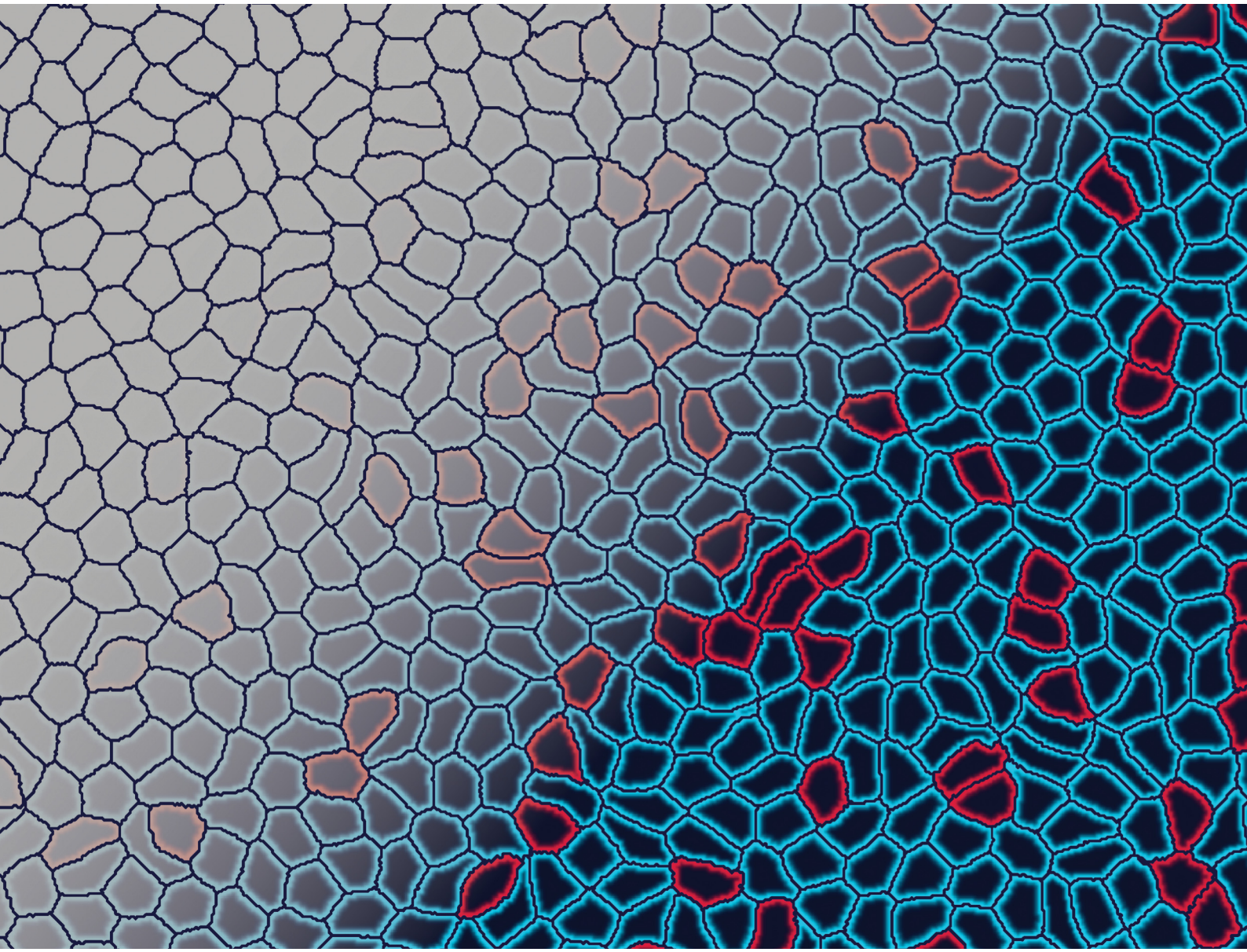
Volume 21
Number 33
7 September 2025
Pages 6473–6650

Soft Matter

rsc.li/soft-matter-journal



ISSN 1744-6848

ROYAL SOCIETY OF CHEMISTRY

PAPER
Liesbeth M. C. Janssen *et al.*
Shape matters: inferring the motility of confluent cells from static images

20 YEARS ANNIVERSARY

Soft Matter

PAPER

View Article Online
View Journal | View Issue

Cite this: *Soft Matter*, 2025, **21**, 6504

Shape matters: inferring the motility of confluent cells from static images†

Quirine J. S. Braat, ‡[a] Giulia Janzen, ‡[ab] Bas C. Jansen,‡[a] Vincent E. Debets, [a] Simone Ciarella [cd] and Liesbeth M. C. Janssen *[ae]

Cell motility in dense cell collectives is pivotal in various diseases like cancer metastasis and asthma. A central aspect in these phenomena is the heterogeneity in cell motility, but identifying the motility of individual cells is challenging. Previous work has established the importance of the average cell shape in predicting cell dynamics. Here, we aim to identify the importance of individual cell shape features, rather than collective features, to distinguish between high-motility and low-motility (or zero-motility) cells in heterogeneous cell layers. Employing the cellular Potts model, we generate simulation snapshots and extract static features as inputs for a simple machine-learning model. Our results show that when cells are either motile or non-motile, this machine-learning model can accurately predict a cell's phenotype using only single-cell shape features. Furthermore, we explore scenarios where both cell types exhibit some degree of motility, characterized by high or low motility. In such cases, our findings indicate that a neural network trained on shape features can accurately classify cell motility, particularly when the number of highly motile cells is low, and high-motility cells are significantly more motile compared to low-motility cells. This work offers potential for physics-inspired predictions of single-cell properties with implications for inferring cell dynamics from static histological images.

Received 3rd March 2025,
Accepted 17th June 2025

DOI: 10.1039/d5sm00222b

rsc.li/soft-matter-journal

1 Introduction

Collective cell migration in dense cell layers and tissues is a fundamental process underlying many physiological phenomena including wound healing, embryogenesis, and tissue development, but it also plays a critical role in disease progression such as asthma and cancer.[1,2] In general, cell migration is driven by a dynamic interplay between forces, deformations, and environmental cues, making the process complex from both a biological and physical perspective.[3,4] This inherent complexity hampers our ability to reliably predict the migratory capacity of confluent cells and densely packed cellular aggregates, both in healthy and pathological conditions. For prognostic purposes, especially in the context of cancer metastasis,[5–8] it would be highly desirable if one could infer information on the expected dynamical behaviour of cellular collectives based solely on static information, *i.e.*, from static, microscopic images routinely obtained from histopathology slides.

Recent breakthroughs have already revealed important morphodynamic links that correlate static, structural features with the collective dynamics of multicellular aggregates. Indeed, pioneering work has established that the average cell shape (as quantified by a dimensionless shape index) in confluent cell layers can serve as a remarkably good proxy for collective cell dynamics, including jamming and unjamming behaviour.[9–16] Additional static features such as the shape and size of cell nuclei can further refine the predictive power.[5,6,17] However, these studies have focused mainly on morphodynamic links for the emergent collective cell dynamics. The question to what extent static or structural information can also inform on single-cell dynamical properties, such as individual cell motility, has thus far remained largely unexplored. Gaining knowledge about such single-cell properties is particularly important in heterogeneous cell layers, where the presence of more intrinsically motile cells, as in the context of a partial epithelial-to-mesenchymal transition (EMT), is associated with more aggressive cancer progression.[18–22]

Here, we seek to derive information about individual cell motility from purely static cell data. In particular, we aim to

[a] *Department of Applied Physics, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands. E-mail: l.m.c.janssen@tue.nl*
[b] *Department of Theoretical Physics, Complutense University of Madrid, 28040 Madrid, Spain*
[c] *Netherlands eScience Center, Amsterdam 1098 XG, The Netherlands*
[d] *Laboratoire de Physique de l'Ecole Normale Supérieure, ENS, Université PSL, CNRS, Sorbonne Université, Université de Paris, F-75005 Paris, France*
[e] *Institute for Complex Molecular Systems, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands*

† Electronic supplementary information (ESI) available: Details on additional shape and structure characterization, additional figures about accuracy of the machine-learning algorithm and results for SHAP and PCA analysis on the data. See DOI: https://doi.org/10.1039/d5sm00222b

‡ These authors contributed equally to this work.

discriminate between two different cellular phenotypes, high-motility and low-motility cells, based on static images of a minimally heterogeneous *in silico* confluent cell layer. The static information that is extracted includes both single-cell geometric shape features and structural properties of the neighbouring cells surrounding a given cell. Our work draws inspiration from Janzen *et al.*,[23] who recently investigated the possibility of predicting particle motility in a dense, heterogeneous mixture of spherical active and passive colloidal particles. Briefly, they showed that the shapes of the Voronoi polygons surrounding active particles exhibit distinct characteristics which can serve as sufficient static information to accurately classify different particle motilities. In the present work, we expand upon this approach to study the more challenging, and more biologically realistic, case of a heterogeneous confluent cell layer. Our primary goal is to infer the phenotype of individual cells based on their static properties. This approach allows us to make predictions about single-cell behaviour without relying on collective cell data.

Our confluent cell model is based on the cellular Potts model (CPM), a simulation technique that allows for cell-resolved dynamics with controllable single-cell motilities.[13,24–29] The CPM, despite its simplicity, has been used successfully in the past to capture the behaviour of biological systems.[27,30,31] To distinguish between high-motility and low-motility cells, we employ a machine-learning (ML) approach that takes as input instantaneous static information derived from CPM simulation snapshots. Our choice to invoke machine learning stems from the fact that, in recent years, ML has emerged as a powerful tool for identifying structure-dynamics relations in dense disordered passive systems,[32–51] purely active systems,[51–59] and active–passive colloidal mixtures.[23] Moreover, it has been successfully employed in experimental studies to predict information about the properties of cell collectives.[60–63] We therefore envision that this work could not only advance our understanding of distinguishing motile and non-motile cells in simulations, but also find future applications in studying the behaviour of individual cells in biological confluent cell layers.

A schematic overview of our methodology is shown in Fig. 1. Briefly, we extract different static features for a given cell from an instantaneous CPM configuration, from which a simple ML algorithm subsequently seeks to classify the cell's motility phenotype. The static input features are subdivided into four categories, namely single-cell (local) shape features, neighbouring-cell (non-local) shape features, local structural features and non-local structural features. The shape features refer to the geometric properties of the cells, such as their size, aspect ratio, and perimeter. Structural features, on the other hand, encompass the spatial arrangement and include metrics such as the cell's position relative to the neighbouring cells. The distinction between shape and structural features allows us to identify how much information regarding a cell's intrinsic motility is captured by its shape. By comparing the predictive power of shape features with structural features, we can determine if the intrinsic motility of a cell can be accurately classified solely on the basis of its shape or if the structural context provides essential additional information. Additionally, focusing on shape features helps minimize the number of parameters to be extracted from images, as pinpointing structural features that require an accurate centre of mass position can be more challenging and computationally intensive. To test the validity range of our ML model, we vary the number of motile cells and their motility strength, thus allowing us to control the cell properties in the heterogeneous confluent layer.

Our analysis reveals that local (single-cell) shape features alone are sufficient to predict whether a cell is highly motile or non-motile for the computational model at hand. The local shape features work particularly well in the regime where the number of motile cells is small and the difference in cell motility between the two cell types is large. In this regime, the cells have a clearly distinct phenotype and local distortions due to a small number of motile cells can be more easily detected. These results illustrate that the shape of a single cell contains a significant amount of information about the motility of an individual cell. We also investigate how the ML algorithm performs with different cell parameters and show that the

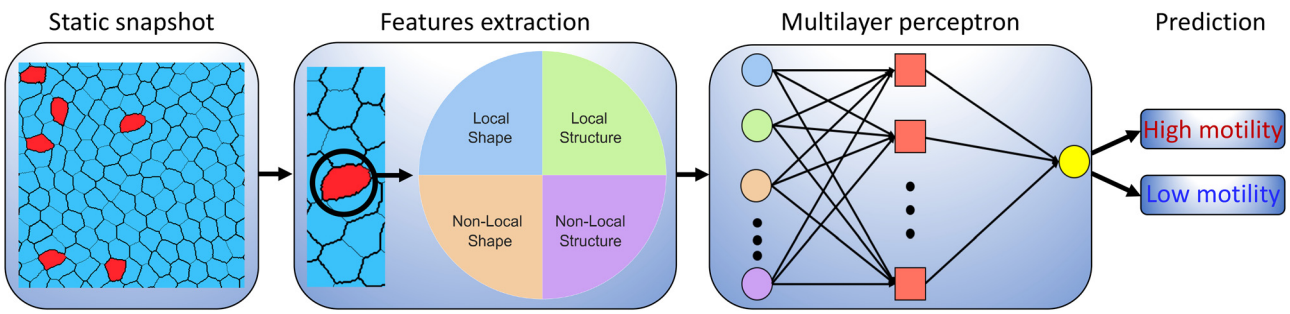


Fig. 1 Schematic overview of our machine learning approach for identifying active cells within a mixture of active and passive cells. The cellular Potts model generates static cell snapshots, and from each snapshot, a set of shape and structure features is extracted. These features include both local and non-local characteristics. Local features are determined by information on individual cells, while non-local features depend on the cell's neighbours, encompassing neighbour averages, neighbour maximum, and neighbour minimum values. Table 1 shows the complete list of features and the corresponding formulas used to compute them. Local shape features are highlighted in blue, local structure features in green, non-local shape features in orange, and non-local structure features in violet. Following feature extraction, a multilayer perceptron is trained to classify cell types, distinguishing between low motility and high motility cells solely based on the features extracted from a snapshot.

model trained only on local shape features is also successful in generalising to data with a different number of motile cells.

2 Methods

2.1 Simulation model

The ML prediction of a cell's phenotype is derived from static images produced using the cellular Potts model.[24,25] The CPM is a coarse-grained, lattice-based computational model that simulates cell dynamics *via* a Monte Carlo algorithm.[64] Briefly, cells are represented as pixelated domains on a square lattice, and their dynamics are driven by pixel-copy attempts that minimise the Hamiltonian. We note that recent work has also extended the CPM to disordered lattices.[65] For our study, we utilise the open-source CPM implementation in CompuCell3D.[66]

We simulate a two-dimensional confluent layer composed of cells with either a high or low motility. The reference Hamiltonian without motility is defined as follows[24,25]

$$\begin{aligned} H_0 &= H_{\text{adhesion}} + H_{\text{area}} + H_{\text{perimeter}} \\ &= \sum_{i,j} J_{\alpha_i,\alpha_j}\left(1-\delta(\sigma_i,\sigma_j)\right) + \sum_{\sigma}\lambda_{\text{A}}(A_\sigma - A_{\text{t}})^2 \\ &\quad + \sum_{\sigma}\lambda_{\text{P}}(P_\sigma - P_{\text{t}})^2. \end{aligned} \tag{1}$$

The individual pixels are indicated with i, j. All cells can be identified with their cell number σ and have an associated cell type α; the cell type is either active or passive, indicating high and zero (or low) motility of the cells, respectively. Each of the terms in the Hamiltonian corresponds to a different physical aspect of the cells. The first term, H_{adhesion}, accounts for the change in adhesion energy associated with cell–cell adhesion contacts. The magnitude of the cell–cell adhesion term between the cell type is set by J_{α_i,α_j}. The Kronecker delta function ($\delta(\sigma_i,\sigma_j)$) ensures that cells do not experience adhesion interactions with themselves. The second term, H_{area}, penalises large differences between a cell's actual area A_σ and its preferred area A_{t} and maintains a cell's size. Similar to the area constraint, an energy penalty term is included for large variations of a cell's perimeter, $H_{\text{perimeter}}$. Contrary to the area constraint, this penalty is only accounted for if the cell's perimeter P_σ exceeds a threshold value P_{t}. When a cell's perimeter is below the threshold P_{t}, no perimeter constraint is applied. We include the perimeter constraint to avoid cell shapes with non-physically large perimeters, which we observed primarily when the motility of the cells was large. We only include the perimeter constraint for these non-physical cell shapes such that the term does not affect the emerging cell shapes otherwise.

To implement the motility of the cells, we include an energy bias[26,67] in the Monte Carlo algorithm using

$$\Delta H = \Delta H_0 - \sum_{\sigma}\kappa_\alpha \vec{p}_\sigma \cdot \Delta\vec{R}. \tag{2}$$

Here, $\Delta\vec{R}$ is the centre-of-mass displacement due to the proposed pixel-copy attempt. The strength of the cell motility is given by κ_α and depends on the specific cell type α (either active or passive) for each individual cell. The unit vector $\vec{p}_\sigma$ represents the directional persistence of the cell. When the centre-of-mass displacement is in the direction of the unit vector $\vec{p}_\sigma$, the cell is biased to migrate in that direction. The dynamics of $\vec{p}_\sigma$ is governed by rotational diffusion, *i.e.*, the direction gets updated every Monte Carlo step (MCS) with a random angular perturbation η. We set η in the range from $-\pi/36$ to $\pi/36$, which is sufficiently long to allow the cells to escape their local environment and explore space. Overall, for a single motile cell, this implementation effectively amounts to a persistent random walk akin to *e.g.* active Brownian particles.[68]

Phenotypic heterogeneity is included *via* the motility term. In biology, motility is controlled by many intrinsic and external factors,[3,4,69] but here we reduce this complexity to a single parameter. The key difference between the high-motility and low-motility cells is the strength of the active force κ_α (which depends on the cell type α). We distinguish between two different scenarios in the simulations, namely

(1) zero-motility cells ($\kappa_{\text{p}} = 0$; passive) combined with high-motility cells ($\kappa_{\text{a}} = 1500$; active);

(2) both cell types are motile, but the high-motility cells are more motile than the low-motility ones ($\kappa_{\text{a}} > \kappa_{\text{p}} > 0$).

The first situation allows us to investigate how active cells distort the cellular arrangements in a purely passive cellular environment. The second resembles a more realistic representation of confluent cell layers, as the motility of cells shows heterogeneity even within confluent tissue.[70] The actual heterogeneity in cell motility can vary significantly between different biological systems and experimental conditions. In this study, we aim to provide a proof of principle by varying the number of highly motile cells (N_{a}) and the ratio γ, which represents the ratio between low-motility (κ_{p}) and high-motility cells (κ_{a}). This approach allows us to explore the effects of motility heterogeneity in a controlled manner.

For the numerical implementation, we employ a two-dimensional square lattice of 300 by 300 pixels with periodic boundary conditions. The simulation contains 144 cells where a number N_{a} of these cells are randomly chosen to be active, creating a mixture of active and passive cells. We vary the number of active cells between 1 and 60. We set the adhesion strength $J_{\alpha_i,\alpha_j} = 5.0$ for all cells, and each cell has a target area A_{t} of 625 pixels which is enforced with an energy penalty constraint of $\lambda_{\text{A}} = 1.0$. To avoid any cell fragmentation, the pixels of an individual cells are forced to remain connected throughout the entire simulation. This can cause artefacts in the cell shapes (long tails are formed). To circumvent this problem, the perimeter constraint (with $\lambda_{\text{P}} = 1.0$) is applied when the cell perimeter exceeds a value of $P_{\text{t}} = 150$ pixels. The complete set of simulation parameters is provided in Table S1 in the ESI.†[71] The same simulation set-up is used for the heterogeneous mixture of high-motility and low-motility cells.

After equilibration, the static snapshots are stored every 1000 mcs. This time interval is chosen such that the high-motility cells can move sufficiently between consecutive snapshots. Snapshots for different parameters are shown in Fig. 2.

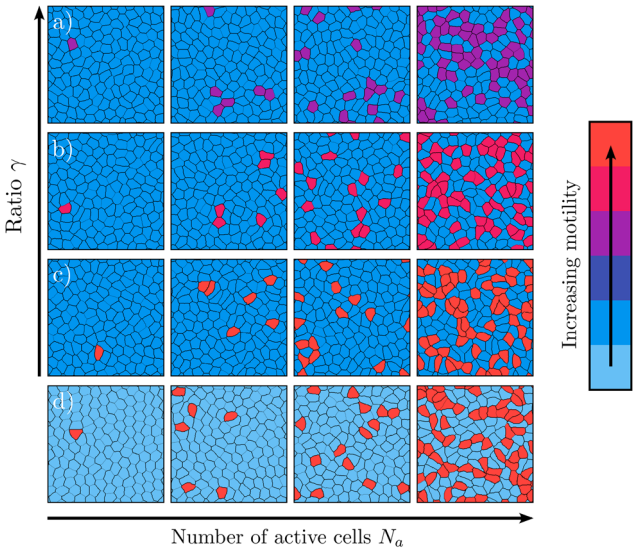


Fig. 2 Overview of the static images extracted from the cellular Potts simulations. The colours indicate the motility of the different cell types. (a)–(c) The snapshots in which the number of active cells N_a and the motility of active cells κ_a is varied for constant passive cell motility, κ_p = 150. The ratio γ is defined as κ_p/κ_a, (d) the snapshots in which the number of active cells N_a is varied for constant cell motility, κ_p = 0 and κ_a = 1500.

It is challenging to distinguish between the two cell types in the static images by visual inspection only. We therefore extract physical features from these snapshots to determine whether a machine-learning algorithm can predict the phenotype of the cell based on a set of simple physical properties.

2.2 Classification model

We approach the task of identifying highly motile cells as a binary classification problem. To accomplish this, we employ a multilayer perceptron,[72,73] as implemented in Scikit-learn,[74] which consists of interconnected neurons in multiple layers. The first layer, *i.e.* the input layer, receives the input vector, while the output layer provides output signals or classifications with assigned weights. The hidden layers adjust the weights until the neural network's margin of error is minimised.[75] In this work, we use a simple neural network with a single hidden layer containing a number of nodes equal to the input features. We update the weights using the ADAM algorithm.[76]

To evaluate the model's performance, we calculate the accuracy, defined as the number of correct predictions divided by the total predictions. Correct predictions include both accurately identifying high-motility cells as motile and low-motility cells as non-motile. Here, a prediction is a single classification attempt (motile or non-motile) based on the static input features for one given cell from one simulation snapshot (detailed in the section below). When the number of motile cells N_a deviates from the number of non-motile cells, indicating an imbalanced dataset, we address this by randomly selecting a subset of non-motile cells and excluding them. This method ensures a balanced dataset with the same number of motile and non-motile cells. We use multiple independent snapshots to obtain a total of 120 000 cells. Note that the number of snapshots used depends on N_a, but the overall number of cells remains fixed. We randomly divide the dataset into training and test sets, allocating 80% of the data to the training set and 20% to the test set. We train 20 independent neural networks, and the reported accuracy is the average accuracy obtained from these neural networks. Consequently, while the single-cell features used for training are extracted from multiple snapshots, the trained model can be tested on features extracted from a single cell. This means that although multiple snapshots are used during the training phase to improve the model's robustness, the properties of an individual cell are sufficient for making predictions during the testing phase.

While this paper presents results based on the application of a multilayer perceptron, we have confirmed that similar results can be achieved using a more sophisticated ML algorithm, specifically a gradient-boosting model, which is a machine-learning method based on decision trees.[77] Additionally, our results show that a simple logistic regression model[78–80] exhibits markedly lower accuracy in predicting cell motility than either the multilayer perceptron or the gradient-boosting algorithm (see Table S2 in the ESI†[71]). Consequently, we can conclude that, for this classification problem, a more advanced non-linear model such as a multilayer perceptron is necessary.

2.3 Input features

Rather than using simulation snapshots as input features, we extract single-cell features from each snapshot to use as input for our simple machine-learning model. This approach is preferred because it provides interpretable results. We employ a total of 145 possible features as input for our machine-learning model, categorising them into structure and shape features. Shape features are defined by the geometric properties of the cells. In contrast, structural features pertain to the spatial organization, incorporating metrics such as the cell's position relative to neighbouring cells. The comprehensive list of these features and the formulas used for their computation are shown in Table 1. A visual representation of the variables used in the computation is provided in Fig. S1 in the ESI.†[71]

Structural features are derived from the centres of mass (COM) of cells and are based solely on properties akin to local structural metrics commonly used for dense, disordered particle systems. These features encompass bond order parameters ψ_n with $n = 2, \ldots, 12$,[81] along with the first and second moment of the neighbour distance and its standard deviation. The single-cell shape features, instead, are computed based on the pixels that constitute each cell. These geometric features include cell size, border length, semi-minor and semi-major axes, parallel and perpendicular alignment, number of neighbouring cells (calculated for each cell to determine how many other cells are adjacent to it), and eccentricity. The eccentricity is determined by fitting cells with an ellipse using a least squares approach.[82]

For both shape and structural features, we further divide these types into two categories: local features, derived from

Table 1 Features employed for the machine learning model with their corresponding formulas. The colour-coded distinctions represent the four feature subsets: local shape features in blue, local structure features in green, non-local shape features in orange, and non-local structure features in violet. The 'boundary pixels' of a cell are defined as pixels with at least one first-order neighbouring pixel belonging to a different cell. The definitions of the variable names are discussed and shown visually in the ESI[71]

Parameter	Formula	Neighbouring cells		
		Avg.	Max.	Min.
Volume (#pixels)	V	$\langle V \rangle$	$V_{\max}$	$V_{\min}$
Surface (# boundary pixels)	A	$\langle A \rangle$	$A_{\max}$	$A_{\min}$
Surface-to-volume ratio	A/V	$\langle A/V \rangle$	$(A/V)_{\max}$	$(A/V)_{\min}$
Number of neighbours	N			
1st moment of mass	$\overline{\lvert r \rvert} = \frac{1}{V}\sum_i^V \lvert \vec{r}_i \rvert$	$\langle \overline{\lvert r \rvert} \rangle$	$\overline{\lvert r \rvert}_{\max}$	$\overline{\lvert r \rvert}_{\min}$
2nd moment of mass	$\overline{\lvert r \rvert^2} = \frac{1}{V}\sum_i^V \lvert \vec{r}_i \rvert^2$	$\langle \overline{\lvert r \rvert^2} \rangle$	$\overline{\lvert r \rvert^2}_{\max}$	$\overline{\lvert r \rvert^2}_{\min}$
3rd moment of mass	$\overline{\lvert r \rvert^3} = \frac{1}{V}\sum_i^V \lvert \vec{r}_i \rvert^3$	$\langle \overline{\lvert r \rvert^3} \rangle$	$\overline{\lvert r \rvert^3}_{\max}$	$\overline{\lvert r \rvert^3}_{\min}$
Stand. var. of mass	$\sigma = \sqrt{\overline{\lvert r \rvert^2} - \overline{\lvert r \rvert}^2}$	$\langle \sigma \rangle$	$\sigma_{\max}$	$\sigma_{\min}$
Skewness of mass	$\mu = \left(\overline{\lvert r \rvert^3} - 3\overline{\lvert r \rvert}\sigma^2 - \overline{\lvert r \rvert}^3\right)/\sigma^3$	$\langle \mu \rangle$	$\mu_{\max}$	$\mu_{\min}$
Total border length	$B = \sum_j \beta_j$	$\langle B \rangle$	$B_{\max}$	$B_{\min}$
Longest border length	$\beta_l = \max_j(\beta_j)$			
Shortest border length	$\beta_s = \min_j(\beta_j)$			
1st moment of border length	$\bar{\beta} = \frac{1}{N}\sum_j \beta_j$	$\langle \bar{\beta} \rangle$	$\bar{\beta}_{\max}$	$\bar{\beta}_{\min}$
2nd moment of border length	$\overline{\beta^2} = \frac{1}{N}\sum_j \beta_j^2$	$\langle \overline{\beta^2} \rangle$	$\overline{\beta^2}_{\max}$	$\overline{\beta^2}_{\min}$
Stand. var. of border length	$\sigma_\beta = \sqrt{\overline{\beta^2} - \bar{\beta}^2}$	$\langle \sigma_\beta \rangle$	$\sigma_{\beta,\max}$	$\sigma_{\beta,\min}$
Semi-major axis	a			
Semi-minor axis	b			
Eccentricity	$e = \sqrt{1 - b^2/a^2}$	$\langle e \rangle$	$e_{\max}$	$e_{\min}$
Summed squared residual (between fitted ellipse and boundary pixels)	$R = \sum_i \left(ax_i^2 + bx_iy_i + cy_i^2 + dx_i + ey_i + f\right)^2$	$\langle R \rangle$	$R_{\max}$	$R_{\min}$
Parallel alignment	$\Gamma_\parallel = \frac{1}{N}\sum_j \left(1 - \frac{bb_j}{aa_j}\right)\lvert\cos(\theta - \theta_j)\rvert$	$\langle \Gamma_\parallel \rangle$	$\Gamma_{\parallel,\max}$	$\Gamma_{\parallel,\min}$
Perpendicular alignment	$\Gamma_\perp = \frac{1}{N}\sum_j \left(1 - \frac{bb_j}{aa_j}\right)\lvert\sin(\theta - \theta_j)\rvert$	$\langle \Gamma_\perp \rangle$	$\Gamma_{\perp,\max}$	$\Gamma_{\perp,\min}$
Parallel front-end alignment	$\Gamma_{\parallel,\mathrm{FE}} = \frac{1}{N}\sum_j \left(1 - \frac{bb_j}{aa_j}\right)\lvert\cos(\theta - \theta_j)\cos(\theta - \phi_j)\rvert$	$\langle \Gamma_{\parallel,\mathrm{FE}} \rangle$	$\Gamma_{\parallel,\mathrm{FE},\max}$	$\Gamma_{\parallel,\mathrm{FE},\min}$
Perpendicular front-end alignment	$\Gamma_{\perp,\mathrm{FE}} = \frac{1}{N}\sum_j \left(1 - \frac{b_j}{a_j}\right)\lvert\sin(\theta - \theta_j)\cos(\theta - \phi_j)\rvert$	$\langle \Gamma_{\perp,\mathrm{FE}} \rangle$	$\Gamma_{\perp,\mathrm{FE},\max}$	$\Gamma_{\perp,\mathrm{FE},\min}$
Parallel side alignment	$\Gamma_{\parallel,\mathrm{side}} = \frac{1}{N}\sum_j \left(1 - \frac{bb_j}{aa_j}\right)\lvert\cos(\theta - \theta_j)\sin(\theta - \phi_j)\rvert$	$\langle \Gamma_{\parallel,\mathrm{side}} \rangle$	$\Gamma_{\parallel,\mathrm{side},\max}$	$\Gamma_{\parallel,\mathrm{side},\min}$
Perpendicular side alignment	$\Gamma_{\perp,\mathrm{side}} = \frac{1}{N}\sum_j \left(1 - \frac{bb_j}{aa_j}\right)\lvert\sin(\theta - \theta_j)\sin(\theta - \phi_j)\rvert$	$\langle \Gamma_{\perp,\mathrm{side}} \rangle$	$\Gamma_{\perp,\mathrm{side},\max}$	$\Gamma_{\perp,\mathrm{side},\min}$
1st moment of neighbour distance	$\overline{\lvert R \rvert} = \frac{1}{N}\sum_j \lvert \vec{R}_j \rvert$	$\langle \overline{\lvert R \rvert} \rangle$	$\overline{\lvert R \rvert}_{\max}$	$\overline{\lvert R \rvert}_{\min}$
2nd moment of neighbour distance	$\overline{\lvert R \rvert^2} = \frac{1}{N}\sum_j \lvert \vec{R}_j \rvert^2$	$\langle \overline{\lvert R \rvert^2} \rangle$	$\overline{\lvert R \rvert^2}_{\max}$	$\overline{\lvert R \rvert^2}_{\min}$
Standard variation of neighbour distance	$\sigma_R = \sqrt{\overline{\lvert R \rvert^2} - \overline{\lvert R \rvert}^2}$	$\langle \sigma_R \rangle$	$\sigma_{R,\max}$	$\sigma_{R,\min}$
Bond order parameters, for $n = 2$, …, 12	$\psi_n = \left\lvert \frac{1}{N}\sum_j \cos(n\phi_j)^2 + \sin(n\phi_j)^2 \right\rvert$	$\langle \psi_n \rangle$	$\psi_{n,\max}$	$\psi_{n,\min}$
Bond order parameters, for $n = N$	$\psi_N = \left\lvert \frac{1}{N}\sum_j \cos(N\phi_j)^2 + \sin(N\phi_j)^2 \right\rvert$	$\langle \psi_N \rangle$	$\psi_{N,\max}$	$\psi_{N,\min}$

information about individual cells, and non-local features, which depend on the properties of a cell's neighbours. Since motile cells tend to deform their neighbourhoods more significantly, examining non-local features provides additional insights. These non-local properties include neighbour averages, and maximum and minimum distances between the centres of mass of a cell and its neighbouring cells. Cells are classified as neighbours when they share at least one pixel. Similar to previous work,[83] local shape alignment between neighbouring cells has also been included. Table 1 illustrates local shape features in blue, local structure features in green, non-local shape features in orange, and non-local structure features in violet. The distributions of various features used in the ML model are provided in the ESI.†[71]

Note that the list of features used here is by no means complete. Depending on the specific biological situations, other features could be relevant as well. For example, individual human bone marrow stromal cells (hBMSCs) exhibit strong surface curvature, which can also be a relevant shape characteristic to include.[84] These features have not been included here, since the cells in the simulations do not exhibit strong curvature. Moreover, it is worth noting that additional radial and angular descriptors can be incorporated into the structural features, as outlined previously.[42] However, we choose to focus on a simpler approach for computational efficiency[23] and because, as will be explained in the results section, our approach, though simple, is robust and provides sufficiently accurate results.

2.4 Feature selection

To achieve optimal performance and gain physical insight from the ML predictions, we evaluate the importance of input features using three different approaches: manually removing some features, Shapley additive explanation (SHAP), and principal component analysis (PCA). The first approach involves training seven different neural networks: one with all the features and the remaining six with subsets of the entire dataset. These subsets include shape features (both local and non-local), local shape features, non-local shape features, structural features (both local and non-local), local structural features, and non-local structural features. After training, we evaluate which neural network achieves the highest accuracy on the test set.

Our second approach involves using SHAP[85] to determine the relative contribution of each feature to the prediction. In essence, the SHAP explanation method computes Shapley values by integrating concepts from cooperative game theory. The objective of this analysis is to distribute the total payoff among players, considering the significance of their contributions to the final outcome. In this context, the feature values act as players, the model represents the coalition, and the payoff corresponds to the model's prediction.

Lastly, our third approach involves applying PCA[86] on our dataset, including shape and structural features. PCA is a valuable tool for condensing multidimensional data with correlated variables into new variables, representing linear combinations of the original ones. Essentially, PCA serves as a method to reduce the dimensionality of high-dimensional data. By identifying the features with significant variances, we can reveal the inherent characteristics within our dataset. The first component corresponds to the projection axis that maximises variance in a particular direction, whereas the second principal component represents an orthogonal projection axis that maximises variance along the subsequent leading direction. This iterative process can be continued to identify additional components.

3 Results and discussion

3.1 Distinguishing motile and non-motile cells

Let us first focus on the situation in which non-motile cells are passive ($\kappa_p = 0$) and the high-motility cells are active ($\kappa_a = 1500$). This system represents a purely active–passive mixture. Fig. 3 shows the accuracy as a function of the number of active cells, N_a. The neural network is trained with different feature configurations, encompassing either all 145 features (black dots), all shape features only (both local and non-local, represented by red stars), solely local shape features (blue triangles), exclusively non-local shape features (orange inverted triangles), and only structural features (both local and non-local, represented by green squares). All five curves produce comparable accuracies, approaching unity when a single active cell moves through a non-motile confluent layer. This result is expected given that the active cell, characterised by a more elongated shape compared to the passive cells (see Fig. S2 in the ESI†[71]), is the only one present, making it easily distinguishable even to the naked eye (see Fig. 2). The elongated shape is accompanied by local distortions in an otherwise ordered confluent layer (see Fig. S4 in ESI†[71]), which explains why the accuracy is highest for $N_a = 1$. Across all four datasets, as the number of motile cells increases, the accuracy decreases. This can be attributed to a change in shape and structure for both the motile and the non-

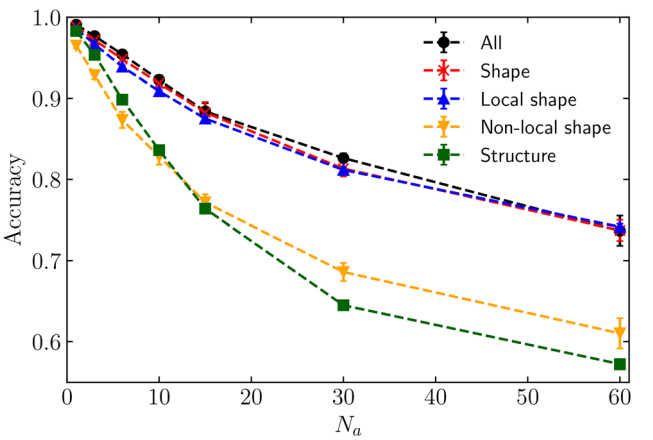


Fig. 3 Accuracy as a function of the number of active particles N_a, with $\kappa_p = 0$ and $\kappa_a = 1500$. The black dots, red stars, blue triangles, inverted triangles and green squares correspond to a neural network trained on all the 145 features, all the shape features, local shape features, non-local shape features, and structural features respectively. The lines are used as a guide to the eye.

motile cells (see Fig. S2–S11 in ESI†[71]). The cell shapes become more similar and the differences in features are more difficult to measure directly. Nevertheless, when the neural network is trained with all features, all shape features, or only local shape features, the accuracy remains above 0.7, suggesting that the algorithm can classify the cell's motility with reasonable accuracy even when cells become more similar. Some of the shape features (see *e.g.* Fig. S8 in ESI†[71]) still possess unique characteristics that allows the ML model to distinguish between the motile and non-motile cells, even for larger N_a. On the contrary, employing structural features alone results in significantly lower accuracy compared to the full dataset or the shape features. Hence, the shape of individual cells contains a substantial amount of information regarding the cell's motility in a purely active–passive mixture.

To gain a deeper understanding of the importance of shape features, we have further subdivided the shape features into local features (single-cell information) and non-local features (information from neighbouring cells). As shown in Fig. 3, it is noteworthy that relying solely on local shape features predicts the correct cell phenotype with almost the same accuracy as using the full set of features.

Apart from the accuracy, we have also investigated the types of errors made by our machine learning model. The model can generate false negatives (high-motility cell is not identified as motile) or false positives (zero-motility cells identified as motile). These results for the ML models trained on all features and local shape features are presented in Fig. S13 in the ESI.†[71] This analysis shows that when all features are used, both types of error occur with approximately the same frequency, with a slight bias toward not identifying the active cell as N_a increases. When only local shape features are used, active cells are missed more frequently, while false identification of passive cells as active is less likely. This suggests that the local environment actually contains some information to improve the prediction of an active cell in a confluent layer. Despite these differences, overall performance remains similar and is still reliable for predicting cell phenotype.

Lastly, we perform analyses using both SHAP and PCA. Both reveal that the list of important features is not limited to local shape features but rather encompasses a combination of the four feature groups: shape (both local and non-local) and structural (both local and non-local) features. Retraining the neural network with the features selected by SHAP or with the principal components obtained from the PCA yields an accuracy almost identical to that obtained with a neural network trained with all features.[71] As shown in Fig. S16 and Table S3 in the ESI,†[71] these analyses indicate that features related to neighbour distance (*e.g.*, standard deviation of neighbour distance) are often the most important ones. Since the neighbour distance-related features are indirectly connected to the shape of the cells, it is perhaps not surprising that these analyses identify these features as the most important ones.

Although SHAP and PCA reveal that the most important features are a combination of both shape and structure, the list of relevant features selected by these machine-learning approaches changes with N_a, making these analyses less computationally efficient. This inefficiency arises from the need to repeat these analyses (SHAP or PCA) for each specific configuration to obtain this list of most important features. Therefore, we can conclude that our simpler approach of selecting only shape features is sufficient for achieving reasonable accuracy for our simplest CP model and is the most robust, consistently yielding results almost identical to those obtained using all features, regardless of N_a.

3.2 Distinguishing cells with different motility

In the previous section, we have shown that a neural network trained with local shape features can correctly predict the cells motility when the passive cells are non-motile ($\kappa_p = 0$) and the active cells are significantly more motile ($\kappa_a = 1500$). However, in realistic heterogeneous biological tissues, cells are expected to have different but finite degrees of motility.[10,87,88] To study a system that more closely resembles actual biological systems, albeit still simplified, we focus on a binary mixture of high-motility and low-motility cells. The low-motility cells have a fixed motility $\kappa_p = 150$. Their motility remains lower than that of highly-motile cells ($\kappa_p < \kappa_a$), where κ_a is varied between 300 and 1500 to represent a wide range of potentially relevant biological systems.

Following a similar approach as in the previous section, we train a neural network for each dataset using static properties, as introduced in Section 2.3. Here, each dataset corresponds to a distinct ratio between low and high cell motility, denoted as $\gamma = \kappa_p/\kappa_a$, along with the number of highly motile cells N_a. Fig. 4 shows the accuracy within the (γ, N_a)-plane for a neural network trained with only local shape features. Consistent with the results observed for non-motile (passive) cells in the previous section, the neural network exclusively trained on local shape features has nearly identical accuracy compared to the one trained with all 145 features (see Fig. S14 in ESI†[71]). This figure shows that when the number of highly motile cells N_a is low, and the ratio between cell motility γ is small, indicating a substantial difference between high-motility and low-motility cells, the model can accurately classify the cell motility.

While the machine learning model relies on individual static images, our numerical CPM simulations also enable the explicit tracking of the emergent dynamics. Notably, we find that our machine learning model tends to fail only when the emergent dynamics, specifically the long-time diffusion coefficients, of high-motility and low-motility cells are very similar (see Fig. S15 in the ESI†).[71] These findings align with those presented in earlier work,[23] where it was shown that in an active–passive mixture of spherical, rigid particles, a machine learning model can correctly classify particle types when the number of active particles is low, and the activity is high.

Finally, invoking a SHAP analysis or PCA, we achieve accurate predictions using only the most important SHAP- or PCA-selected features (see Fig. S16 and Table S3 in the ESI†[71]). Similarly to the previous section, where the cells are passive, we observe that the most important features identified by these

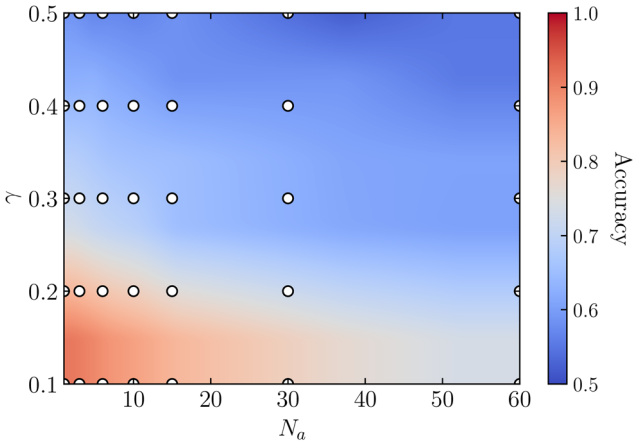


Fig. 4 Accuracy map of a neural network trained on local shape features in the (γ, N_a)-plane, where N_a ranges from 1 to 60 and $\gamma = \kappa_p/\kappa_a$, with $\kappa_p = 150$ and $300 \leq \kappa_a \leq 1500$. The data points are shown in white, and the accuracy is interpolated using a linear interpolation method.

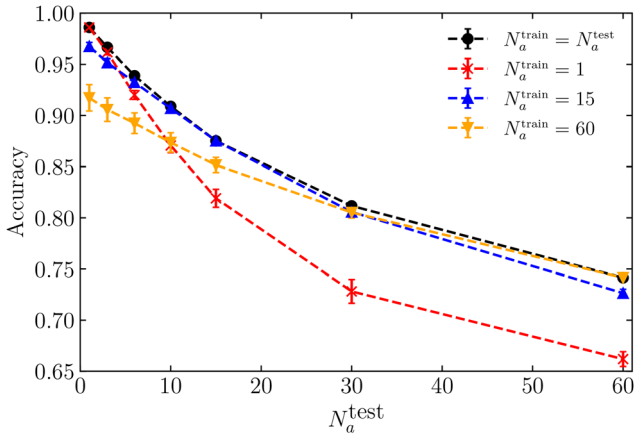


Fig. 5 Accuracy as a function of the number of active cells $N_a = N_a^{\text{test}}$, with $\kappa_p = 0$ and $\kappa_a = 1500$. The black dots represent accuracy obtained from individual models, each trained using $N_a^{\text{train}} = N_a^{\text{test}}$. The red stars, blue triangles, and orange inverted triangles represent accuracy obtained from a single model trained with data for $N_a = 1$, $N_a = 15$ or $N_a = 60$, respectively. Each neural network is trained exclusively with local shape features.

analyses are a combination of shape (both local and non-local) and structural (both local and non-local) features. While this feature list remains consistent for fixed $\gamma > 0$ and different N_a, it varies for different γ. Consequently, as discussed in the previous section, this approach is less computationally efficient compared to the case of using local shape features, which yields accurate results for different configurations.

In summary, our findings indicate that our machine-learning model can accurately classify cell motility when the number of motile cells is low, and the motility of highly-motile cells significantly surpasses that of low-motility cells. Comparable to the case in which the low-motility cells are passive, accurate predictions can be achieved using local shape features alone. While this approach shows its potential for the simplified CP model that we have developed, we speculate that these results can generalize to other computational models and potentially to experimental results.

3.3 Generalisation of the model

We now aim to assess how effectively our machine-learning model generalises to different data featuring either a distinct number of motile cells or varying motility. First, we explore the generalisation capability of the machine learning model when the low-motility cells are passive ($\kappa_p = 0$), and the number of motile cells N_a varies. In the previous section, we established that a model trained exclusively on local shape features achieves an accuracy almost identical to that of a model trained with all 145 features. Therefore, in the remainder of this paper, we present the results from neural networks trained exclusively on local shape features.

In Fig. 5, we compare the accuracy obtained when the ML model is trained and tested on a singular specific value of the number of motile cells N_a (black dots), with the accuracy of models trained with $N_a = 1$ (red stars), $N_a = 15$ (blue triangles), or $N_a = 60$ (orange inverted triangles). As expected, this figure shows that the model performs best when trained and tested on a singular, specific value of N_a. Nevertheless, all four curves yield an accuracy surpassing 0.7, suggesting that a single model trained at a fixed N_a can provide accurate predictions for unseen parameter regions.

Fig. 5 reveals that the ML model trained with an intermediate number of motile cells, $N_a = 15$, yields nearly identical results compared to the model trained and tested on a single, specific value of N_a. This model is the most effective across the entire range of N_a. We attribute this to the fact that a system with an intermediate number of motile cells shares similarities with both low and high numbers of motile cells, contributing to its robust performance. Additionally, while the model trained on one motile cell generalises better to different data associated with a small number of motile cells ($N_a < 10$), the model trained on $N_a = 60$ generalises better to different data corresponding to a high number of motile cells ($N_a > 30$). This discrepancy arises from the distinctive system structures (see feature distributions in ESI†[71]) between scenarios with only one motile cell and those with a substantial number, respectively.

Lastly, we explore whether the machine learning approach can generalise to a different data set when the number of motile cells is constant, and the ratio between cell motilities γ varies. For each value of N_a, we train four distinct models: one with $\gamma = 0$ (where $\kappa_p = 0$ and $\kappa_a = 1500$), another with $\gamma = 0.1$ (where $\kappa_p = 150$ and $\kappa_a = 1500$), a third with $\gamma = 0.2$ (where $\kappa_p = 150$ and $\kappa_a = 750$), and the last one with $\gamma = 0.4$ (where $\kappa_p = 150$ and $\kappa_a = 375$). Subsequently, each of these four models is tested with the local shape features corresponding to the dataset with a fixed ratio $\gamma = 0.1$. We have decided to test the generalization of the ML model with $\gamma = 0.1$, as the previous section has shown that the accuracy is highest for this mixture of high-motility and low-motility cells.

Fig. 6 demonstrates that, as expected, the highest performance is achieved by a model trained and tested on the

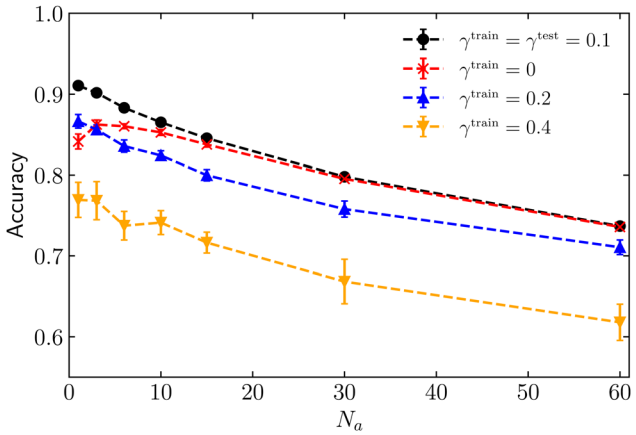


Fig. 6 Accuracy as a function of the number of active cells N_a, for neural networks solely trained on local shape features. The black dots represent accuracy obtained from individual models, each trained using $\gamma^{train} = \gamma^{test} = 0.1$. The red stars, blue triangles, and orange inverted triangles represent accuracy obtained from models trained with data for $\gamma = 0$, $\gamma = 0.2$ or $\gamma = 0.4$, respectively. The accuracy for each of these lines corresponds to the neural network tested on $\gamma = 0.1$.

identical ratio between cell motility, $\gamma = 0.1$ (black dots). Additionally, the figure shows that a model trained on $\gamma = 0$ (represented by red stars) yields an accuracy nearly indistinguishable from the model trained and tested on $\gamma = 0.1$ when the number of motile cells is high ($N_a > 10$). In both datasets corresponding to $\gamma = 0$ and 0.1, the motility of the highly-motile cells remains constant. Consequently, the model exhibits effective generalisation within this parameter range, even if is trained on data associated with different low motility. This generalisation can be attributed to the similarity in behaviour between the two systems, given the abundant high-motile cells sharing the same motility.

When the model is trained on $\gamma = 0.2$ and tested on $\gamma = 0.1$ (blue triangles), the accuracy is always lower than that of the model trained and tested on $\gamma = 0.1$. Nonetheless, the accuracy is consistently higher than 0.7, indicating that this model can reasonably generalise unseen data. Lastly, when the model is trained on $\gamma = 0.4$ (orange inverted triangles), the accuracy significantly diminishes compared to the model trained and tested on $\gamma = 0.1$. Furthermore, the accuracy drops below 0.7 as the number of motile cells increases ($N_a > 20$). These results indicate that a decrease in the motility of motile cells corresponds to a lower predictive power of the model when tested on unseen data. We expect that testing the trained networks with larger values for γ also makes it more difficult to generalize as the baseline prediction (see Fig. 4) is significantly worse for this parameter in the first place.

In summary, we find that the generalisation capability of our machine-learning approach to different unseen data is reasonable. The model is capable of making fairly accurate predictions when the number of motile cells is unknown, but its predictive power diminishes when the motility of the highly-motile cells in the training and testing sets are significantly different.

4 Conclusions

This study establishes proof-of-concept for discriminating between highly motile (active) cells and less motile or non-motile (passive) cells within a heterogeneous confluent cell layer, using only static information of a cell's instantaneous shape and structural environment. We have developed the confluent layer in a cellular Potts model with minimal ingredients such that we could control the motility of the two cell phenotypes in a simplistic manner. Our results are valid for our CP model, but we expect these findings to hold among other computational models and real-world biological systems with similar characteristics. Our results show that a simple machine-learning model trained on local, single-cell shape features alone can predict the cellular motility phenotype with reasonably good accuracy, and excels especially when the fraction of highly-motile cells is low and their motility is significantly higher than that of low-motility cells.[23] While prior studies have also highlighted the importance of cell shape and morphology,[9,60–62] most notably in strongly anisotropic tissues, other measures, such as the alignment between cells, may be necessary for a fully accurate prediction[83] as these tissues behave differently from the cells in our confluent layer model.

Common limitations of machine-learning approaches are that they may generalise poorly to unseen data, and that they may offer limited physical insight. We find that our model exhibits reasonably good generalisation when the number of motile cells or the motility ratio is unknown, provided that the motility strengths in the training and testing sets do not differ greatly. This reaffirms that the power of machine-learning methods relies heavily on the use of a sufficiently diverse data set. Additionally, to gain some physical insight from our machine-learning predictions, we have employed three different methods to assess the importance of the various input features. Of these, the analyses based on SHAP and PCA reveal that there is not a universal list of most important static features: in general, the most important features combine cellular shape and structural characteristics, and the list varies with different heterogeneity settings (N_a). Nonetheless, if we restrict the data set to local single-cell shape features alone, we find that this simple approach leads to remarkably robust predictions across the different settings studied in this work. This suggests that the full list of structural input features may contain some redundancies. Importantly, it also allows us to conclude that a cell's instantaneous shape, though not perfect, can serve as a remarkably useful informant on a cell's phenotype.

Our work, which establishes a morphodynamic link for individual cells, is complementary to recent research on morphodynamic links at the collective cell level. In particular, previous studies have demonstrated that the average cell shape within confluent tissue can be used as a static order parameter for emergent, collective cell jamming and unjamming dynamics.[9–16] By integrating these insights, our work not only reinforces the significance of cell shape in understanding collective behaviour, but it also provides a more nuanced perspective

on how intrinsic single-cell properties are coupled to a cell's morphology. This study also opens up avenues for further research on the role of heterogeneity in dense cell collectives, following previous work that has studied the heterogeneity in size and softness of cells.[89,90] Given the simplicity, performance, and computational efficiency of our machine-learning approach, we anticipate that a similar approach could ultimately prove valuable in analyzing experimental cell data—particularly for diagnostic tasks like assessing the progression of partial or complete EMT in tumors or tissues.

Author contributions

QJSB, GJ, BCJ, VED, SC, LMCJ designed the research. QJSB, GJ, BCJ contributed equally and conducted the research, establishing the methodology, carrying out the analysis and validating the data. BCJ, SC developed the software. QJSB, GJ wrote the initial manuscript. LMCJ supervised the project. All authors contributed to editing the manuscript, to the discussions, and provided meaningful insights.

Conflicts of interest

There are no conflicts to declare.

Data availability

The dataset of extracted features and the machine learning results supporting the findings of this study are openly available on Zenodo at **https://doi.org/10.5281/zenodo.15699073**.

Acknowledgements

QJSB and LMCJ thank the Dutch Research Council (NWO) for financial support through the ENW-XL project "Active Matter Physics of Collective Metastasis" (OCENW.GROOT.2019.022). GJ acknowledges support from the Comunidad de Madrid and the Complutense University of Madrid (Spain) through the Atracción de Talento program (Grant No. 2022-T1/TIC-24007).

References

1 C. H. Stuelten, C. A. Parent and D. J. Montell, *Nat. Rev. Cancer*, 2018, **18**, 296–312.
2 P. Friedl and D. Gilmour, *Nat. Rev. Mol. Cell Biol.*, 2009, **10**, 445–457.
3 X. Trepat, Z. Chen and K. Jacobson, *Cell Migration*, John Wiley & Sons, Ltd, 2012, pp. 2369–2392.
4 F. Merino-Casallo, M. J. Gomez-Benito, S. Hervas-Raluy and J. M. Garcia-Aznar, *Cell Adhes. Migr.*, 2022, **16**, 25–64.
5 P. Gottheil, J. Lippoldt, S. Grosser, F. Renner, M. Saibah, D. Tschodu, A.-K. Poßögel, A.-S. Wegscheider, B. Ulm, K. Friedrichs, C. Lindner, C. Engel, M. Löffler, B. Wolf, M. Höckel, B. Aktas, H. Kubitschke, A. Niendorf and J. A. Käs, *Phys. Rev. X*, 2023, **13**, 031003.
6 S. Grosser, J. Lippoldt, L. Oswald, M. Merkel, D. M. Sussman, F. Renner, P. Gottheil, E. W. Morawetz, T. Fuhs, X. Xie, S. Pawlizak, A. W. Fritsch, B. Wolf, L.-C. Horn, S. Briest, B. Aktas, M. L. Manning and J. A. Käs, *Phys. Rev. X*, 2021, **11**, 011033.
7 J.-A. Park, L. Atia, J. A. Mitchel, J. J. Fredberg and J. P. Butler, *J. Cell Sci.*, 2016, **129**, 3375–3383.
8 J. Käs, J. Lippoldt, S. Grosser, D. Tschodu, F. Renner, H. Kubitschke, A.-K. Poßögel, A.-S. Wegscheider, B. Ulm, K. Friedrichs, *et al.*, *Cancer cell motility through unjamming impacts metastatic risk*, 2022.
9 D. Bi, J. H. Lopez, J. M. Schwarz and M. L. Manning, *Nat. Phys.*, 2015, **11**, 1074–1079.
10 D. Bi, X. Yang, M. C. Marchetti and M. L. Manning, *Phys. Rev. X*, 2016, **6**, 021011.
11 M. Chiang and D. Marenduzzo, *Europhys. Lett.*, 2016, **116**, 28009.
12 M. Czajkowski, D. Bi, M. L. Manning and M. C. Marchetti, *Soft Matter*, 2018, **14**, 5628–5642.
13 A. J. Devanny, D. J. Lee, L. Kampman and L. J. Kaufman, *bioRxiv*, 2023, preprint, DOI: **10.1101/2023.07.10.548321**.
14 L. Atia, D. Bi, Y. Sharma, J. A. Mitchel, B. Gweon, S. A. Koehler, S. J. DeCamp, B. Lan, J. H. Kim, R. Hirsch, A. F. Pegoraro, K. H. Lee, J. R. Starr, D. A. Weitz, A. C. Martin, J.-A. Park, J. P. Butler and J. J. Fredberg, *Nat. Phys.*, 2018, **14**, 613–620.
15 J.-A. Park, J. H. Kim, D. Bi, J. A. Mitchel, N. T. Qazvini, K. Tantisira, C. Y. Park, M. McGill, S.-H. Kim, B. Gweon, J. Notbohm, R. Steward Jr, S. Burger, S. H. Randell, A. T. Kho, D. T. Tambe, C. Hardin, S. A. Shore, E. Israel, D. A. Weitz, D. J. Tschumperlin, E. P. Henske, S. T. Weiss, M. L. Manning, J. P. Butler, J. M. Drazen and J. J. Fredberg, *Nat. Mater.*, 2015, **14**, 1040–1048.
16 J. H. Kim, A. F. Pegoraro, A. Das, S. A. Koehler, S. A. Ujwary, B. Lan, J. A. Mitchel, L. Atia, S. He, K. Wang, D. Bi, M. H. Zaman, J.-A. Park, J. P. Butler, K. H. Lee, J. R. Starr and J. J. Fredberg, *Biochem. Biophys. Res. Commun.*, 2020, **521**, 706–715.
17 P.-H. Wu, D. M. Gilkes, J. M. Phillip, A. Narkar, T. W.-T. Cheng, J. Marchand, M.-H. Lee, R. Li and D. Wirtz, *Sci. Adv.*, 2020, **6**, eaaw6938.
18 S. P. Carey, A. Starchenko, A. L. McGregor and C. A. Reinhart-King, *Clin. Exp. Metastasis*, 2013, **30**, 615–630.
19 R. Kalluri and R. Weinberg, *J. Clin. Invest.*, 2009, **119**, 1420–1428.
20 M. K. Jolly and T. Celià-Terrassa, *J. Clin. Med.*, 2019, **8**, 1542.
21 L. A. Hapach, S. P. Carey, S. C. Schwager, P. V. Taufalele, W. Wang, J. A. Mosier, N. Ortiz-Otero, T. J. McArdle, Z. E. Goldblatt, M. C. Lampi, F. Bordeleau, J. R. Marshall, I. M. Richardson, J. Li, M. R. King and C. A. Reinhart-King, *Cancer Res.*, 2021, **81**, 3649–3663.
22 C. E. Meacham and S. J. Morrison, *Nature*, 2013, **501**, 328–337.
23 G. Janzen, X. L. J. A. Smeets, V. E. Debets, C. Luo, C. Storm, L. M. C. Janssen and S. Ciarella, *Europhys. Lett.*, 2023, **143**, 17004.

24 F. Graner and J. A. Glazier, *Phys. Rev. Lett.*, 1992, **69**, 2013–2016.
25 J. A. Glazier and F. Graner, *Phys. Rev. E: Stat. Phys., Plasmas, Fluids, Relat. Interdiscip. Top.*, 1993, **47**, 2128–2154.
26 A. F. M. Marée, V. A. Grieneisen and P. Hogeweg, *The Cellular Potts Model and Biophysical Properties of Cells, Tissues and Morphogenesis*, Birkhäuser Basel, Basel, 2007, pp. 107–136.
27 P. Albert and U. Schwarz, *Biophys. J.*, 2014, **106**, 2340–2352.
28 A. Voss-Böhme, *PLoS One*, 2012, **7**, 1–14.
29 S. Sadhukhan and S. K. Nandi, *Phys. Rev. E*, 2021, **103**, 062403.
30 N. Sepúlveda, L. Petitjean, O. Cochet, E. Grasland-Mongrain, P. Silberzan and V. Hakim, *PLoS Comput. Biol.*, 2013, **9**, e1002944.
31 M. Scianna and L. Preziosi, *Axioms*, 2021, **10**, 32.
32 E. D. Cubuk, S. S. Schoenholz, J. M. Rieser, B. D. Malone, J. Rottler, D. J. Durian, E. Kaxiras and A. J. Liu, *Phys. Rev. Lett.*, 2015, **114**, 108001.
33 E. D. Cubuk, R. J. S. Ivancic, S. S. Schoenholz, D. J. Strickland, A. Basu, Z. S. Davidson, J. Fontaine, J. L. Hor, Y.-R. Huang, Y. Jiang, N. C. Keim, K. D. Koshigan, J. A. Lefever, T. Liu, X.-G. Ma, D. J. Magagnosc, E. Morrow, C. P. Ortiz, J. M. Rieser, A. Shavit, T. Still, Y. Xu, Y. Zhang, K. N. Nordstrom, P. E. Arratia, R. W. Carpick, D. J. Durian, Z. Fakhraai, D. J. Jerolmack, D. Lee, J. Li, R. Riggleman, K. T. Turner, A. G. Yodh, D. S. Gianola and A. J. Liu, *Science*, 2017, **358**, 1033–1037.
34 E. D. Cubuk, S. S. Schoenholz, E. Kaxiras and A. J. Liu, *J. Phys. Chem. B*, 2016, **120**, 6139–6146.
35 D. M. Sussman, M. Paoluzzi, M. C. Marchetti and M. L. Manning, *EPL*, 2018, **121**, 36001.
36 E. Boattini, M. Ram, F. Smallenburg and L. Filion, *Mol. Phys.*, 2018, **116**, 3066–3075.
37 E. Boattini, M. Dijkstra and L. Filion, *J. Chem. Phys.*, 2019, **151**, 154901.
38 S. S. Schoenholz, E. D. Cubuk, D. M. Sussman, E. Kaxiras and A. J. Liu, *Nat. Phys.*, 2016, **12**, 469–471.
39 V. Bapst, T. Keck, A. Grabska-Barwińska, C. Donner, E. D. Cubuk, S. S. Schoenholz, A. Obika, A. W. R. Nelson, T. Back, D. Hassabis and P. Kohli, *Nat. Phys.*, 2020, **16**, 448–454.
40 J. Paret, R. L. Jack and D. Coslovich, *J. Chem. Phys.*, 2020, **152**, 144502.
41 F. M. C. P. Landes, G. Biroli, O. Dauchot, A. J. Liu and D. R. Reichman, *Phys. Rev. E*, 2020, **101**, 010602.
42 E. Boattini, F. Smallenburg and L. Filion, *Phys. Rev. Lett.*, 2021, **127**, 88007.
43 R. M. Alkemade, E. Boattini, L. Filion and F. Smallenburg, *J. Chem. Phys.*, 2022, **156**, 204503.
44 N. Oyama, S. Koyama and T. Kawasaki, *Front. Phys.*, 2023, **10**, 1–17.
45 I. Tah, S. A. Ridout and A. J. Liu, *J. Chem. Phys.*, 2022, **157**, 124501.
46 G. Jung, G. Biroli and L. Berthier, *Phys. Rev. Lett.*, 2023, **130**, 238202.
47 S. Ciarella, M. Chiappini, E. Boattini, M. Dijkstra and L. M. C. Janssen, *Mach. Learn.: Sci. Technol.*, 2023, **4**, 025010.
48 D. Coslovich, R. L. Jack and J. Paret, *J. Chem. Phys.*, 2022, **157**, 204503.
49 S. Ciarella, D. Khomenko, L. Berthier, F. C. Mocanu, D. R. Reichman, C. Scalliet and F. Zamponi, *Nat. Commun.*, 2023, **14**, 4229.
50 R. M. Alkemade, F. Smallenburg and L. Filion, *J. Chem. Phys.*, 2023, **158**, 134512.
51 G. Janzen, C. Smit, S. Visbeek, V. E. Debets, C. Luo, C. Storm, S. Ciarella and L. M. C. Janssen, *Phys. Rev. Mater.*, 2024, **8**, 025602.
52 F. Cichos, K. Gustavsson, B. Mehlig and G. Volpe, *Nat. Mach. Intell.*, 2020, **2**, 94–103.
53 J. M. Newby, A. M. Schaefer, P. T. Lee, M. G. Forest and S. K. Lai, *Proc. Natl. Acad. Sci. U. S. A.*, 2018, **115**, 9026–9031.
54 H. Jeckel, E. Jelli, R. Hartmann, P. K. Singh, R. Mok, J. F. Totz, L. Vidakovic, B. Eckhardt, J. Dunkel and K. Drescher, *Proc. Natl. Acad. Sci. U. S. A.*, 2019, **116**, 1489–1494.
55 S. Bo, F. Schmidt, R. Eichhorn and G. Volpe, *Phys. Rev. E*, 2019, **100**, 10102.
56 G. M. Gil, M. A. Garcia-March, C. Manzo, J. D. Martín-Guerrero and M. Lewenstein, *New J. Phys.*, 2020, **22**, 013010.
57 I. Tah, T. A. Sharp, A. J. Liu and D. M. Sussman, *Soft Matter*, 2021, **17**, 10242–10253.
58 S. Bag and R. Mandal, *Soft Matter*, 2021, **17**, 8322–8330.
59 M. Ruiz-Garcia, C. M. B. Gutierrez, L. C. Alexander, D. G. A. L. Aarts, L. Ghiringhelli and C. Valeriani, *Phys. Rev. E*, 2024, **109**, 06461.
60 S. M. Lyons, E. Alizadeh, J. Mannheimer, K. Schuamberg, J. Castle, B. Schroder, P. Turk, D. Thamm and A. Prasad, *Biol. Open*, 2016, **5**, 289–299.
61 M. D'Orazio, F. Corsi, A. Mencattini, D. Di Giuseppe, M. Colomba Comes, P. Casti, J. Filippi, C. Di Natale, L. Ghibelli and E. Martinelli, *Front. Oncol.*, 2020, **10**(1–11), 580698.
62 M. Kim, Y. Namkung, D. Hyun and S. Hong, *Adv. Intell. Syst.*, 2023, **5**, 2300017.
63 H. Yang, F. Meyer, S. Huang, L. Yang, C. Lungu, M. A. Olayioye, M. J. Buehler and M. Guo, *PRX Life*, 2024, **2**, 043010.
64 D. Frenkel and B. Smit, *Understanding Molecular Simulation*, Academic Press, San Diego, 2nd edn, 2002, pp. 23–61.
65 H. Nemati and J. de Graaf, *The Cellular Potts Model on Disordered Lattices*, 2024, **https://arxiv.org/abs/2404.09055**.
66 M. H. Swat, G. L. Thomas, J. M. Belmonte, A. Shirinifard, D. Hmeljak and J. A. Glazier, *Computational Methods in Cell Biology*, Academic Press, 2012, vol. 110, pp. 325–366.
67 N. Guisoni, K. I. Mazzitello and L. Diambra, *Front. Phys.*, 2018, **6**(1–11), 61.
68 V. E. Debets, L. M. Janssen and C. Storm, *Biophys. J.*, 2021, **120**, 1483–1497.
69 S. SenGupta, C. A. Parent and J. E. Bear, *Nat. Rev. Mol. Cell Biol.*, 2021, **22**, 259–547.
70 T. E. Angelini, E. Hannezo, X. Trepat, M. Marquez, J. J. Fredberg and D. A. Weitz, *Proc. Natl. Acad. Sci. U. S. A.*, 2011, **108**, 4714–4719.
71 ESI†.
72 M. W. Gardner and S. R. Dorling, *Atmos. Environ.*, 1998, **32**, 2627–2636.

73 S. Pal and S. Mitra, *IEEE Trans. Neural Networks*, 1992, **3**, 683–697.
74 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.
75 S. Haykin, *Neural Networks and Learning Machines*, Pearson, 2009.
76 D. P. Kingma and J. Ba, *arXiv*, 2014, preprint, arXiv: 1412.6980, DOI: **10.48550/arXiv.1412.6980**.
77 G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye and T.-Y. Liu, Proceedings of the 31st International Conference on Neural Information Processing Systems, Red Hook, NY, USA, 2017, pp. 3149–3157.
78 C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer-Verlag, Berlin, Heidelberg, 2006.
79 H. F. Mahmoud, *arXiv*, 2019, preprint, arXiv:1906.10221, DOI: **10.48550/arXiv.1906.10221**.
80 Y. Tokuda, M. Fujisawa, D. M. Packwood, M. Kambayashi and Y. Ueda, *AIP Adv.*, 2020, **10**, 105110.
81 W. Mickel, S. C. Kapfer, G. E. Schröder-Turk and K. Mecke, *J. Chem. Phys.*, 2013, **138**, 044501.
82 M. Pilu, A. W. Fitzgibbon and R. B. Fisher, Proceedings of 3rd IEEE International Conference on Image Processing, 1996, pp. 599–602.
83 X. Wang, M. Merkel, L. B. Sutter, G. Erdemci-Tandogan, M. L. Manning and K. E. Kasza, *Proc. Natl. Acad. Sci. U. S. A.*, 2020, **117**, 13541–13551.
84 D. Chen, S. Sarkar, J. Candia, S. J. Florczyk, S. Bodhak, M. K. Driscoll, C. G. Simon, J. P. Dunkers and W. Losert, *Biomaterials*, 2016, **104**, 104–118.
85 S. M. Lundberg and S.-I. Lee, Proceedings of the 31st International Conference on Neural Information Processing Systems, Red Hook, NY, USA, 2017, pp. 4768–4777.
86 I. T. Jolliffe, *Principal component analysis for special types of data*, Springer, 2002.
87 A. E. Cerchiari, J. C. Garbe, N. Y. Jee, M. E. Todhunter, K. E. Broaders, D. M. Peehl, T. A. Desai, M. A. LaBarge, M. Thomson and Z. J. Gartner, *Proc. Natl. Acad. Sci. U. S. A.*, 2015, **112**, 2287–2292.
88 S. Ravindran, S. Rasool and C. Maccalli, *Cancer Microenviron.*, 2019, **12**, 133–148.
89 T. Fuhs, F. Wetzel, A. W. Fritsch, X. Li, R. Stange, S. Pawlizak, T. R. Kießling, E. Morawetz, S. Grosser and F. Sauer, *et al.*, *Nat. Phys.*, 2022, **18**, 1510–1519.
90 Asadullah, S. Kumar, N. Saxena, M. Sarkar, A. Barai and S. Sen, *J. Cell Sci.*, 2021, **134**, jcs250225.