{g})$ to $(\Sigma_{p+1}, \{\tau|_{\Sigma}, \check{h}, \check{m}\})$ via $(\widehat{\Sigma}_{p+2}, \hat{\gamma})$. The procedure to go from $(\widehat{\mathcal{M}}_{d+2}, \hat{g})$ to $(\widehat{\Sigma}_{p+2}, \hat{\gamma})$ is nothing but the theory of submanifolds in Lorentzian geometry and is well known (see, e.g., Refs. [46,53]). We coordinatise $\widehat{\mathcal{M}}_{d+2}$

with $x^{\hat{\mu}} = (u, x^\mu)$ and $\widehat{\Sigma}_{p+2}$ with $\hat{\sigma}^{\hat{a}} = (w, \sigma^a)$. The metric on $\widehat{\mathcal{M}}_{d+2}$ can, by assumption, be written in null reduction form

$$ds_{\widehat{\mathcal{M}}_{d+2}}^2 = \hat{g}_{\hat{\mu}\hat{\nu}} dx^{\hat{\mu}} dx^{\hat{\nu}} = 2\tau_\mu dx^\mu (du - m_\nu dx^\nu) + h_{\mu\nu} dx^\mu dx^\nu. \tag{A2}$$

This line element is invariant under the Newton-Cartan gauge transformations (2.4) and conversely all gauge invariance of this line element are of the form (2.4). The invariance under the $U(1)$ transformation with parameter $\sigma(x^\mu)$ requires that we vary the higher-dimensional coordinate u as $\delta u = \sigma$. From the higher-dimensional perspective this corresponds to a diffeomorphism that leaves the x^μ unaffected but that shifts u by some function of x^μ .

The Lorentzian submanifold is defined via a set of embedding maps $\hat{X}^{\hat{\mu}}(\sigma^{\hat{a}})$ in the usual way. We define the projector

$$\hat{P}_\nu^{\hat{\mu}} = \hat{u}_a^{\hat{\mu}} \hat{v}_\nu^{\hat{a}} = \delta_\nu^{\hat{\mu}} - \hat{n}_\rho^{\hat{\mu}} \hat{n}_\nu^{\hat{\rho}} \delta_{IJ} \hat{g}^{\hat{\rho}\hat{\mu}}, \tag{A3}$$

where $\hat{n}_\rho^{\hat{\mu}}$ are the normal 1-forms to $\widehat{\Sigma}_{p+2}$ and where $\hat{u}_a^{\hat{\mu}} = \partial_{\hat{\sigma}^{\hat{a}}} \hat{X}^{\hat{\mu}}$. We require that the null direction is shared between $\widehat{\mathcal{M}}_{d+2}$ and $\widehat{\Sigma}_{p+2}$, which can be expressed as the requirements

$$\hat{u}_w^\mu = 1, \hat{u}_a^\mu = 0, \tag{A4}$$

where the null direction on the submanifold is described by w . Further, we want to impose a null reduction analog of the timelike requirement (2.24). To this end, we introduce a vector $U^{\hat{\mu}} = (\frac{\partial}{\partial u})^{\hat{\mu}} = \delta_u^{\hat{\mu}}$ so that $U_{\hat{\mu}} = (0, \tau_\mu)$. Requiring that the null Killing vector field is tangential to the submanifold $\hat{n}_\rho^{\hat{\mu}} = U^{\hat{\mu}} \hat{n}_\rho^{\hat{\mu}} = U_{\hat{\mu}} \hat{n}^{\hat{\rho}\hat{\mu}} = 0$ for all I implies the desired relation $\tau_\mu n_I^\mu = 0$ where we have identified $\hat{n}_I^{\hat{\mu}} = n_I^\mu$. This further implies that $n^{\mu\nu} = \hat{g}^{\hat{\mu}\hat{\nu}} \hat{n}_\nu^{\hat{\mu}} = h^{\mu\nu} n_\nu^\mu$ in agreement with the timelike constraint. This also implies that $\hat{P}_\nu^{\hat{\mu}} = P_\nu^\mu$, as well as the normalization $\hat{g}^{\hat{\mu}\hat{\nu}} \hat{n}_\mu^{\hat{\rho}} \hat{n}_\nu^{\hat{\sigma}} = h^{\mu\nu} n_\mu^{\hat{\rho}} n_\nu^{\hat{\sigma}} = \delta^{IJ}$. Further, the above considerations lead us to conclude that

$$\hat{n}^{\mu I} = \hat{g}^{\hat{\mu}\hat{\nu}} \hat{n}_\nu^{\hat{\rho}} = -\hat{v}^\mu n_\mu^I = -v^I. \tag{A5}$$

The metric on $\widehat{\Sigma}_{p+2}$ can also be written in null reduction form

$$\begin{aligned}
 ds_{\widehat{\Sigma}_{p+2}}^2 &= \hat{\gamma}_{\hat{a}\hat{b}} d\hat{x}^{\hat{a}} d\hat{x}^{\hat{b}} = 2\tau_a dx^a (dw - m_b dx^b) + h_{ab} dx^a dx^b \\
 &= 2\tau_a dx^a (dw - \check{m}_b dx^b) + \check{h}_{ab} dx^a dx^b,
 \end{aligned} \tag{A6}$$

where we recall the definitions of \check{h}_{ab} and \check{m}_a in (2.34) and (2.39), respectively. As manifested in the equations above, the null reduction form of the metric is Galilean boost invariant and does not distinguish between *checked* and *unchecked* metric data. In turn, the Lorentzian metric $\hat{\gamma}$ on $\widehat{\Sigma}_{p+2}$ is the pullback of the metric \hat{g} on $\widehat{\mathcal{M}}_{d+2}$, that is

$$\hat{\gamma}_{\hat{a}\hat{b}} = \hat{u}_a^{\hat{\mu}} \hat{u}_b^{\hat{\nu}} \hat{g}_{\hat{\mu}\hat{\nu}}, \tag{A7}$$

which implies that

$$\begin{aligned}
 \tau_a &= \hat{\gamma}_{aw} = \hat{u}_a^{\hat{\mu}} \hat{u}_w^{\hat{\nu}} \hat{g}_{\hat{\mu}\hat{\nu}} = \hat{u}_a^{\hat{\mu}} \hat{u}_w^{\hat{\nu}} \hat{g}_{\mu\nu} + \hat{u}_a^{\hat{\mu}} \hat{u}_w^{\hat{\nu}} \hat{g}_{\mu\nu} \\
 &= \hat{u}_a^\mu \tau_\mu + \hat{u}_a^\mu \hat{u}_w^\nu \check{h}_{\mu\nu}.
 \end{aligned} \tag{A8}$$

Thus, taking

$$\hat{u}_w^\mu = 0, \quad (\text{A9})$$

and identifying $\hat{u}_a^\mu = u_a^\mu$ we get the desired relation between the two clock 1-forms, namely, $\tau_a = u_a^\mu \tau_\mu$. Next, we consider

$$\begin{aligned} \bar{h}_{ab} &= \hat{\gamma}_{ab} = \hat{u}_a^\mu \hat{u}_b^\nu \hat{g}_{\mu\nu} = \hat{u}_a^\mu \hat{u}_b^\nu \hat{g}_{\mu\nu} + \hat{u}_a^\mu \hat{u}_b^\nu \hat{g}_{\mu\mu} + \hat{u}_a^\mu \hat{u}_b^\nu \hat{g}_{\nu\nu} \\ &= u_a^\mu u_b^\nu \bar{h}_{\mu\nu}, \end{aligned} \quad (\text{A10})$$

where we have used (A4), which again agrees with the results of Sec. II B. The relation $\hat{u}_w^\mu \hat{u}_\mu^w = \hat{u}_u^w = 1$ where we used (A9), fixes $\hat{u}_u^w = 1$. To determine \hat{u}_μ^w we bring into play the orthogonality requirement $\hat{u}_\mu^w \hat{n}^{\mu I} = \hat{g}^{\mu\nu} \hat{u}_\mu^w \hat{n}_\nu^I = 0$, which translates into the relation

$$\hat{v}^\mu n_\mu^I = \hat{u}_\mu^w n_\mu^I, \quad (\text{A11})$$

where we have used that $\hat{u}_u^w = 1$ and $n_I^\mu = \delta_{IJ} h^{\mu\nu} n_\nu^J$. This is possible only if

$$\hat{u}_\mu^w = \hat{v}^I n_\mu^I. \quad (\text{A12})$$

The null reduction of the ambient inverse metric is

$$\hat{g}^{\mu\nu} = 2\check{\Phi}, \quad \hat{g}^{\mu\mu} = -\hat{v}^\mu, \quad \hat{g}^{\mu\nu} = h^{\mu\nu}, \quad (\text{A13})$$

while the relation between \hat{g}^{-1} and $\hat{\gamma}^{-1}$ is given by $\hat{\gamma}^{\hat{a}\hat{b}} = \hat{u}_\mu^{\hat{a}} \hat{u}_\nu^{\hat{b}} \hat{g}^{\mu\nu}$. In turn, the relation $\hat{\gamma}^{ab} = h^{ab}$ requires that $\hat{u}_u^a = 0$. Using this, we can write

$$\hat{\gamma}^{wa} = \hat{u}_\mu^w \hat{u}_\nu^a \hat{g}^{\mu\nu} = \hat{v}^I n_\mu^I \hat{u}_\nu^a h^{\mu\nu} + \hat{u}_\nu^a \hat{g}^{\mu\nu}, \quad (\text{A14})$$

where we have used (A12), which leads us to identify $\hat{u}_\mu^a = u_\mu^a$ and, by the orthogonality relation (2.25), leads to $\hat{v}^a = u_\mu^a \hat{v}^\mu$ as desired. The relation (A12) furthermore implies that

$$\hat{\gamma}^{ww} = \hat{u}_\mu^w \hat{u}_\nu^w \hat{g}^{\mu\nu} = 2\check{\Phi} - \hat{v}^I \hat{v}_I = \check{\Phi}. \quad (\text{A15})$$

In summary, the Lorentzian objects arrange themselves under submanifold null reduction according to

$$\hat{u}_a^\mu \xrightarrow{\text{null red.}} \hat{u}_a^\mu = u_a^\mu, \quad \hat{u}_w^\mu = 1, \quad \hat{u}_w^\mu = 0, \quad \hat{u}_a^\mu = 0, \quad (\text{A16})$$

$$\hat{u}_\mu^a \xrightarrow{\text{null red.}} \hat{u}_\mu^a = u_\mu^a, \quad \hat{u}_u^w = 1, \quad \hat{u}_\mu^w = \hat{v}^I n_\mu^I, \quad \hat{u}_u^a = 0, \quad (\text{A17})$$

$$\hat{n}_\mu^I \xrightarrow{\text{null red.}} \hat{n}_\mu^I = n_\mu^I, \quad \hat{n}_u^I = 0, \quad (\text{A18})$$

$$\hat{n}_I^\mu \xrightarrow{\text{null red.}} \hat{n}_I^\mu = n_I^\mu, \quad \hat{n}_I^\mu = -\hat{v}_I. \quad (\text{A19})$$

The metric on $\widehat{\Sigma}_{p+2}$ is

$$ds_{\widehat{\Sigma}_{p+2}}^2 = 2\tau_a dx^a (dw - \check{m}_b dx^b) + \check{h}_{ab} dx^a dx^b, \quad (\text{A20})$$

while the components of the inverse metric on $\widehat{\Sigma}_{p+2}$ are

$$\hat{\gamma}^{ww} = 2\check{\Phi} = 2\check{\Phi} - \hat{v}^I \hat{v}_I, \quad \hat{\gamma}^{wa} = -\hat{v}^a, \quad \hat{\gamma}^{ab} = h^{ab}. \quad (\text{A21})$$

a. Null reduction of the connection the extrinsic curvature

We now consider the null reduction of the Lorentzian connection. The nonzero components of the higher-dimensional

Christoffel symbols are

$$\hat{\Gamma}_{\mu\nu}^\rho = \bar{\Gamma}_{(\mu\nu)}^\rho = \bar{\Gamma}_{\mu\nu}^\rho + \frac{1}{2} \hat{v}^\rho \tau_{\mu\nu}, \quad (\text{A22})$$

$$\hat{\Gamma}_{\mu\nu}^u = -\bar{\mathcal{K}}_{\mu\nu} - 2\tau_{(\mu} \partial_{\nu)} \check{\Phi}, \quad (\text{A23})$$

$$\hat{\Gamma}_{u\mu}^\rho = \frac{1}{2} h^{\rho\sigma} \tau_{\mu\sigma}, \quad (\text{A24})$$

$$\hat{\Gamma}_{u\mu}^u = \frac{1}{2} a_\mu, \quad (\text{A25})$$

where

$$\bar{\mathcal{K}}_{\mu\nu} = -\frac{1}{2} \mathcal{L}_{\hat{v}} \bar{h}_{\mu\nu}, \quad a_\mu = \mathcal{L}_{\hat{v}} \tau_\mu = \hat{v}^\rho \tau_{\rho\mu}. \quad (\text{A26})$$

The NC extrinsic curvature $\bar{\mathcal{K}}_{\mu\nu}$ should not be confused with the submanifold extrinsic curvature K_{ab}^I . The pullback of the ambient TNC extrinsic curvature, $\bar{\mathcal{K}}_{ab} = u_a^\mu u_b^\nu \bar{\mathcal{K}}_{\mu\nu}$, is related to the TNC extrinsic curvature on the submanifold Σ_{p+1} ,

$$\bar{\mathcal{K}}_{ab}^\Sigma = -\frac{1}{2} \mathcal{L}_{\hat{v}}^\Sigma \bar{h}_{ab}, \quad (\text{A27})$$

where $\mathcal{L}_{\hat{v}}^\Sigma$ denotes the Lie derivative along \hat{v}^a on Σ_{p+1} , in the following way:

$$\bar{\mathcal{K}}_{ab} = \bar{\mathcal{K}}_{ab}^\Sigma - \tau_{(a} \partial_{b)} (\hat{v}^I \hat{v}^I) + \hat{v}^I K_{ab}^I. \quad (\text{A28})$$

This can be shown by starting with $\bar{\mathcal{K}}_{ab} = u_a^\mu u_b^\nu \bar{\mathcal{K}}_{\mu\nu}$ and using $\hat{v}^\rho = \hat{v}^c u_c^\rho + \hat{v}^I n_I^\rho$ in (A26). The identity

$$\mathcal{L}_{n^I} \bar{h}_{\mu\nu} = 2\nabla_{(\mu} n_{\nu)}^I + 2\tau_{(\mu} \partial_{\nu)} \hat{v}^I \quad (\text{A29})$$

together with Eq. (2.61) can then be used to derive (A28).

The higher-dimensional extrinsic curvature \hat{K}_{ab}^I is determined in terms of the higher-dimensional analog of the surface covariant derivative of (2.56), which we will call $\hat{D}_{\hat{a}}$. It acts on a mixed tensor $\hat{T}^{\hat{b}\hat{\mu}}$ according to

$$\hat{D}_{\hat{a}} \hat{T}^{\hat{b}\hat{\mu}} = \partial_{\hat{a}} \hat{T}^{\hat{b}\hat{\mu}} + \hat{\gamma}_{\hat{a}\hat{c}}^{\hat{b}} \hat{T}^{\hat{c}\hat{\mu}} + \hat{u}_{\hat{a}}^{\hat{\nu}} \hat{\Gamma}_{\hat{\nu}\hat{\lambda}}^{\hat{\mu}} \hat{T}^{\hat{b}\hat{\lambda}}, \quad (\text{A30})$$

where $\hat{\gamma}_{\hat{a}\hat{c}}^{\hat{b}}$ is the Levi-Civita connection of $\hat{\gamma}$, while $\hat{\Gamma}_{\hat{\nu}\hat{\lambda}}^{\hat{\mu}}$ is the Levi-Civita connection of \hat{g} . The higher-dimensional extrinsic curvature is

$$\hat{K}_{ab}^I = \hat{n}_\mu^I \hat{D}_a \hat{u}_b^\mu = \hat{n}_\mu^I (\partial_a \hat{u}_b^\mu + \hat{u}_a^{\hat{\nu}} \hat{\Gamma}_{\hat{\nu}\hat{\lambda}}^{\hat{\mu}} \hat{u}_b^{\hat{\lambda}}), \quad (\text{A31})$$

which using (A16) and (A18) means that

$$\hat{K}_{ab}^I = n_\mu^I D_a u_b^\mu + \frac{1}{2} \hat{v}^I \tau_{ab} = K_{ab}^I, \quad (\text{A32})$$

where we have recognized the extrinsic curvature of (2.61). This is invariant under both gauge transformations and Galilean boosts. The other nonzero components of the higher-dimensional extrinsic curvature are $\hat{K}_{wb}^I = -\frac{1}{2} \tau_{Ib}$.

Below Eq. (A2), we have shown that the $U(1)$ gauge transformation is a specific diffeomorphism in the higher-dimensional description. This is a useful way to find out how various objects transform under the σ gauge transformation. This also applies to tensors defined on the submanifold Σ_{p+1} , since they descend from the Lorentzian manifold $\widehat{\Sigma}_{p+2}$. A diffeomorphism of a generic tensor $X_a^{\hat{b}}$ is given by

$$\delta X_a^{\hat{b}} = \hat{\xi}^{\hat{c}} \partial_{\hat{c}} X_a^{\hat{b}} + X_c^{\hat{b}} \partial_a \hat{\xi}^{\hat{c}} - X_a^{\hat{c}} \partial_{\hat{c}} \hat{\xi}^{\hat{b}}. \quad (\text{A33})$$

In order to find the $U(1)$ transformation, we need to choose a diffeomorphism for which $\hat{\xi}^{\hat{a}} = -\sigma \delta_w^{\hat{a}}$. Since all objects

are independent of u we find that a 1-form X_a in this case transforms as

$$\delta X_a = -X_w \partial_a \sigma, \quad (\text{A34})$$

while a vector X^b is $U(1)$ invariant. Applying this to the extrinsic curvature \hat{K}_{ab}^I we find

$$\delta_\sigma \hat{K}_{ab}^I = -\hat{K}_{aw}^I \partial_b \sigma - \hat{K}_{wb}^I \partial_a \sigma. \quad (\text{A35})$$

Using that $\hat{K}_{wb}^I = -\frac{1}{2} \tau_{Ib}$ we recover the transformation rule (2.62).

b. Variations from null reduction

Here we obtain some of the results of Sec. III A using null reduction. We begin with the variations of the normal 1-forms. In the relativistic case, the normal 1-forms can be shown to transform as [53]

$$\delta \hat{n}_\mu^I = \frac{1}{2} \hat{n}_J^{\hat{\nu}} \hat{n}_\mu^{\hat{\rho}I} \delta \hat{g}_{\hat{\nu}\hat{\rho}} - \hat{n}_\nu^{\hat{\rho}} \hat{u}_\mu^{\hat{\alpha}} \delta \hat{u}_\alpha^{\hat{\nu}} + \frac{1}{2} \hat{n}_{\hat{\mu}J} (\hat{n}^{\hat{\nu}J} \delta \hat{n}_\nu^I - \hat{n}^{\hat{\nu}I} \delta \hat{n}_\nu^J). \quad (\text{A36})$$

Restricting to $\hat{\mu} = \mu$, the last term simply reduces to $\lambda^I J n_\mu^J$. This follows from demanding that $\hat{n}_\mu^I = 0$ is preserved under transformations, implying that $\delta \hat{n}_\mu^I = 0$. Ignoring rotations of the normal 1-forms, we get

$$\delta n_\mu^I = -\frac{1}{2} \hat{v}^\rho n_\rho^J n_{\mu J} n^{vI} \delta \tau_\nu - \frac{1}{2} n^{vJ} n_{\mu J} \hat{v}^\rho n_\rho^I \delta \tau_\nu + \frac{1}{2} n^{\rho J} n_{\mu J} n^{vI} \delta \bar{h}_{\rho\nu} - n_\nu^I u_\mu^a \delta u_a^\nu, \quad (\text{A37})$$

where we have used that $\hat{n}^\mu = -\hat{v}^\mu n_\mu^I$. Using the definitions of \hat{v} and \bar{h} , we find that the variation can be written as

$$\delta n_\mu^I = -v^{(I} n^{J)v} n_{\mu J} \delta \tau_\nu + \frac{1}{2} n^{\rho J} n_{\mu J} n^{vI} \delta h_{\rho\nu} - n_\nu^I u_\mu^a \delta u_a^\nu, \quad (\text{A38})$$

in agreement with the result (3.4) [up to a local $\mathfrak{so}(d-p)$ transformation that we ignored].

With this at hand, we rederive (3.10) using the method of null reduction. The relativistic result reads [53]

$$\delta_{\hat{X}} \hat{K}_{ab}^I = -\hat{n}_\mu^I \hat{D}_a \hat{D}_b \hat{\xi}^{\hat{\mu}} + \hat{n}_\mu^I \hat{\xi}^{\hat{\lambda}} \hat{u}_a^{\hat{\nu}} \hat{u}_b^{\hat{\rho}} \hat{R}_{\hat{\lambda}\hat{\nu}\hat{\rho}}^{\hat{\mu}} + \hat{\lambda}^I J \hat{K}_{ab}^J, \quad (\text{A39})$$

where

$$\hat{\lambda}^{IJ} = \hat{n}^{\hat{\mu}I} \hat{n}^{J\hat{\nu}} \hat{\xi}^{\hat{\rho}} \partial_{\hat{\nu}} \hat{g}_{\hat{\mu}\hat{\rho}} = \hat{n}_{\hat{\rho}}^{[I} \hat{n}^{J]\hat{\nu}} \hat{\Gamma}_{\hat{\nu}\hat{\sigma}}^{\hat{\rho}} \hat{\xi}^{\hat{\sigma}}. \quad (\text{A40})$$

We keep the null direction fixed, so that

$$\hat{\xi}^{\hat{\mu}} = -\delta \hat{X}^{\hat{\mu}}, \quad \hat{\xi}^u = 0. \quad (\text{A41})$$

We are interested in $(\hat{a}, \hat{b}) = (a, b)$ and since $\hat{n}_u^I = 0 = \hat{u}_a^u$, (A39) reduces to

$$\delta_X \hat{K}_{ab}^I = -n_\mu^I \hat{D}_a \hat{D}_b \xi^\mu + n_\mu^I \xi^\lambda u_a^\nu u_b^\sigma \hat{R}_{\lambda\nu\sigma}^\mu + \hat{\lambda}^I J \hat{K}_{ab}^J, \quad (\text{A42})$$

where $\hat{\xi}^\mu = \xi^\mu$ so that $\delta \hat{X}^\mu = \delta X^\mu$. In the absence of torsion, the null reduction of the Riemann tensor gives

$$\hat{R}_{\lambda\nu\sigma}^\mu = -\partial_\lambda \hat{\Gamma}_{\nu\sigma}^\mu + \partial_\nu \hat{\Gamma}_{\lambda\sigma}^\mu - \hat{\Gamma}_{\lambda\rho}^\mu \hat{\Gamma}_{\nu\sigma}^{\hat{\rho}} + \hat{\Gamma}_{\nu\rho}^\mu \hat{\Gamma}_{\lambda\sigma}^{\hat{\rho}} = R_{\lambda\nu\sigma}^\mu. \quad (\text{A43})$$

Since in the absence of torsion $\hat{D}_w \xi^\mu = 0$ and $\hat{D}_b \xi^\mu = D_b \xi^\mu$, we find that

$$\hat{D}_a \hat{D}_b \xi^\mu = D_a D_b \xi^\mu, \quad (\text{A44})$$

while the null reduction of (A40) gives $\hat{\lambda}^{IJ} = n_\rho^I n^{J\rho} \Gamma_{\nu\sigma}^\rho \xi^\sigma$ and so we obtain (3.10), as expected.

c. Note on the reduction of the Lorentzian action

The variational principle for NC surfaces in Sec. III B 1 can be obtained from null reduction of the relativistic variational principle [46]:

$$\delta S = \int_\Sigma d^{p+1} \sigma \sqrt{-\hat{\gamma}} \left(\frac{1}{2} \hat{T}^{\hat{a}\hat{b}} \delta \hat{\gamma}_{\hat{a}\hat{b}} + \hat{D}^{\hat{a}\hat{b}} \delta \hat{K}_{\hat{a}\hat{b}}^I \right). \quad (\text{A45})$$

The null reduction formulas of the previous section, for instance, (A42), imply that the null reduction of (A45) will include a dependence on variations of $\hat{K}_{wa}^I = -\frac{1}{2} \tau_{Ia}$. Such torsion dependent terms were not included in (3.23). The reason, as mentioned throughout the paper is that we have assumed to be working without torsion, that is, $\tau_{\mu\nu} = 0$ at the expense of only being able to extract the divergence of the energy current instead of the energy current itself.

2. Perfect fluid from null reduction

In this section, we consider the null reduction of the equilibrium partition function of a relativistic space-filling perfect fluid, that is a fluid that is not living on a surface. The case in which the fluid is confined to the surface (i.e., a fluid membrane) considered in Sec. IV A is a straightforward modification of this analysis. The result provides us with the hydrostatic partition function of a Galilean-invariant perfect fluid.

We begin with the null reduction of the unit normalized relativistic fluid velocity $\hat{u}^{\hat{\mu}}$, which obeys $\hat{g}_{\hat{\mu}\hat{\nu}} \hat{u}^{\hat{\mu}} \hat{u}^{\hat{\nu}} = -1$. We define the nonrelativistic fluid velocity u^μ as follows [36]:

$$u^\mu = \frac{\hat{u}^\mu}{\hat{u}_u}, \quad (\text{A46})$$

where $\hat{u}_u = \hat{g}_{u\hat{\mu}} \hat{u}^{\hat{\mu}} = \tau_\mu \hat{u}^\mu$. This implies that $\tau_\mu u^\mu = 1$ which is the standard normalization of the contravariant velocity of a nonrelativistic fluid. The relativistic condition

$$\hat{g}_{\hat{\mu}\hat{\nu}} \hat{u}^{\hat{\mu}} \hat{u}^{\hat{\nu}} = \bar{h}_{\mu\nu} \hat{u}^\mu \hat{u}^\nu + 2\tau_\mu \hat{u}^\mu \hat{u}^u = -1, \quad (\text{A47})$$

can be used to solve for \hat{u}^u , leading to

$$\hat{u}^u = -\frac{1}{2\hat{u}_u} - \frac{1}{2} \hat{u}_u \bar{h}_{\mu\nu} u^\mu u^\nu. \quad (\text{A48})$$

We still need to find a lower-dimensional interpretation of \hat{u}_u . This can be achieved as follows. Let $\hat{T}^{\hat{\mu}\hat{\nu}}$ be the energy-momentum tensor of the higher-dimensional relativistic theory. For a perfect fluid this is $\hat{T}^{\hat{\mu}\hat{\nu}} = (\hat{E} + \hat{P}) \hat{u}^{\hat{\mu}} \hat{u}^{\hat{\nu}} + \hat{P} \delta_{\hat{\mu}\hat{\nu}}$. The mass current of the null reduced theory is given by $\hat{T}^{\mu u}$ (see, e.g., Ref. [36]). In the lower-dimensional theory, this is equal to nu^μ , where n is the mass density. Comparing the two expressions yields

$$\hat{u}_u^2 = \frac{n}{\hat{E} + \hat{P}}. \quad (\text{A49})$$

We will later find expressions for \hat{E} and \hat{P} in terms of the nonrelativistic energy and pressure.

In the hydrostatic partition function approach for a relativistic fluid, one identifies the intensive fluid variables such as temperature and velocity with a timelike Killing vector of an otherwise arbitrary Lorentzian curved background geometry. By varying the metric while keeping the Killing vector fixed, one extracts the fluid energy-momentum tensor. This approach has been applied to nonrelativistic fluids on a NC background in Refs. [33,92] and here we will show how this follows from null reduction. In the higher-dimensional Lorentzian geometry, we assume the existence of a Killing vector $\hat{k}^{\hat{\mu}}$ such that

$$\hat{k}^{\hat{\mu}} = \hat{\beta}\hat{u}^{\hat{\mu}}, \quad (\text{A50})$$

where $\hat{\beta}$ is the relativistic (inverse) temperature, and $\hat{u}^{\hat{\mu}}$ the relativistic fluid velocity. Just like in the Lorentzian setting, we will introduce a Newton-Cartan Killing vector k^{μ} that is proportional to the nonrelativistic fluid velocity u^{μ} and that is timelike, where $\tau_{\mu}k^{\mu}$ relates to the nonrelativistic temperature. Hence we write

$$k^{\mu} = \beta u^{\mu}, \quad (\text{A51})$$

where $\beta = \tau_{\mu}k^{\mu}$ is the nonrelativistic (inverse) temperature. The null reduction of $\hat{k}^{\hat{\mu}}$ is just $\hat{k}^{\hat{\mu}} = (\hat{k}^{\mu}, k^{\mu}) = \beta(\hat{\mu}, u^{\mu})$, where we write $\hat{k}^{\hat{\mu}} = \beta\hat{\mu}$ with $\hat{\mu}$ a parameter to be determined. This means that

$$\beta u^{\mu} = \hat{\beta}\hat{u}^{\mu}. \quad (\text{A52})$$

Now, since $\hat{k}^{\hat{\mu}}$ is a Killing vector, we have that

$$\mathcal{L}_{\hat{k}}\hat{g}_{\hat{\mu}\hat{\nu}} = 0, \quad (\text{A53})$$

which, after null reduction, turns into the statements

$$\mathcal{L}_k\tau_{\mu} = 0, \quad \mathcal{L}_k\bar{h}_{\mu\nu} = -2\tau_{(\mu}\partial_{\nu)}\hat{k}^{\mu}. \quad (\text{A54})$$

In a NC geometry a Killing vector is defined by setting to zero the transformations in (2.4) [and thus also implying that the variations in (2.7) give zero]. Here $\hat{k}^{\hat{\mu}}$ is thus a specific $U(1)$ gauge transformation parameter that is associated with the existence of a Killing vector.

The relativistic hydrostatic partition function at ideal order in derivatives is an integral of the pressure which depends on the intensive variables, i.e., scalar quantities built from the Killing vector. One of these is the norm of $\hat{k}^{\hat{\mu}}$ which relates to the relativistic temperature. However, in the case of null reduction we actually have, besides $\hat{k}^{\hat{\mu}}$, another Killing vector which is $U^{\hat{\mu}} = (\frac{\partial}{\partial u})^{\hat{\mu}}$. Since $U^{\hat{\mu}}$ is null, we can form only one other scalar,

$$\hat{g}_{\hat{\mu}\hat{\nu}}U^{\hat{\mu}}\hat{k}^{\hat{\nu}} = \tau_{\mu}k^{\mu} = \beta, \quad (\text{A55})$$

which is the nonrelativistic (inverse) temperature. The other scalar is of course

$$-\hat{\beta}^2 = \hat{g}_{\hat{\mu}\hat{\nu}}\hat{k}^{\hat{\mu}}\hat{k}^{\hat{\nu}} = \beta^2(2\hat{\mu} + \bar{h}_{\mu\nu}u^{\mu}u^{\nu}). \quad (\text{A56})$$

This determines the proportionality between the relativistic and nonrelativistic temperatures. We define

$$\mu = \hat{\mu} + \frac{1}{2}\bar{h}_{\mu\nu}u^{\mu}u^{\nu}. \quad (\text{A57})$$

We will see below that μ is a chemical potential related to the mass conservation, which is a consequence of the null Killing

vector and we note that its definition implies $\mu < 0$. In the grand canonical ensemble for a system at rest, the partition function is of the form $\mathcal{Z} = \text{Tr} e^{-\beta H + \beta\mu N}$, where H is the Hamiltonian and N the conserved mass of the system.

a. Null reduction of the hydrostatic partition function

At the end of Sec. A 1 a, we discussed the role of the $U(1)$ transformation from the null reduction point of view, and we showed that such a transformation corresponds to a diffeomorphism generated by $\hat{\xi}^{\hat{\mu}} = -\sigma\delta_{\hat{u}}^{\hat{\mu}}$. Applying this to our Killing vector $\hat{k}^{\hat{\mu}}$, we learn that under $\delta_{\hat{\xi}}\hat{k}^{\hat{\mu}} = \mathcal{L}_{\hat{\xi}}\hat{k}^{\hat{\mu}}$, the NC Killing vector k^{μ} is left inert and that $\hat{k}^{\hat{\mu}}$ transforms as

$$\delta_{\sigma}\hat{k}^{\hat{\mu}} = k^{\mu}\partial_{\mu}\sigma. \quad (\text{A58})$$

Since τ_{μ} is also invariant it follows that β also does not transform. Hence, using $\hat{k}^{\hat{\mu}} = \beta\hat{\mu}$ and $k^{\mu} = \beta u^{\mu}$, we can write

$$\delta_{\sigma}\hat{\mu} = u^{\mu}\partial_{\mu}\sigma. \quad (\text{A59})$$

It then follows that μ defined in Eq. (A57) is $U(1)$ invariant, making μ together with β the two parameters on which the lower dimensional pressure in the hydrostatic partition function should depend.

In a $d+1$ -dimensional theory, the hydrostatic partition function is given by

$$S = \int d^{d+1}x eP(T, \mu), \quad (\text{A60})$$

where P is the fluid pressure. Next, we vary S keeping the Killing vector fixed, i.e., $\delta k^{\mu} = 0 = \delta\hat{k}^{\hat{\mu}}$. The variation of the temperature is then given by

$$\delta T = \delta(\tau_{\mu}k^{\mu})^{-1} = -(\tau_{\nu}k^{\nu})^{-2}k^{\mu}\delta\tau_{\mu} = -T u^{\mu}\delta\tau_{\mu}, \quad (\text{A61})$$

while the variation of the chemical potential reads

$$\begin{aligned} \delta\mu &= \delta\hat{\mu} + \frac{1}{2}u^{\mu}u^{\nu}\delta\bar{h}_{\mu\nu} + \bar{h}_{\mu\nu}u^{\nu}\delta u^{\mu} \\ &= \hat{\mu}\frac{\delta T}{T} + \frac{1}{2}u^{\mu}u^{\nu}\delta\bar{h}_{\mu\nu} \\ &\quad + \bar{u}^2\frac{\delta T}{T}. \end{aligned} \quad (\text{A62})$$

This allows us to compute

$$\delta P = \left(\frac{\partial P}{\partial T}\right)_{\mu}\delta T + \left(\frac{\partial P}{\partial\mu}\right)_{T}\delta\mu = s\delta T + n\delta\mu \quad (\text{A63})$$

$$= -\left(sT + n\mu + \frac{1}{2}n\bar{u}^2\right)u^{\mu}\delta\tau_{\mu} + \frac{1}{2}nu^{\mu}u^{\nu}\delta\bar{h}_{\mu\nu}, \quad (\text{A64})$$

where s is the entropy density and n the mass density. Thus, combining our findings, we obtain

$$\begin{aligned} \delta S &= \int d^{d+1}x e\left(\mathcal{T}^{\mu}\delta\tau_{\mu} + \frac{1}{2}\mathcal{T}^{\mu\nu}\delta\bar{h}_{\mu\nu}\right) \\ &= \int d^{d+1}x e\left[\frac{1}{2}(Ph^{\mu\nu} + nu^{\mu}u^{\nu})\delta\bar{h}_{\mu\nu} - P\hat{v}^{\mu}\delta\tau_{\mu} \right. \\ &\quad \left. - \left(sT + n\mu + \frac{1}{2}n\bar{u}^2\right)u^{\mu}\delta\tau_{\mu}\right], \end{aligned} \quad (\text{A65})$$

TABLE I. The three classes of Newton-Cartan geometries and their properties.

Geometry	Constraint on τ	Causality	Torsion
TNC	None	Acausal	Yes
TTNC	$\tau \wedge d\tau = 0$	Surfaces of absolute simultaneity	Yes
NC	$d\tau = 0$	Absolute time	No

leading us to identify the energy current and the Cauchy stress-mass tensor as

$$\begin{aligned} \mathcal{T}^\mu &= -P\hat{v}^\mu - \left(sT + n\mu + \frac{1}{2}n\bar{u}^2 \right) u^\mu \\ &= -P\hat{v}^\mu - \left(\mathcal{E} + P + \frac{1}{2}n\bar{u}^2 \right) u^\mu, \end{aligned} \quad (\text{A66})$$

$$\mathcal{T}^{\mu\nu} = Ph^{\mu\nu} + nu^\mu u^\nu, \quad (\text{A67})$$

where we defined \mathcal{E} , the internal energy, via the relation $\mathcal{E} + P = sT + n\mu$. This matches the results of Ref. [36], where these equations were obtained by directly null reducing the expression for the relativistic energy-momentum tensor.

The relation between the relativistic and nonrelativistic currents can be found from

$$\frac{1}{2}\sqrt{-\hat{g}}\hat{T}^{\hat{\mu}\hat{\nu}}\delta\hat{g}_{\hat{\mu}\hat{\nu}} = e\left(\mathcal{T}^\mu\delta\tau_\mu + \frac{1}{2}\mathcal{T}^{\mu\nu}\delta\bar{h}_{\mu\nu}\right). \quad (\text{A68})$$

Hence the energy current is given by $\mathcal{T}^\mu = \hat{T}^{\mu\mu}$. For a perfect fluid, this is $\mathcal{T}^\mu = (\hat{E} + \hat{P})\hat{u}^\mu\hat{u}^\mu - \hat{P}\hat{v}^\mu$. Comparing this with (A66) implies that we have the identification $\hat{P} = P$, as well as

$$\mathcal{E} + P + \frac{1}{2}n\bar{u}^2 = -(\hat{E} + \hat{P})\hat{u}_\mu\hat{u}^\mu = \frac{1}{2}(\hat{E} + \hat{P}) + \frac{1}{2}n\bar{u}^2, \quad (\text{A69})$$

where $\bar{u}^2 = \bar{h}_{\mu\nu}u^\mu u^\nu$ and where we used (A46), (A48), and (A49). Hence we conclude that, since $\hat{P} = P$, we have $\hat{E} = 2\mathcal{E} + P$. Finally, we note that Eq. (A49) can be obtained from comparing $\hat{T}^{\mu\nu} = \mathcal{T}^{\mu\nu}$. Replacing P in (A60) by χ and confining the fluid to a surface leads to (4.5) upon Wick rotation.

APPENDIX B: CLASSES OF NEWTON-CARTAN GEOMETRIES

As mentioned in Sec. II A 3, while it is not necessary to work with torsion for relevant systems, it is nevertheless formally necessary to introduce it in order to obtain the correct variational calculus [see discussion around (2.13)]. Thus it is instructive to briefly mention other types of Newton-Cartan geometry for which different conditions on τ_μ are considered. In the most general version of NC geometry, ‘‘torsional Newton-Cartan geometry’’ (TNC geometry [27–29]), the clock 1-form is completely unconstrained. A more moderate version, referred to as ‘‘twistless torsional Newton-Cartan geometry’’ (TTNC geometry), requires that the clock 1-form be hypersurface-orthogonal (i.e., it satisfies the Frobenius integrability condition $\tau \wedge d\tau = 0$). We summarize these different notions in Table I. In fact, these conditions are intimately linked with torsion. In particular if τ is closed ($d\tau = 0$), there is no torsion, but if τ is hypersurface-orthogonal ($\tau \wedge d\tau = 0$) the *twist* vanishes, $\omega^2 = h^{\mu\rho}h^{\nu\sigma}\omega_{\mu\nu}\omega_{\rho\sigma} = 0$, where the twist tensor is given by $\omega_{\mu\nu} = h^{\rho\sigma}h_{\sigma\mu}h^{\lambda\kappa}h_{\kappa\nu}\tau_{\rho\lambda}$. Finally, if the clock 1-form is completely unconstrained, so is the torsion.

When there is no constraint on τ_μ , it was shown in Ref. [44] that the spacetime becomes acausal in the sense that given a point P there exists a neighborhood of P such that all points in the neighborhood are separated from P by curves that are spacelike, i.e., their tangent vectors are orthogonal to τ_μ . When τ_μ is hypersurface orthogonal, the spacetime admits a foliation in terms of constant time slices. At different points on such a hypersurface clocks may tick at a slower or faster rate as time evolves, although all observers on such a constant time slices agree that they are simultaneous with each other. When there is no torsion (and τ is exact) the rate at which time evolves is the same for all points on the constant time slices and we are dealing with absolute time. In this case the interval between two events P and Q connected by a curve γ joining P and Q , i.e., $\int_\gamma \tau$, is independent of the choice of γ .

APPENDIX C: CONNECTIONS ON THE SUBMANIFOLD

The purpose of this Appendix is to find the relation between the NC connections of the ambient spacetime and the submanifold as described in Sec. II B 5.

Consider first the projection of the submanifold covariant derivative acting on a vector V^ν ,

$$\begin{aligned} u_a^\mu u_\nu^b \nabla_\mu V^\nu &= u_a^\mu u_\nu^b (\partial_\mu V^\nu + \Gamma_{\mu\rho}^\nu V^\rho) = \partial_a (u_\nu^b V^\nu) - V^\nu \partial_a u_\nu^b + u_a^\mu u_\nu^b \Gamma_{\mu\rho}^\nu V^\rho \\ &= \partial_a V^b - V^\sigma (u_c^\nu u_\sigma^c + n_I^\nu n_\sigma^I) \partial_a u_\nu^b + u_a^\mu u_\nu^b \Gamma_{\mu\rho}^\nu (u_c^\rho u_\sigma^c + n_I^\rho n_\sigma^I) V^\sigma \\ &= \partial_a V^b + \Gamma_{ac}^b V^c - V^c u_c^\nu \partial_a u_\nu^b - V_I h^{bc} \tilde{K}_{ac}^I, \end{aligned} \quad (\text{C1})$$

where we defined

$$\Gamma_{ca}^b = u_a^\mu u_\nu^b u_c^\rho \Gamma_{\rho\mu}^\nu. \quad (\text{C2})$$

Now, if the vector is a pushforward of a submanifold vector as in $V^\mu = u_a^\mu V^a$, the last term in the expression above vanishes, which leads us to define

$$\gamma_{ac}^b = \Gamma_{ac}^b - u_c^\mu \partial_a u_\mu^b. \quad (\text{C3})$$

The connection on the submanifold is also given by (2.57), which we can write using the ambient structures as

$$-\hat{v}^c \partial_a \tau_b = -u_\mu^c u_a^\nu u_b^\rho \hat{v}^\mu \partial_\nu \tau_\rho - \hat{v}^c u_a^\nu \tau_\rho \partial_\nu u_b^\rho, \quad (C4)$$

$$h^{cd} \partial_a \bar{h}_{bd} = u_\mu^c u_\nu^d u_a^\rho u_b^\lambda u_\rho^\sigma \partial_\sigma \bar{h}_{\lambda\sigma} + h^{cd} \bar{h}_{b\lambda} \partial_a u_d^\lambda + h^{c\lambda} \bar{h}_{\lambda\sigma} \partial_a u_b^\sigma. \quad (C5)$$

Substituting these back into (2.57), we find that

$$\begin{aligned} \gamma_{ab}^c - \Gamma_{ab}^c &= -\hat{v}^c \tau_\rho \partial_a u_b^\rho + \frac{1}{2} h^{cd} \bar{h}_{b\lambda} \partial_a u_d^\lambda + \frac{1}{2} h^{c\lambda} \bar{h}_{\lambda\sigma} \partial_a u_b^\sigma + \frac{1}{2} h^{cd} \bar{h}_{a\lambda} \partial_b u_d^\lambda + \frac{1}{2} h^{c\lambda} \bar{h}_{\lambda\sigma} \partial_b u_a^\sigma - \frac{1}{2} h^{cd} \bar{h}_{a\lambda} \partial_d u_b^\lambda - \frac{1}{2} h^{cd} \bar{h}_{\lambda b} \partial_d u_a^\lambda \\ &= u_\sigma^c \partial_a u_b^\sigma, \end{aligned} \quad (C6)$$

obtaining the result (2.58).

APPENDIX D: GAUSS-BONNET AND (2 + 1)-DIMENSIONAL MEMBRANES

For a closed co-dimension one surface embedded in flat (3 + 1)-dimensional Newton-Cartan geometry, the Gauss-Codazzi equation (2.73) relates K^2 and $K \cdot K$ according to

$$K^2 - K \cdot K = \mathcal{R}, \quad (D1)$$

where \mathcal{R} is the spatial Ricci scalar $\mathcal{R} = h^{ab} \mathcal{R}_{ac}{}^c$. This is the Ricci scalar of a two-dimensional spatial metric on constant time slices of Σ . This can be seen from the perspective of gauging the Bargmann algebra (see, e.g., Refs. [23,55,56]) as we will briefly review.

In this section we will denote surface tangent space indices as $\bar{a}, \bar{b}, \dots = 1, 2$. It is well known that (2 + 1)-dimensional Newton-Cartan geometry arises as a gauging of $\text{barg}(2, 1)$, which is generated by $(H, P_{\bar{a}}, G_{\bar{a}}, J_{\bar{a}\bar{b}}, N)$ with the following nonvanishing brackets:

$$\begin{aligned} [H, G_{\bar{a}}] &= P_{\bar{a}}, [J_{\bar{a}\bar{b}}, G_{\bar{c}}] = 2\delta_{\bar{c}[\bar{a}} G_{\bar{b}]}, [J_{\bar{a}\bar{b}}, P_{\bar{c}}] = 2\delta_{\bar{c}[\bar{a}} P_{\bar{b}]}, \\ [J_{\bar{a}\bar{b}}, J_{\bar{c}\bar{d}}] &= 4\delta_{[\bar{a}[\bar{d}} J_{\bar{c}]\bar{b}]}, [P_{\bar{a}}, G_{\bar{b}}] = N\delta_{\bar{a}\bar{b}}. \end{aligned} \quad (D2)$$

The gauging procedure then proceeds as follows. We introduce a Lie algebra valued connection

$$\mathcal{A}_a = H\tau_a + P_{\bar{a}} e_a^{\bar{a}} + Nm_a + G_{\bar{a}} \omega_\mu^{\bar{a}} + \frac{1}{2} J_{\bar{a}\bar{b}} \omega_a^{\bar{a}\bar{b}}, \quad (D3)$$

with an associated curvature two-form $\mathcal{F} = d\mathcal{A} + \mathcal{A} \wedge \mathcal{A}$ whose Lie algebra expansion is given by

$$\begin{aligned} \mathcal{F}_{ab} &= HR_{ab}(H) + P_{\bar{a}} \bar{\mathcal{R}}_{ab}{}^{\bar{a}}(P) + N \bar{\mathcal{R}}_{ab}(N) \\ &+ G_{\bar{a}} \bar{\mathcal{R}}_{ab}{}^{\bar{a}}(G) + \frac{1}{2} J_{\bar{a}\bar{b}} \bar{\mathcal{R}}_{ab}{}^{\bar{a}\bar{b}}(J). \end{aligned} \quad (D4)$$

In Ref. [58] it is shown that the Riemann tensor is related to the curvatures appearing in the gauging procedure as follows:

$$\mathcal{R}_{abd}{}^c = e_{\bar{a}}^c \tau_d \bar{\mathcal{R}}_{ab}{}^{\bar{a}}(G) - e_{d\bar{a}} e_{\bar{b}}^c \bar{\mathcal{R}}_{ab}{}^{\bar{a}\bar{b}}(J). \quad (D5)$$

The curvature of the spatial rotations $\bar{\mathcal{R}}_{ab}{}^{\bar{a}\bar{b}}(J)$ is the curvature 2-form of the constant time slices which for (twistless torsional) NC geometry is Riemannian. In (2 + 1)-dimensional Newton-Cartan geometry, therefore, the spatial Ricci scalar \mathcal{R} only depends on the curvature two-form $\bar{\mathcal{R}}_{ab}{}^{\bar{a}\bar{b}}(J)$ and we have the usual identities from two-dimensional Riemannian geometry for the spatial projections of $\mathcal{R}_{abd}{}^c$. For example, the vanishing of the two-dimensional Einstein tensor would read

$$h^{ac} h^{be} \mathcal{R}_{abc}{}^d - \frac{1}{2} \mathcal{R} h^{de} = 0. \quad (D6)$$

In the case of torsionless NC geometry the (2+1)-dimensional integration measure e is just the integration measure on the constant time slices (since the time direction has a trivial measure when we are dealing with absolute time). The Gauss-Bonnet theorem then tells us that

$$\int_{\Sigma} d^3 \sigma e \mathcal{R} = 4\pi \int d\sigma^0 \chi(\Sigma_s), \quad (D7)$$

where $\chi(\Sigma_s)$ is the Euler characteristic of the constant time slices Σ_s . Hence, the Gauss-Codazzi equation (D1) gives us a relation between the coefficients a_2, a_3 of (4.19), allowing us to set either a_2 or a_3 equal to zero (but only when both a_2 and a_3 are constant). In (4.24), we have chosen to set a_3 to zero.

-
- [1] P. B. Canham, The minimum energy of bending as a possible explanation of the biconcave shape of the human red blood cell, *J. Theor. Biol.* **26**, 61 (1970).
- [2] W. Helfrich, Elastic properties of lipid bilayers: Theory and possible experiments, *Z. Naturforsch. C* **28**, 693 (1973).
- [3] F. C. Keber, E. Loiseau, T. Sanchez, S. J. DeCamp, L. Giomi, M. J. Bowick, M. C. Marchetti, Z. Dogic, and A. R. Bausch, Topology and dynamics of active nematic vesicles, *Science* **345**, 1135 (2014).
- [4] C. L. Kane and E. J. Mele, Z_2 Topological Order and the Quantum Spin Hall Effect, *Phys. Rev. Lett.* **95**, 146802 (2005).
- [5] U. Seifert, Configurations of fluid membranes and vesicles, *Adv. Phys.* **46**, 13 (1997).
- [6] Z. C. Tu and Z. C. Ou-Yang, Recent theoretical advances in elasticity of membranes following Helfrich's spontaneous curvature model, *Adv. Colloid. Interface Sci.* **208**, 66 (2014).
- [7] A. Guckenberger and S. Gekle, Theory and algorithms to compute Helfrich bending forces: A review, *J. Phys.: Condens. Matter* **29**, 203001 (2017).
- [8] D. J. Steigmann, Mechanics and physics of lipid bilayers, in *The Role of Mechanics in the Study of Lipid Bilayers*, edited by D. J. Steigmann (Springer International Publishing, Cham, 2018), p. 1.

- [9] J. Guven and P. Vázquez-Montejo, The geometry of fluid membranes: Variational principles, symmetries and conservation laws, in *The Role of Mechanics in the Study of Lipid Bilayers*, edited by D. J. Steigmann (Springer International Publishing, Cham, 2018), pp. 167–219.
- [10] S. J. Streichan, M. F. Lefebvre, N. Noll, E. F. Wieschaus, and B. I. Shraiman, Quantification of myosin distribution predicts global morphogenetic flow in the fly embryo, [arXiv:1701.07100](https://arxiv.org/abs/1701.07100) (2017).
- [11] L. Ritsma *et al.*, Intestinal crypt homeostasis revealed at single-stem-cell level by in vivo live imaging, *Nature (London)* **507**, 362 (2014).
- [12] P. Delpierre, J. B. Marston, and A. Venaille, Topological origin of equatorial waves, *Science* **358**, 1075 (2017).
- [13] S. Shankar, M. J. Bowick, and M. C. Marchetti, Topological Sound and Flocking on Curved Surfaces, *Phys. Rev. X* **7**, 031039 (2017).
- [14] D. J. G. Pearce, P. W. Ellis, A. Fernandez-Nieves, and L. Giomi, Geometrical Control of Active Turbulence in Curved Topographies, *Phys. Rev. Lett.* **122**, 168002 (2019).
- [15] S. Henkes, M. C. Marchetti, and R. Sknepnek, Dynamical patterns in nematic active matter on a sphere, *Phys. Rev. E* **97**, 042605 (2018).
- [16] F. Alaïmo, C. Köhler, and A. Voigt, Curvature controlled defect dynamics in topological active nematics, *Sci. Rep.* **7**, 5211 (2017).
- [17] R. Capovilla and J. Guven, Stresses in lipid membranes, *J. Phys. A: Math. Gen.* **35**, 6233 (2002).
- [18] R. Capovilla and J. Guven, Second variation of the Helfrich–Canham Hamiltonian and reparametrization invariance, *J. Phys. A: Math. Gen.* **37**, 5983 (2004).
- [19] M. M. Terzi and M. Deserno, Lipid membranes: From self-assembly to elasticity, in *The Role of Mechanics in the Study of Lipid Bilayers*, edited by D. J. Steigmann (Springer International Publishing, Cham, 2018), pp. 105–166.
- [20] F. C. Frank, I. Liquid crystals. On the theory of liquid crystals, *Discuss. Faraday Soc.* **25**, 19 (1958).
- [21] E. Cartan, Sur les variétés à connexion affine et la théorie de la relativité généralisée (première partie), *Ann. Éc. Norm. Super.* **40**, 325 (1923).
- [22] E. Cartan, Sur les variétés à connexion affine et la théorie de la relativité généralisée (première partie) (suite), *Ann. Éc. Norm. Super.* **41**, 1 (1924).
- [23] R. Andringa, E. Bergshoeff, S. Panda, and M. de Roo, Newtonian gravity and the Bargmann algebra, *Class. Quant. Grav.* **28**, 105011 (2011).
- [24] D. Hansen, J. Hartong, and N. A. Obers, Action Principle for Newtonian Gravity, *Phys. Rev. Lett.* **122**, 061106 (2019).
- [25] K. Jensen, On the coupling of Galilean-invariant field theories to curved spacetime, *SciPost Phys.* **5**, 011 (2018).
- [26] J. Hartong, E. Kiritsis, and N. A. Obers, Schrödinger invariance from Lifshitz isometries in holography and field theory, *Phys. Rev. D* **92**, 066003 (2015).
- [27] M. H. Christensen, J. Hartong, N. A. Obers, and B. Rollier, Torsional Newton–Cartan geometry and Lifshitz holography, *Phys. Rev. D* **89**, 061901(R) (2014).
- [28] M. H. Christensen, J. Hartong, N. A. Obers, and B. Rollier, Boundary stress-energy tensor and Newton–Cartan geometry in Lifshitz holography, *J. High Energy Phys.* **01** (2014) 057.
- [29] J. Hartong, E. Kiritsis, and N. A. Obers, Lifshitz space-times for Schrödinger holography, *Phys. Lett. B* **746**, 318 (2015).
- [30] T. Harmark, J. Hartong, and N. A. Obers, Nonrelativistic strings and limits of the AdS/CFT correspondence, *Phys. Rev. D* **96**, 086019 (2017).
- [31] T. Harmark, J. Hartong, L. Menculini, N. A. Obers, and Z. Yan, Strings with non-relativistic conformal symmetry and limits of the AdS/CFT correspondence, *J. High Energy Phys.* **11** (2018) 190.
- [32] T. Harmark, J. Hartong, L. Menculini, N. A. Obers, and G. Oling, Relating non-relativistic string theories, *J. High Energy Phys.* **11** (2019) 071.
- [33] K. Jensen, Aspects of hot Galilean field theory, *J. High Energy Phys.* **04** (2015) 123.
- [34] N. Banerjee, S. Dutta, and A. Jain, Null fluids—A new viewpoint of Galilean fluids, *Phys. Rev. D* **93**, 105020 (2016).
- [35] E. Kiritsis and Y. Matsuo, Charge-hyperscaling violating Lifshitz hydrodynamics from black-holes, *J. High Energy Phys.* **12** (2015) 076.
- [36] J. Hartong, N. A. Obers, and M. Sanchioni, Lifshitz Hydrodynamics from Lifshitz black branes with linear momentum, *J. High Energy Phys.* **10** (2016) 120.
- [37] J. de Boer, J. Hartong, N. A. Obers, W. Sybesma, and S. Vandoren, Perfect fluids, *SciPost Phys.* **5**, 003 (2018).
- [38] J. de Boer, J. Hartong, N. A. Obers, W. Sybesma, and S. Vandoren, Hydrodynamic modes of homogeneous and isotropic fluids, *SciPost Phys.* **5**, 014 (2018).
- [39] I. Novak, J. Sonner, and B. Withers, Hydrodynamics without boosts, [arXiv:1911.02578](https://arxiv.org/abs/1911.02578) (2019).
- [40] J. de Boer, J. Hartong, E. Have, N. A. Obers, and W. Sybesma, Non-boost invariant fluid dynamics, [arXiv:2004.10759](https://arxiv.org/abs/2004.10759) (2020).
- [41] D. T. Son, Newton–Cartan geometry and the quantum Hall effect, [arXiv:1306.0638](https://arxiv.org/abs/1306.0638) [cond-mat.mes-hall] (2013).
- [42] M. Geracie, D. T. Son, C. Wu, and S.-F. Wu, Spacetime Symmetries of the quantum Hall effect, *Phys. Rev. D* **91**, 045030 (2015).
- [43] A. Gromov and A. G. Abanov, Thermal Hall Effect and Geometry with Torsion, *Phys. Rev. Lett.* **114**, 016802 (2015).
- [44] M. Geracie, K. Prabhu, and M. M. Roberts, Curved non-relativistic spacetimes, Newtonian gravitation and massive matter, *J. Math. Phys.* **56**, 103505 (2015).
- [45] J. Armas, J. Bhattacharya, A. Jain, and N. Kundu, On the surface of superfluids, *J. High Energy Phys.* **06** (2017) 090.
- [46] J. Armas, How fluids bend: The elastic expansion for higher-dimensional black holes, *J. High Energy Phys.* **09** (2013) 073.
- [47] J. Armas and T. Harmark, Black holes and biophysical (mem)branes, *Phys. Rev. D* **90**, 124022 (2014).
- [48] Y. Aminov, *The Geometry of Submanifolds* (CRC Press, Boca Raton, FL, 2014).
- [49] B. Carter, Perturbation dynamics for membranes and strings governed by Dirac–Goto–Nambu action in curved space, *Phys. Rev. D* **48**, 4835 (1993).
- [50] Recent developments in gravitation and mathematical physics, in *Proceedings of the 2nd Mexican School, Tlaxcala, Mexico, December 1–7, 1996*, edited by A. Garcia, C. Lammerzahl, A. Macias, T. Matos, and D. Nunez (Science Network Publ., Konstanz, Germany, 1998), p. 506.
- [51] B. Carter, Essentials of classical brane dynamics, *5th Peyresq Meeting on Quantum Spacetime, Brane Cosmology, and*

- Stochastic Effective Theories Peyresq, Haute-Provence, France, June 25–30, 2000*; *Int. J. Theor. Phys.* **40**, 2099 (2001).
- [52] R. Capovilla and J. Guven, Geometry of deformations of relativistic membranes, *Phys. Rev. D* **51**, 6736 (1995).
- [53] J. Armas and J. Tarrío, On actions for (entangling) surfaces and DCFTs, *J. High Energy Phys.* **04** (2018) 100.
- [54] J. Armas, J. Hartong, E. Have, B. F. Nielsen, and N. Obers (unpublished).
- [55] G. Festuccia, D. Hansen, J. Hartong, and N. A. Obers, Torsional Newton-Cartan geometry from the Noether procedure, *Phys. Rev. D* **94**, 105023 (2016).
- [56] E. A. Bergshoeff, J. Hartong, and J. Rosseel, Torsional Newton-Cartan geometry and the Schrödinger algebra, *Class. Quant. Grav.* **32**, 135017 (2015).
- [57] X. Bekaert and K. Morand, Connections and dynamical trajectories in generalised Newton-Cartan gravity I. An intrinsic view, *J. Math. Phys.* **57**, 022507 (2016).
- [58] J. Hartong and N. A. Obers, Hořava-Lifshitz gravity from dynamical Newton-Cartan geometry, *J. High Energy Phys.* **07** (2015) 155.
- [59] M. Geracie, K. Prabhu, and M. M. Roberts, Physical stress, mass, and energy for non-relativistic matter, *J. High Energy Phys.* **06** (2017) 089.
- [60] M. Geracie and D. T. Son, Hydrodynamics on the lowest Landau level, *J. High Energy Phys.* **06** (2015) 044.
- [61] N. Banerjee, J. Bhattacharya, S. Bhattacharyya, S. Jain, S. Minwalla, and T. Sharma, Constraints on fluid dynamics from equilibrium partition functions, *J. High Energy Phys.* **09** (2012) 046.
- [62] K. Jensen, M. Kaminski, P. Kovtun, R. Meyer, A. Ritz, and A. Yarom, Towards Hydrodynamics Without an Entropy Current, *Phys. Rev. Lett.* **109**, 101601 (2012).
- [63] J. Armas, J. Bhattacharya, and N. Kundu, Surface transport in plasma-balls, *J. High Energy Phys.* **06** (2016) 015.
- [64] F. M. Haehl, R. Loganayagam, and M. Rangamani, Adiabatic hydrodynamics: The eightfold way to dissipation, *J. High Energy Phys.* **05** (2015) 060.
- [65] L. Landau, E. Lifshitz, and J. Sykes, *Theory of Elasticity*, Course of Theoretical Physics (Pergamon Press, 1989).
- [66] H. Naito, M. Okuda, and O.-Y. Zhong-can, Polygonal shape transformation of a circular biconcave vesicle induced by osmotic pressure, *Phys. Rev. E* **54**, 2816 (1996).
- [67] F. M. Goñi, The basic structure and dynamics of cell membranes: An update of the Singer–Nicolson model, *Biochim. Biophys. Acta* **1838**, 1467 (2014).
- [68] J. Armas, J. Camps, T. Harmark, and N. A. Obers, The Young modulus of black strings and the fine structure of blackfolds, *J. High Energy Phys.* **02** (2012) 110.
- [69] B. Różycki and R. Lipowsky, Spontaneous curvature of bilayer membranes from molecular simulations: Asymmetric lipid densities and asymmetric adsorption, *J. Chem. Phys.* **142**, 054101 (2015).
- [70] O.-Y. Zhong-Can, Anchor ring-vesicle membranes, *Phys. Rev. A* **41**, 4517 (1990).
- [71] M. Mutz and D. Bensimon, Observation of toroidal vesicles, *Phys. Rev. A* **43**, 4525 (1991).
- [72] C. Tanford, Hydrostatic pressure in small phospholipid vesicles, *Proc. Natl. Acad. Sci. USA* **76**, 3318 (1979).
- [73] S. N. Solodukhin, Entanglement entropy in non-relativistic field theories, *J. High Energy Phys.* **04** (2010) 101.
- [74] E. Bergshoeff, J. Gomis, and Z. Yan, Nonrelativistic string theory and T-duality, *J. High Energy Phys.* **11** (2018) 133.
- [75] J. Klusoň, Non-relativistic D-brane from T-duality along null direction, [arXiv:1907.05662](https://arxiv.org/abs/1907.05662) (2019).
- [76] D. Pereñiguez, p -brane Newton-Cartan geometry, *J. Math. Phys.* **60**, 112501 (2019).
- [77] A. Gromov, K. Jensen, and A. G. Abanov, Boundary Effective Action for Quantum Hall States, *Phys. Rev. Lett.* **116**, 126802 (2016).
- [78] G. Napoli and L. Vergori, Hydrodynamic theory for nematic shells: The interplay among curvature, flow, and alignment, *Phys. Rev. E* **94**, 020701 (2016).
- [79] F. M. Haehl, R. Loganayagam, and M. Rangamani, The fluid manifesto: Emergent symmetries, hydrodynamics, and black holes, *J. High Energy Phys.* **01** (2016) 184.
- [80] K. Jensen, N. Pinzani-Fokeeva, and A. Yarom, Dissipative hydrodynamics in superspace, *J. High Energy Phys.* **09** (2018) 127.
- [81] H. Liu and P. Glorioso, Lectures on non-equilibrium effective field theories and fluctuating hydrodynamics, in *Proceedings, Theoretical Advanced Study Institute in Elementary Particle Physics: Physics at the Fundamental Frontier (TASI 2017)*, Boulder, CO, USA, June 5–30, 2017 PoS **TASI2017**, 008 (2018).
- [82] P. Kovtun, Lectures on hydrodynamic fluctuations in relativistic theories, *INT Summer School on Applications of String Theory Seattle, Washington, USA, July 18–29, 2011*, *J. Phys. A* **45**, 473001 (2012).
- [83] J. Armas, (Non)-dissipative hydrodynamics on embedded surfaces, *J. High Energy Phys.* **09** (2014) 047.
- [84] J. Guven, Chern-Simons theory and three-dimensional surfaces, *Class. Quant. Grav.* **24**, 1833 (2007).
- [85] J. Armas and T. Harmark, Constraints on the effective fluid theory of stationary branes, *J. High Energy Phys.* **10** (2014) 063.
- [86] J. Guven, Conformal symmetry breaking and self-similar spirals, [arXiv:1904.06876](https://arxiv.org/abs/1904.06876) [cond-mat.soft] (2019).
- [87] J. Guven and G. Manrique, Conformal mechanics of planar curves, [arXiv:1905.00488](https://arxiv.org/abs/1905.00488) (2019).
- [88] J. Guven, Conformal mechanics of space curves, [arXiv:1905.07041](https://arxiv.org/abs/1905.07041) [cond-mat.soft] (2019).
- [89] J. Armas and A. Jain, Viscoelastic hydrodynamics and holography, *J. High Energy Phys.* **01** (2020) 126.
- [90] N. Poovuttikul and W. Sybesma, First order non-Lorentzian fluids, entropy production and linear instabilities, [arXiv:1911.00010](https://arxiv.org/abs/1911.00010) (2019).
- [91] B. Julia and H. Nicolai, Null Killing vector dimensional reduction and Galilean geometrodynamics, *Nucl. Phys. B* **439**, 291 (1995).
- [92] N. Banerjee, S. Dutta, and A. Jain, Equilibrium partition function for nonrelativistic fluids, *Phys. Rev. D* **92**, 081701(R) (2015).

BIBLIOGRAPHY

- [1] S. B. Nissen, M. Perera, J. M. Gonzalez, S. M. Morgani, M. H. Jensen, K. Sneppen, J. M. Brickman, and A. Trusina, “Four simple rules that are sufficient to generate the mammalian blastocyst,” *PLoS biology* **15** (2017), no. 7, e2000737.
- [2] S. B. Nissen, S. Rønild, A. Trusina, and K. Sneppen, “Theoretical tool bridging cell polarities with development of robust morphologies,” *Elife* **7** (2018) e38407.
- [3] B. F. Nielsen, S. B. Nissen, K. Sneppen, J. Mathiesen, and A. Trusina, “Model to link cell shape and polarity with organogenesis,” *iScience* **23** (2020), no. 2, 100830.
- [4] D. J. Andrew and A. J. Ewald, “Morphogenesis of epithelial tubes: Insights into tube formation, elongation, and elaboration,” *Developmental biology* **341** (2010), no. 1, 34–55.
- [5] J. M. Sawyer, J. R. Harrell, G. Shemer, J. Sullivan-Brown, M. Roh-Johnson, and B. Goldstein, “Apical constriction: a cell shape change that can drive morphogenesis,” *Developmental biology* **341** (2010), no. 1, 5–19.
- [6] J. H. Gutzman, E. Graeden, I. Brachmann, S. Yamazoe, J. K. Chen, and H. Sive, “Basal constriction during midbrain–hindbrain boundary morphogenesis is mediated by *wnt5b* and focal adhesion kinase,” *Biology open* **7** (2018), no. 11, bio034520.
- [7] M. R. Visetsouk, E. J. Falat, R. J. Garde, J. L. Wendlick, and J. H. Gutzman, “Basal epithelial tissue folding is mediated by differential regulation of microtubules,” *Development* **145** (2018), no. 22, dev167031.
- [8] S. Chung, S. Kim, and D. J. Andrew, “Uncoupling apical constriction from tissue invagination,” *Elife* **6** (2017) e22235.
- [9] Y. E. Sanchez-Corrales, G. B. Blanchard, and K. Röper, “Radially patterned cell behaviours during tube budding from an epithelium,” *Elife* **7** (2018) e35717.
- [10] S. G. McShane, M. A. Molè, D. Savery, N. D. Greene, P. P. Tam, and A. J. Copp, “Cellular basis of neuroepithelial bending during mouse spinal neural tube closure,” *Developmental biology* **404** (2015), no. 2, 113–124.
- [11] A. J. Copp, F. A. Brook, and H. J. Roberts, “A cell-type-specific abnormality of cell proliferation in mutant (curly tail) mouse embryos developing spinal neural tube defects,” *Development* **104** (1988), no. 2, 285–295.
- [12] C. T. Baldwin, N. R. Lipsky, C. F. Hoth, T. Cohen, W. Mamuya, and A. Milunsky, “Mutations in *pax3* associated with waardenburg syndrome type i,” *Human mutation* **3** (1994), no. 3, 205–211.
- [13] M. Tassabehji, A. P. Read, V. E. Newton, M. Patton, P. Gruss, R. Harris, and T. Strachan, “Mutations in the *pax3* gene causing waardenburg syndrome type 1 and type 2,” *Nature genetics* **3** (1993), no. 1, 26–30.

- [14] T.-F. Wu, Y.-L. Yao, I.-L. Lai, C.-C. Lai, P.-L. Lin, and W.-M. Yang, "Loading of pax3 to mitotic chromosomes is mediated by arginine methylation and associated with waardenburg syndrome," *Journal of Biological Chemistry* **290** (2015), no. 33, 20556–20564.
- [15] J. B. Kirkegaard, B. F. Nielsen, A. Trusina, and K. Sneppen, "Self-assembly, buckling and density-invariant growth of three-dimensional vascular networks," *Journal of the Royal Society Interface* **16** (2019), no. 159, 20190517.
- [16] B. Strilić, T. Kučera, J. Eglinger, M. R. Hughes, K. M. McNagny, S. Tsukita, E. Dejana, N. Ferrara, and E. Lammert, "The molecular basis of vascular lumen formation in the developing mouse aorta," *Developmental cell* **17** (2009), no. 4, 505–515.
- [17] C. Berclaz, D. Szlag, D. Nguyen, J. Extermann, A. Bouwens, P. J. Marchand, J. Nilsson, A. Schmidt-Christensen, D. Holmberg, A. Grapin-Botton, *et. al.*, "Label-free fast 3d coherent imaging reveals pancreatic islet micro-vascularization and dynamic blood flow," *Biomedical optics express* **7** (2016), no. 11, 4569–4580.
- [18] Y. Inoue, M. Suzuki, T. Watanabe, N. Yasue, I. Tateo, T. Adachi, and N. Ueno, "Mechanical roles of apical constriction, cell elongation, and cell migration during neural tube formation in xenopus," *Biomechanics and Modeling in Mechanobiology* **15** (2016), no. 6, 1733–1746.
- [19] H. Y. Kim, V. D. Varner, and C. M. Nelson, "Apical constriction initiates new bud formation during monopodial branching of the embryonic chicken lung," *Development* **140** (2013), no. 15, 3146–3155.
- [20] B. F. Nielsen, G. Linga, A. Christensen, and J. Mathiesen, "Substrate curvature governs texture orientation in thin films of smectic block copolymers," *Soft Matter* **16** (2020), no. 14, 3395–3406.
- [21] A. D. Pezzutti, L. R. Gomez, and D. A. Vega, "Smectic block copolymer thin films on corrugated substrates," *Soft Matter* **11** (Mar., 2015) 2866–2873.
- [22] E. A. Matsumoto, D. A. Vega, A. D. Pezzutti, N. A. Garcia, P. M. Chaikin, and R. A. Register, "Wrinkles and splay conspire to give positive disclinations negative curvature," *Proceedings of the National Academy of Sciences* **112** (Oct., 2015) 12639–12644.
- [23] K. Yamada and S. Komura, "The dynamics of order–order phase separation," *Journal of Physics: Condensed Matter* **20** (2008), no. 15, 155107.
- [24] S. Villain-Guillot and D. Andelman, "The lamellar-disorder interface: one-dimensional modulated profiles," *The European Physical Journal B - Condensed Matter and Complex Systems* **4** (July, 1998) 95–101.
- [25] L. Zhang, L. Wang, and J. Lin, "Defect structures and ordering behaviours of diblock copolymers self-assembling on spherical substrates," *Soft Matter* **10** (June, 2014) 6713.
- [26] R. D. Kamien, D. R. Nelson, C. D. Santangelo, and V. Vitelli, "Extrinsic curvature, geometric optics, and lamellar order on curved substrates," *Physical Review E* **80** (Nov., 2009) 051703.
- [27] D. R. Nelson, "Toward a Tetravalent Chemistry of Colloids," *Nano Letters* **2** (Oct., 2002) 1125–1129.
- [28] T. Lopez-Leon, A. Fernandez-Nieves, M. Nobili, and C. Blanc, "Nematic-Smectic Transition in Spherical Shells," *Physical Review Letters* **106** (June, 2011) 247802.
- [29] R. Backofen, A. Voigt, and T. Witkowski, "Particles on curved surfaces: A dynamic approach by a phase-field-crystal model," *Physical Review E* **81** (Feb., 2010) 025701.

- [30] E. A. Matsumoto, D. A. Vega, A. D. Pezzutti, N. A. García, P. M. Chaikin, and R. A. Register, “Wrinkles and splay conspire to give positive disclinations negative curvature,” *Proceedings of the National Academy of Sciences* **112** (2015), no. 41, 12639–12644.
- [31] D. Miller, M. A. Martin, N. Harel, O. Tirosh, T. Kustin, M. Meir, N. Sorek, S. Gefen-Halevi, S. Amit, O. Vorontsov, *et. al.*, “Full genome viral sequences inform patterns of SARS-CoV-2 spread into and within Israel,” *Nature communications* **11** (2020), no. 1, 1–10.
- [32] A. Endo, S. Abbott, A. J. Kucharski, S. Funk, *et. al.*, “Estimating the overdispersion in covid-19 transmission using outbreak sizes outside china,” *Wellcome Open Research* **5** (2020), no. 67, 67.
- [33] C. Pozderac and B. Skinner, “Superspreading of sars-cov-2 in the usa,” *Plos one* **16** (2021), no. 3, e0248808.
- [34] J. B. Kirkegaard and K. Sneppen, “Variability of individual infectiousness derived from aggregate statistics of covid-19,” *medRxiv* **0** (2021).
- [35] Apple Inc., “Building an App to Notify Users of COVID-19 Exposure.” https://developer.apple.com/documentation/exposurenotification/building_an_app_to_notify_users_of_covid-19_exposure, 2020. [Online; accessed 31-May-2020].
- [36] L. Ferretti, C. Wymant, M. Kendall, L. Zhao, A. Nurtay, L. Abeler-Dorner, M. Parker, D. G. Bonsall, and C. Fraser, “Quantifying sars-cov-2 transmission suggests epidemic control with digital contact tracing,” *medRxiv* (2020).
- [37] K. Sneppen, B. F. Nielsen, R. J. Taylor, and L. Simonsen, “Overdispersion in COVID-19 increases the effectiveness of limiting nonrepetitive contacts for transmission control,” *Proceedings of the National Academy of Sciences* **118** (2021), no. 14,.
- [38] B. F. Nielsen, L. Simonsen, and K. Sneppen, “COVID-19 superspreading suggests mitigation by social network modulation,” *Physical Review Letters* **126** (2021), no. 11, 118301.
- [39] B. F. Nielsen, A. Eilersen, L. Simonsen, and K. Sneppen, “Lockdowns exert selection pressure on overdispersion of sars-cov-2 variants,” *medRxiv* (2021).
- [40] S. Ørskov, B. F. Nielsen, S. Føns, K. Sneppen, and L. Simonsen, “The COVID-19 pandemic: Key considerations for the epidemic and its control,” *APMIS* (2021).
- [41] A. Sandford, “Coronavirus: Half of humanity now on lockdown as 90 countries call for confinement,” *Euronews* (2020).
- [42] C. Fraser, D. A. Cummings, D. Klinkenberg, D. S. Burke, and N. M. Ferguson, “Influenza transmission in households during the 1918 pandemic,” *American journal of epidemiology* **174** (2011), no. 5, 505–514.
- [43] J. Brugger and C. L. Althaus, “Transmission of and susceptibility to seasonal influenza in switzerland from 2003 to 2015,” *Epidemics* **30** (2020) 100373.
- [44] M. G. Roberts and H. Nishiura, “Early estimation of the reproduction number in the presence of imported cases: pandemic influenza h1n1-2009 in new zealand,” *PloS one* **6** (2011), no. 5, e17835.
- [45] B. M. Althouse, E. A. Wenger, J. C. Miller, S. V. Scarpino, A. Allard, L. Hébert-Dufresne, and H. Hu, “Superspreading events in the transmission dynamics of sars-cov-2: Opportunities for interventions and control,” *PLoS biology* **18** (2020), no. 11, e3000897.

- [46] Q. Bi, Y. Wu, S. Mei, C. Ye, X. Zou, Z. Zhang, X. Liu, L. Wei, S. A. Truelove, T. Zhang, *et. al.*, “Epidemiology and transmission of covid-19 in 391 cases and 1286 of their close contacts in shenzhen, china: a retrospective cohort study,” *The Lancet Infectious Diseases* (2020).
- [47] Q.-L. Jing, M.-J. Liu, Z.-B. Zhang, L.-Q. Fang, J. Yuan, A.-R. Zhang, N. E. Dean, L. Luo, M.-M. Ma, I. Longini, *et. al.*, “Household secondary attack rate of covid-19 and associated determinants in guangzhou, china: a retrospective cohort study,” *The Lancet Infectious Diseases* **20** (2020), no. 10, 1141–1150.
- [48] J. O. Lloyd-Smith, S. J. Schreiber, P. E. Kopp, and W. M. Getz, “Superspreading and the effect of individual variation on disease emergence,” *Nature* **438** (2005), no. 7066, 355–359.
- [49] Q. Yang, T. K. Saldi, P. K. Gonzales, E. Lasda, C. J. Decker, K. L. Tat, M. R. Fink, C. R. Hager, J. C. Davis, C. D. Ozeroff, D. Muhrad, S. K. Clark, W. T. Fattor, N. R. Meyerson, C. L. Paige, A. R. Gilchrist, A. Barbachano-Guerrero, E. R. Worden-Sapper, S. S. Wu, G. R. Brisson, M. B. McQueen, R. D. Dowell, L. Leinwand, R. Parker, and S. L. Sawyer, “Just 2% of sars-cov-2-positive individuals carry 90% of the virus circulating in communities,” *Proceedings of the National Academy of Sciences* **118** (2021), no. 21, <https://www.pnas.org/content/118/21/e2104547118.full.pdf>.
- [50] P. Z. Chen, N. Bobrovitz, Z. Premji, M. Koopmans, D. N. Fisman, and F. X. Gu, “Heterogeneity in transmissibility and shedding sars-cov-2 via droplets and aerosols,” *Elife* **10** (2021) e65774.
- [51] J. B. Kirkegaard, J. Mathiesen, and K. Sneppen, “Superspreading of airborne pathogens in a heterogeneous world,” *Scientific reports* **11** (2021), no. 1, 1–9.
- [52] M. Kidd, A. Richter, A. Best, N. Cumley, J. Mirza, B. Percival, M. Mayhew, O. Megram, F. Ashford, T. White, *et. al.*, “S-variant SARS-CoV-2 lineage B.1.1.7 is associated with significantly higher viral loads in samples tested by ThermoFisher TaqPath RT-qPCR,” *The Journal of infectious diseases* **223** (2021), no. 10,.
- [53] T. Golubchik, K. A. Lythgoe, M. D. Hall, L. Ferretti, H. R. Fryer, G. MacInyre-Cockett, M. de Cesare, A. Trebes, P. Piazza, D. Buck, *et. al.*, “Early analysis of a potential link between viral load and the N501Y mutation in the SARS-COV-2 spike protein,” *medRxiv* **0** (2021).
- [54] A. Kucharski and C. L. Althaus, “The role of superspreading in middle east respiratory syndrome coronavirus (mers-cov) transmission,” *Eurosurveillance* **20** (2015), no. 25, 21167.
- [55] A. Remuzzi and G. Remuzzi, “Covid-19 and italy: what next?,” *The Lancet* (2020).
- [56] Our World In Data and European Centre for Disease Prevention and Control, “covid-19-data (Deaths).” <https://github.com/owid/covid-19-data/>, 2020.
- [57] A. Billah, M. Miah, and N. Khan, “Reproductive number of coronavirus: A systematic review and meta-analysis based on global level evidence,” *PLOS ONE* **15** (11, 2020) 1–17.
- [58] DST, “Population in Denmark (May 1, 2020),” *Publication of Statistics Denmark* (2020). <https://www.dst.dk/en/Statistik/emner/befolkning-og-valg/befolkning-og-befolkningsfremskrivning/folketal>. Accessed Nov. 19, 2020.
- [59] J. Mossong, N. Hens, M. Jit, P. Beutels, K. Auranen, R. Mikolajczyk, M. Massari, S. Salmaso, G. S. Tomba, J. Wallinga, *et. al.*, “Social contacts and mixing patterns relevant to the spread of infectious diseases,” *PLoS medicine* **5** (2008), no. 3,.

- [60] Eurostat, “Eu statistics on income and living conditions microdata 2004-2019, release 2 in 2020,” 2020.
- [61] M. S. Graham, C. H. Sudre, A. May, M. Antonelli, B. Murray, T. Varsavsky, K. Kläser, L. S. Canas, E. Molteni, M. Modat, *et. al.*, “Changes in symptomatology, reinfection, and transmissibility associated with the sars-cov-2 variant b. 1.1. 7: an ecological study,” *The Lancet Public Health* **6** (2021), no. 5, e335–e345.
- [62] E. Volz, S. Mishra, M. Chand, J. C. Barrett, R. Johnson, L. Geidelberg, W. R. Hinsley, D. J. Laydon, G. Dabrera, Á. O’Toole, *et. al.*, “Assessing transmissibility of sars-cov-2 lineage b. 1.1. 7 in england,” *Nature* **593** (2021), no. 7858, 266–269.
- [63] N. G. Davies, S. Abbott, R. C. Barnard, C. I. Jarvis, A. J. Kucharski, J. D. Munday, C. A. Pearson, T. W. Russell, D. C. Tully, A. D. Washburne, *et. al.*, “Estimated transmissibility and impact of sars-cov-2 lineage b. 1.1. 7 in england,” *Science* **372** (2021), no. 6538,.
- [64] M. Worobey, J. Pekar, B. B. Larsen, M. I. Nelson, V. Hill, J. B. Joy, A. Rambaut, M. A. Suchard, J. O. Wertheim, and P. Lemey, “The emergence of sars-cov-2 in europe and north america,” *Science* **370** (2020), no. 6516, 564–570.
- [65] B. F. Nielsen, K. Sneppen, L. Simonsen, and J. Mathiesen, “Social network heterogeneity is essential for contact tracing,” *medRxiv* (2020).
- [66] A. Mollgaard, I. Zettler, J. Dammeyer, M. H. Jensen, S. Lehmann, and J. Mathiesen, “Measure of node similarity in multilayer networks,” *PloS one* **11** (2016), no. 6,.
- [67] P. Sapiezynski, A. Stopczynski, D. D. Lassen, and S. Lehmann, “Interaction data from the copenhagen networks study,” *Scientific Data* **6** (2019), no. 1, 1–10.
- [68] A. Endo *et. al.*, “Implication of backward contact tracing in the presence of overdispersed transmission in covid-19 outbreaks,” *Wellcome open research* **5** (2020).
- [69] A. V. Tkachenko, S. Maslov, A. Elbanna, G. N. Wong, Z. J. Weiner, and N. Goldenfeld, “Time-dependent heterogeneity leads to transient suppression of the covid-19 epidemic, not herd immunity,” *Proceedings of the National Academy of Sciences* **118** (2021), no. 17,.
- [70] T. Britton, F. Ball, and P. Trapman, “A mathematical model reveals the influence of population heterogeneity on herd immunity to sars-cov-2,” *Science* **369** (2020), no. 6505, 846–849.
- [71] J. Armas, J. Hartong, E. Have, B. F. Nielsen, and N. A. Obers, “Newton-cartan submanifolds and fluid membranes,” *Physical Review E* **101** (2020), no. 6, 062803.