

Profiling Dynamic Patterns of Single-Cell Motility

Debonil Maity, Nikita Sivakumar, Pratik Kamat, Nahuel Zamponi, Chanhong Min, Wenxuan Du, Hasini Jayatilaka, Adrian Johnston, Bartholomew Starich, Anshika Agrawal, Deanna Riley, Leandro Venturutti, Ari Melnick, Leandro Cerchietti, Jeremy Walston, and Jude M. Phillip*

Cell motility plays an essential role in many biological processes as cells move and interact within their local microenvironments. Current methods for quantifying cell motility typically involve tracking individual cells over time, but the results are often presented as averaged values across cell populations. While informative, these ensemble approaches have limitations in assessing cellular heterogeneity and identifying generalizable patterns of single-cell behaviors, at baseline and in response to perturbations. In this study, CaMI is introduced, a computational framework designed to leverage the single-cell nature of motility data. CaMI identifies and classifies distinct spatio-temporal behaviors of individual cells, enabling robust classification of single-cell motility patterns in a large dataset ($n = 74\,253$ cells). This framework allows quantification of spatial and temporal heterogeneities, determination of single-cell motility behaviors across various biological conditions and provides a visualization scheme for direct interpretation of dynamic cell behaviors. Importantly, CaMI reveals insights that conventional cell motility analyses may overlook, showcasing its utility in uncovering robust biological insights. Together, a multivariate framework is presented to classify emergent patterns of single-cell motility, emphasizing the critical role of cellular heterogeneity in shaping cell behaviors across populations.

1. Introduction

The dynamic movement of mammalian cells, that is, cell motility, underlies critical biological processes that regulate homeostasis and disease, including immune locomotion and surveillance, and the metastasis of cancer cells from primary tumor sites.^[1-4] While cell motility involves a complex coordination of cellular displacements, tortuous paths, and effective diffusivities, the current paradigm for analyzing and reporting cell motility data is often limited to individual motility parameters such as average speeds and displacements. While these approaches are informative, they offer limited capacity to assess cellular heterogeneity, extract single-cell motility patterns, and assess the abundance of subpopulations of cells that may be enriched or depleted in response to perturbations. Furthermore, single-cell profiling is critical to properly understand and interpret cellular heterogeneity of biological phenotypes.

D. Maity, N. Sivakumar, C. Min, J. M. Phillip
 Department of Biomedical Engineering
 Johns Hopkins University
 Baltimore, MD 21212, USA
 E-mail: jphillip@jhu.edu

D. Maity, N. Sivakumar, P. Kamat, C. Min, W. Du, H. Jayatilaka, A. Johnston,
 B. Starich, D. Riley, J. Walston, J. M. Phillip
 Institute for Nanobiotechnology
 Johns Hopkins University
 Baltimore, MD 21212, USA
 P. Kamat, W. Du, A. Johnston, B. Starich, A. Agrawal, J. M. Phillip
 Department of Chemical and Biomolecular Engineering
 Johns Hopkins University
 Baltimore, MD 21212, USA

N. Zamponi, A. Melnick, L. Cerchietti
 Department of Medicine
 Division of Hematology and Medical Oncology
 Weill Cornell Medicine
 New York 10065, USA

L. Venturutti
 Department of Pathology and Laboratory Medicine
 University of British Columbia
 Centre for Lymphoid Cancer
 British Columbia Cancer Research Institute Vancouver
 British Columbia V6T 1Z4, Canada

J. Walston
 Department of Medicine
 Geriatrics and Gerontology
 Johns Hopkins School of Medicine
 Baltimore, MD 21224, USA

J. M. Phillip
 Department of Oncology
 Sidney Kimmel Comprehensive Cancer Center
 Johns Hopkins School of Medicine
 Baltimore, MD 21287, USA

 The ORCID identification number(s) for the author(s) of this article can be found under <https://doi.org/10.1002/advs.202400918>

© 2024 The Author(s). Advanced Science published by Wiley-VCH GmbH. This is an open access article under the terms of the [Creative Commons Attribution License](#), which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

DOI: [10.1002/advs.202400918](https://doi.org/10.1002/advs.202400918)

Recent advances in single-cell technologies and analyses have enabled the detailed quantitative description of cell types, cell states, and cell fates based on genetic, epigenetic, or protein signatures.^[5,6] Building on these recent advances, we rationalized that combining single-cell profiling of cell motility with modern data science approaches could offer new solutions to define functional dynamic cell behaviors and cellular heterogeneities across multiple conditions.^[7–11] Furthermore, performing these single-cell behavior analyses in biological systems with known underlying molecular information (i.e., knockdowns of particular genes, drug treatments) presents the potential to trace the influence of specific molecular programs on dynamic cell behaviors. Despite this promise, robust quantification of motility patterns to inform functional behaviors and heterogeneities of mammalian cells remain challenging. This challenge is partly due to the lack of universal standards for acquiring cell motility data (i.e., image acquisition settings, such as total imaging duration and acquisition intervals), the lack of large publicly available cell motility datasets for comparative interrogation of cell motility patterns, and limited single-cell computational tools designed specifically for cell motility datasets.

To address this gap, we present the computational framework CaMI, which enables the harmonization of motility datasets from multiple sources having different acquisition settings and robust identification of single-cell motility patterns (**Figure 1A**). Using CaMI, we combined non-uniform mammalian cell motility datasets from *in vitro* experiments to profile 74,253 single cells spanning 409 biological conditions^[12–16] to derive emergent cell motility patterns at single-cell resolution for both short (2.5 h) and long (8 h) durations.

To demonstrate the biological utility of this framework, we applied CaMI to study how molecular and environmental perturbations induce fractional redistributions in cell motility behaviors in three test cases: 1) to quantify how local environmental perturbations, such as cell density and drug treatments that modulate cell signaling, can impact emergent single-cell motility properties; 2) to quantify how cytoskeleton-acting drugs modulate cell movements to drive distinct cell motility patterns, and lastly 3) to quantify emergent cell motility patterns in primary human monocytes, demonstrating utility to profile heterogeneity in single-cell responses in donor samples and a potential path for preclinical/clinical applications. Furthermore, to demonstrate the technical utility of this framework, we show how spatial and temporal heterogeneities (computed as the entropies) across distinct motility patterns can predict the dimensionality of cells moving in either 2D or 3D microenvironments. The curated motility dataset ($n = 409$ biological conditions) and the MATLAB code for the CaMI pipeline are available within the supplementary documents and publicly available here: <https://github.com/PhillipLab-JHU/CaMI>.

2. Results

2.1. Description of Data Acquisition and Curation

To develop a computational framework to identify and classify emergent patterns of single-cell motility across multiple conditions, we combined existing and newly acquired cell motility datasets.^[7,17–22] All cells were tracked using commercial cell track-

ing software (Metamorph). Following cell tracking, we generated the time-dependent coordinates (x, y, t) of cells to generate single-cell trajectories for all cells across multiple experimental conditions presented. Cell trajectories were generated based on time-lapse movies collected using brightfield or phase-contrast images. Movies ranged in duration from 2.5 to 16 h, with interval times and temporal resolutions ranging from 2 to 10 min. Below, we discuss a robust harmonization strategy to tune the temporal resolution, thereby enabling direct comparison of mammalian cell movements within short-duration (2.5 h) and long-duration (8 h) videos.

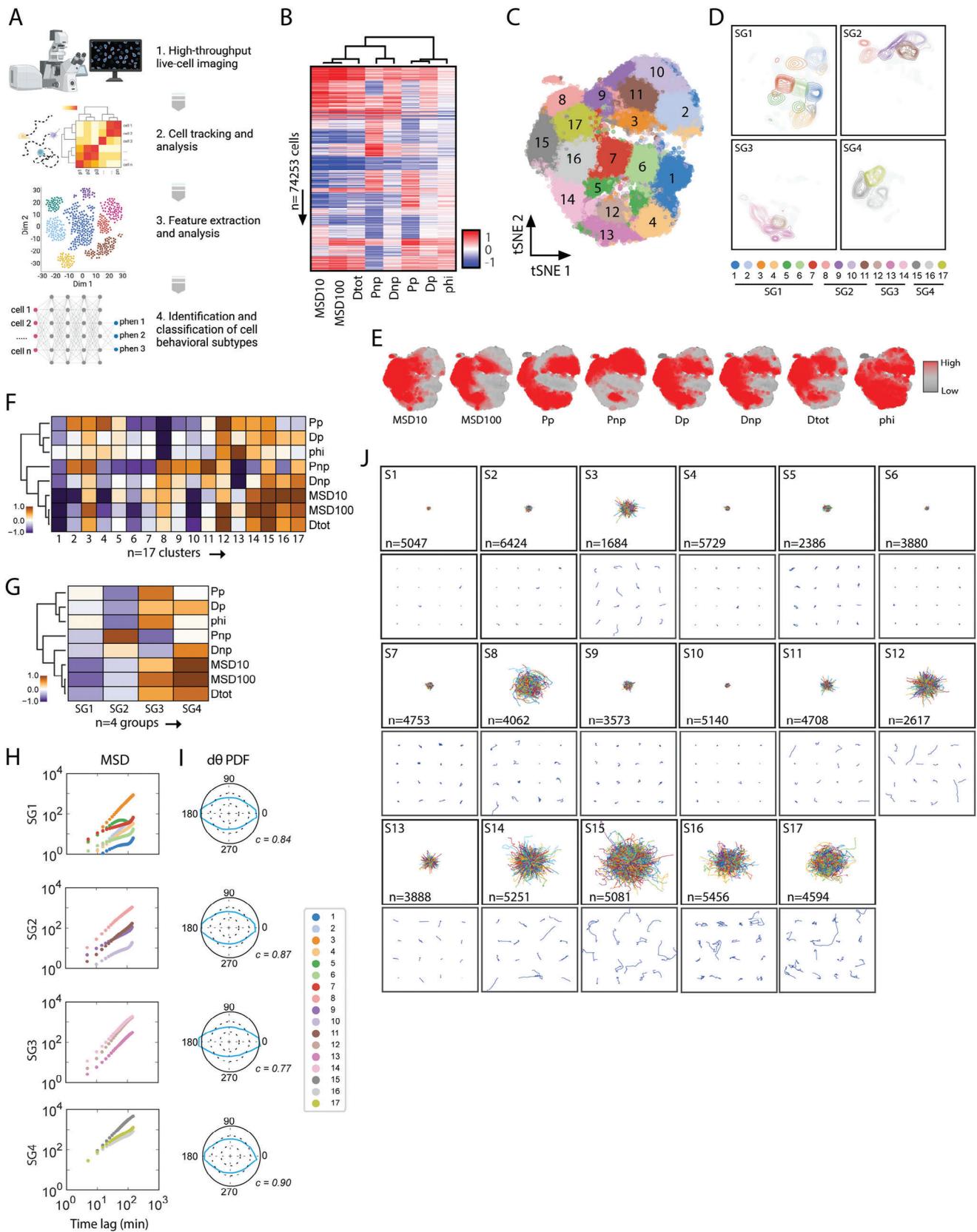
2.2. Imaging Interval Times Determine the Shape of Cell Movement Trajectories

Currently, there are no universal standards for acquiring cell motility data with regard to the total imaging duration or image acquisition intervals. While each researcher chooses imaging parameters based on a set of user-dependent optimization parameters (e.g., convenience), this often presents challenges for downstream data integration across datasets generated across multiple experiments or multiple labs. To combine motility datasets covering multiple biological conditions and from multiple sources, we first needed to integrate the data without introducing significant noise and artifacts into the datasets. We first restricted the total imaging duration to 2.5 h by cutting longer movies into shorter non-overlapping 2.5-h trajectories per single cell. This timescale was chosen not only to maximize the total number of cells within our dataset, but also to have a timescale that is achievable for both *in vitro* and *in vivo* imaging experiments, and to enable downstream mapping of short-duration motility patterns to long-duration motility patterns. After curating the 2.5 h trajectories per cell, we evaluated the effective contributions of image acquisition intervals (i.e., time between subsequent images—temporal resolution) on the magnitude of cell motility.

Using a test case of ht1080 fibrosarcoma cells moving within 3D collagen-I gels, we *in-silico* varied the interval times from the most to least time-resolved trajectories within our dataset, that is, 2 to 10 min (sample trajectories shown in **Figure S1A**, Supporting Information, top). We varied the acquisition time by excluding intervening points from the fully time-resolved trajectory. Results indicated that the interval times significantly impacted the shapes of the cell trajectories and their corresponding motility parameters. The same trajectory acquired across different interval times had significantly different geometric concordance (**Figure S1A**, Supporting Information, bottom), or the average area of triangles formed between intervening points along a given trajectory (see Experimental Section). Together, these results highlight that acquisition intervals contribute significantly to the magnitude of the perceived motility and that temporal resolution matters, that is, cell movements based on 2-min intervals do not equal cell movements at 10-min intervals.

2.3. Harmonizing Single-Cell Motility Data Across Multiple Datasets

Since the interval times determine the shape and magnitude of the cell trajectories, to accurately compare trajectories acquired



at different interval times we needed a robust approach to estimate the intervening x-y coordinates in lower resolution trajectories to harmonize the temporal resolutions of all trajectories within our dataset. Using linear interpolation, we identified spatially weighted cell locations for symmetric (8 to 4-min, i.e., midpoint) and asymmetric (8 to 6-min) multiples of the acquisition-interval times (Figure S1B,C, Supporting Information). We observed that interpolated cell locations showed significant deviations in concordance from the original trajectories in the both the symmetric and asymmetric cases (Figure S1B,C, Supporting Information). To account for this incongruity, we developed a pseudo-Monte Carlo (pMC) approach that applies a random perturbation defined by a pMC factor to the linearly interpolated locations, thereby estimating the location of cells at a given time point (see Experimental Section). We found that this pMC approach generated concordances with no significant difference to the trajectories generated using actual data points for both the symmetric and asymmetric conditions (Figure S1B,C, Supporting Information, bottom).

In addition to the geometric concordance, we tested how higher order motility parameters varied based on interpolation approach. We found that the pMC approach could replicate short and long time-scale MSD behaviors of the original trajectory, while linear interpolation showed deviations at shorter time scales (Figure S2A, Supporting Information). To further compare the motility properties between actual, interpolated, and pMC-estimated trajectories we computed motility parameters using the anisotropic persistent random walk (APRW) model, which corrects for the fact that cells do not always follow random walks on 2D or 3D substrates.^[19,23] Results indicated no statistical differences between actual and reconstructed trajectories using an optimized pMC-2.01 (referred to as pMC-2 moving forward) for any of the eight motility parameters assessed (Figure S2B, Supporting Information). While the pMC-2 reconstructed trajectories do not exactly map the actual/ground truth trajectories, this approach maximizes the geometric concordance, thereby recapitulating the magnitude of the higher order motility parameters. Therefore, we constructed a harmonized dataset for the 409 biological conditions with a total duration of 2.5 h and acquisition time intervals of 5 min (Tables S1–S3, Supporting Information). We selected an acquisition time interval of 5 min because this was the acquisition interval for >70% of trajectories within the dataset.

2.4. Classifying Functional Cell Behaviors at Single-Cell Resolution

With this curated dataset of cell trajectories (coordinates: x , y , t), we developed the CaMI framework to take the coordi-

nates of cell trajectories, compute cell motility parameters using the anisotropic persistence walk (APRW) model,^[16] perform k -means clustering analysis, and lastly identify and classify dynamic cell behaviors (Figure 1A). Using the 74 253 single cells, we computed eight motility parameters using the APRW model (Figure 1B, see Experimental Section). Previously, we showed that the eight motility parameters generated using the APRW model were sufficient to capture the critical motility behaviors in a cohort of primary dermal fibroblasts.^[24] Furthermore, using the 74 253 cell trajectories, we performed factor analysis to compare the communalities score of the eight APRW parameters with >30 other motility parameters computed based on Heteromotility^[25] (Figure S3A, Supporting Information). We found that the motility information captured by the eight APRW parameters (communality = 0.53) was greater than that of the >30 parameters from Heteromotility (communality = 0.39). Given this result, we chose to use the eight APRW parameters for further analysis. The APRW parameters include short and long time-scale mean squared displacements (MSD_{10} , MSD_{100}), persistence times in the primary and secondary axes of migration (P_p , P_{np}), total diffusivities and diffusivities in the primary and secondary axes of migration (D_{tot} , D_p , D_{np}), and the anisotropic index that describes the spatial persistence (ϕ).

Using unsupervised k -means clustering and calculations of the inertia and silhouette values^[26] (Figure S3B, Supporting Information), we identified 17 distinct short-duration cell motility clusters (S1 to S17). To confirm this classification, we computed a coherence factor based on the vector distances between every cell and each cluster centroid (Figure S3C, Supporting Information). Results showed the expected staircase plot, highlighting that cells within each cluster exhibited small distances (blue) to their assigned cluster centroids, and larger distances (white) to unassigned cluster centroids. Using 2D t-stochastic neighbor embedding (tSNE), we plotted all 74,253 cells and painted each cell based on the designated motility cluster ($n = 17$) (Figure 1C). We then assessed cluster groupings using unsupervised hierarchical clustering to describe gross similarities among cell motility clusters, resulting in four behavior classes: SG1, SG2, SG3, and SG4 (Figure 1D). Each cluster and corresponding behavior class showed distinct motility patterns defined by the magnitude of the eight motility parameters (Figure 1E–G; Figure S3D, Supporting Information). Moreover, to verify that the pMC interpolation approach used for harmonization did not introduce clustering artifacts that deviate from trends in the original data, we compared the higher order parameter distributions between original and interpolated trajectories for a random sample of cells from each cluster (Figure S4, Supporting Information). In general, the interpolated distributions match original distributions, suggesting

Figure 1. Identifying and classifying cell motility patterns across short durations. A) Illustration of CaMI workflow. B) Heatmap showing the magnitude of eight motility parameters per cell ($n = 74,253$ cells). Each row represents a single cell, and each column represents a motility parameter. C,D) 2D tSNE plot showing the 17 motility clusters (C) where each dot is a single cell and (D) contour plots that highlight the spatial distribution of cells/clusters that are further separated into four short-duration behavior classes within the reduced dimensionality space. E) 2D t-SNE plot showing the magnitude of the eight motility parameters per cluster region. F,G) Heatmap showing the average magnitude of the eight cell motility parameters among each of the 17 cell motility clusters and the four short duration behavioral classes (G). H) Mean squared displacements per long duration (F) motility clusters separated based on the four short-duration behavior classes (SG1–SG4). I) Probability density function profiles of the angular velocity magnitudes for the four short duration behavior classes. High circularity values (less ellipsoidal profiles) indicate that cells have a similar likelihood of moving in multiple directions. J) Visualization of single-cell trajectories per cluster, the top panel shows the origin-centered trajectories for all cells within a given cluster, bottom panel shows 16 randomly selected cells per cluster. Scalebar denotes 100 μ m.

our harmonization method does not skew the emergent motility clusters reported here.

Overall SG1-4 can be summarized based on the following gross motility behaviors, respectively: SG1: low speed and high persistence; SG2: low speed and tortuous (low persistence); SG3: high speed and high persistence; and SG4: high speed and tortuous (low persistence). We also observed that the cell motility clusters belonging to SG1 and SG2 had low mean-squared displacements (MSDs) relative to motility clusters belonging to SG3 and SG4 (Figure 1H). Simultaneously, cell motility clusters in SG1 and SG3 had a higher persistence as denoted by a lower circularity score based on the probability density functions of their angular velocity magnitudes relative to SG2 and SG4 (Figure 1I). Circular profiled of the angular velocity magnitudes indicate that cells have a similar likelihood to move in multiple directions. To visually confirm that different clusters denoted distinct cell motility patterns, we plotted the trajectories of cells classified within each cluster, highlighting overt differences in cell behaviors per motility cluster (Figure 1J).

Together, this analysis demonstrates a multivariate approach to identify and classify distinct single-cell motility patterns across multiple datasets. All clustered trajectories and higher order parameters are reported in Tables S2 and S3 (Supporting Information).

2.5. Emergent Short-Duration Motility Patterns Among Distinct Cell Types and Biological Conditions

To determine how cell-types drive emergent cell behaviors, we quantified the abundance of cell trajectories across the 17 short-duration motility clusters for baseline conditions (i.e., untreated/unmodified cells) in various cell-types and performed unsupervised hierarchical clustering based on the magnitude of the correlation (Figure 2A). Maintaining the same hierarchical clustering order by short-term trajectory, we repeated this quantification for the four short-duration behavior classes, SG1-SG4 and the eight APRW motility parameters. Results indicated that groups of cells and cell types present a range of heterogeneous motility behaviors, but certain gross cell types tended to cluster together. For example, various immune cell conditions grouped together into one cluster (3D T-cell lymphoma line LY12, germinal center B-cells, 3D Jurkat T-cells, and 3D T-cell lymphoma line Hut78), while mesenchymal/fibroblast-type cells grouped together in a separate dendrogram branch (2D and 3D fibrosarcoma, 2D human embryonic fibroblasts). Together this suggests that certain cell types (e.g., immune, mesenchymal) may share some aspects of their intrinsic patterns of movements. Furthermore, even though groups of cells across conditions were enriched for one behavioral class over another (Figure 2A center), cells were generally biased toward unique combinations of short-term trajectories within each SG behavior class (Figure 2A left), suggesting heterogeneous phenotypes across cell groups.

Together, these results indicate that changes in the overall magnitude of cell motility parameters are reflected by the redistribution of cells among motility clusters and behavior classes (Figure 2A–C). We note that while the CaMI platform harmonizes acquisition interval from different motility experiments, the current algorithm does not account for possible differences

in culture conditions (i.e. local cell density, specific cell type, etc.) between experiments that may impact specific motility behaviors. Therefore, similarities and differences between cell types reported here may still be confounded by these variables. As such, in subsequent sections we will directly apply the CaMI framework to profile cell motility patterns in specific test cases to highlight the biological insights that can be achieved relative to conventional cell motility analyses. Within each of these test cases, culture conditions between biological conditions were similar and enable us to make more direct conclusions about how certain perturbations impact motility behavior.

2.6. Test Case 1: Effect of Cell Density on Emergent Motility Patterns of Breast Adenocarcinoma Cells

To benchmark the ability of CaMI to yield biological insights, we first asked how this framework can elucidate emergent differences in cell motility in response to distinct environmental conditions, such as cell density. A previous study showed that MDA231 breast adenocarcinoma cells cultured at high density exhibit enhanced cell speed. Moreover, blocking IL-8 receptor signaling using the drug Reparixin can reverse this effect, suggesting that enhanced cell motility at high density conditions is due to paracrine IL8 signaling. Here, we were curious as to whether CaMI could further elucidate how Reparixin treatment impacts emergent single-cell motility patterns, that is collections of several bulk motility properties, beyond just displacement.

To test this hypothesis, we first extracted higher order motility parameters for control and 50nM Reparixin-treated cells cultured at low density (LD) and high density (HD) (Figure 2B). We found that on average, HD cells not only had greater displacement, but higher persistence time and diffusivity in the primary axis than their LD counterparts. When treated with 50nM Reparixin, the motility parameters shift such that both LD and HD cells have similarly low displacement, diffusivity, and persistence times. We then projected these cells onto the CaMI space (Figure 2C,D) and found that baseline LD and HD cells existed in distinct motility states of low- and high- speed. Within both density conditions, a subset of cells occupied the low displacement behavioral classes (SG1/2), while another subset occupied the high displacement behavioral classes (SG3/4). The baseline HD condition simply has a higher proportion of cells within the high-motility behavioral classes. This result shows that CaMI provides additional information about the heterogeneity of motility states within each cell condition. Instead of all single cells having HD in high-density conditions, CaMI enables us to see that different subsets of cells take on distinct behaviors. Treatment with Reparixin consistently skewed LD and HD cells toward high enrichment of SG1 motility patterns 6 and 7. Moreover, we quantified the Shannon entropy for each of these conditions and found that baseline conditions had higher entropy than those treated with Reparixin. This result indicates that Reparixin treatment also reduces the heterogeneity of motility states. Correlating the abundance of LD and HD cell types at baseline and in response to Reparixin treatment (Figure 2E) shows that these density conditions have a stronger correlation in response to Reparixin exposure (Spearman R = 0.83, p-value < 0.001) than at baseline (Spearman R = 0.44, p-value = 0.07). These results can also be

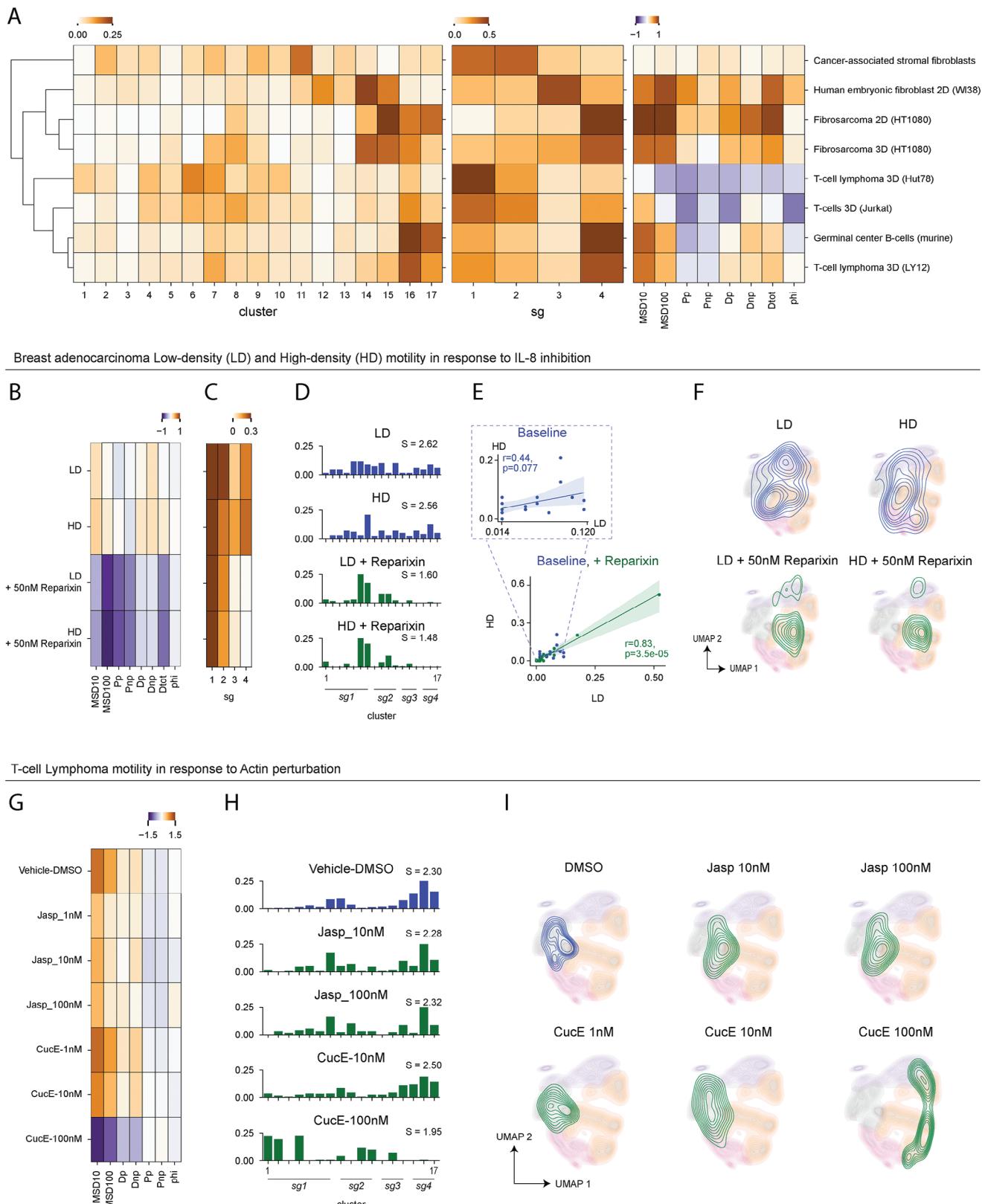


Figure 2. Application of CaMI workflow to short-duration datasets. A) Heatmaps indicating the fractional abundance of distinct cell types in our dataset across the seventeen short-duration motility clusters (left), four short-duration behavior classes (center), and the average magnitude of the eight motility parameters within each cell group (right). B) Heatmap of scaled higher order motility parameters for low-density (LD) and high-density (HD) MDA231 cells at baseline in response to 50nM Reparinix exposure. C) Barplots showing the fractional abundance of MDA231 cells within each of the seventeen

visually demonstrated by projecting each cell type condition onto the CaMI short-duration tSNE space (Figure 2F).

Together, these results demonstrate how CaMI can confirm trends observed through bulk motility parameter calculation while providing additional information about the heterogeneity of motility states in response to environmental and molecular perturbations.

2.7. Test Case 2: Effect of Cytoskeletal Modulators on T-lymphoma Motility Within 3D Microenvironments

Next, we wondered whether CaMI can quantitatively demonstrate how perturbing cellular mechanisms that directly drive cell movement, such as the cytoskeletal network, shift cell motility behaviors. To explore this question, we applied CaMI to test the effect of actin and microtubule modulation in T-lymphoma cells (LY12) within 3D collagen-I microenvironments. To test the effects of actin and microtubule modulation on motility behavior, we exposed cells to escalating concentrations of four drugs: Jasplakinolide (Jasp), an actin polymerizer, Curcubitacin E (CucE), an actin depolymerizer (Figure 2F,G), Taxol, a microtubule stabilizer, and Vincristine (Vinc), a microtubule depolymerizer (Figure S5A–C, Supporting Information).

First, we quantified the bulk motility parameters for each cell type condition (Figure 2G; Figure S5A, Supporting Information). At baseline, LY12 cells have high displacement and low persistence times. We observed that increasing Jasp concentration slightly decreased cell motility, as noted by the slight drops in motility parameters like Dtot and MSD10 (Figure 2G). However, increasing concentrations of CucE appeared to have a much more significant decrease in cell displacement, as indicated by the stronger drops in the values of these motility parameters (Figure 2G). Using the CaMI workflow, we observed that baseline LY12 cells occupied SG1 cluster 6/7 or behavior class SG4 clusters 14–17 with a stronger enrichment toward SG4 (Figure 2H). Treatment with increasing concentrations of Jasp appeared to slightly increase enrichment in SG1 clusters, while decreasing that in SG4. LY12 cells treated with increasing concentrations of CucE, on the other hand, had a much stronger response with cells being depleted from the SG4 behavior and enriched in either SG1 or SG2. Here, CaMI provides additional information that bulk motility parameter analysis by showing that the drop in displacement parameters in response to 100nM CucE treatment was reflected by a divergence toward two distinct cell motility states. The Shannon entropy for the cells treated with 100nM CucE was also much lower than at baseline or in response to Jasp, suggesting that CucE reduces the heterogeneity in cell motility as well. These trends are also visually reflected in the short-duration motility tSNE space (Figure 2I) where we observe the shift of cells treated

with 100nM CucE from the left to right of the low-dimensional space.

Interestingly, exposures to microtubule stabilizer Taxol led to a similar response as exposure to Jasp, wherein cells experience a slight decrease in displacement and enrichment in SG1 motility patterns (Figure S5A–C, Supporting Information). Conversely, treatment with Vincristine led to a strong increase in cell displacement from baseline and greater enrichment in SG4 motility patterns (Figure S5A–C, Supporting Information). These patterns were visually observed by the contour of cells treated with Taxol slightly expanding in the low-dimensional motility space, while contour of cells treated with Vincristine progressively shrinks within SG4 (Figure S5C, Supporting Information).

Together, these results indicate how the CaMI framework can quantitatively demonstrate cellular heterogeneity that are not directly apparent using conventional cell motility analyses. Particularly, CaMI can elucidate instances where change in a bulk motility parameter translates to enrichment of two distinct motility states. Treatment with 100nM CucE decreased motility parameters such as MSD10 and MSD100, but this decrease translated to enrichment of two distinct behavioral classes—SG1 and SG2. This result means that the drug causes cells to take on both low displacement, high persistence, and low displacement, low persistence motility behaviors.

2.8. Using CaMI to Predict Cell Motility Clusters in Previously Unclassified Cells

In addition to using CaMI to decipher how molecular and environmental perturbations drive fractional redistributions across motility states, we wondered whether we could project new data into our dataset to quantify behaviors for previously unclassified cells. To perform this experiment, we trained a linear support vector machine learning model (90:10 train-test split) to predict short-duration motility pattern (S1–S17) based on the underlying eight APRW parameters of individual trajectories. Our trained linear SVM had an average training accuracy of 98.7% across all clusters (Figure S6A, Supporting Information). To validate the model, we simulated 75 000 new trajectories using a model developed by Wu et al.^[16] With these new trajectories, we computed the eight motility parameters and performed the classification using the trained 90:10 model, which yielded an average validation accuracy of 96.7% across all clusters (Figure S6B, Supporting Information). Here, we validated the true motility cluster of each simulated cell by performing an iterative analysis to re-classify all the cells adding one new simulated/unclassified cell each time (i.e., re-training on 74 253 cells + 1 simulated cell; 75 000 times). Having demonstrated a high accuracy of the trained model, we tested the model using cell trajectories of primary monocytes moving within 2 mg mL⁻¹ collagen-I gels.

short-duration motility clusters by condition. D) Linear regression plots indicating the correlation between LD and HD fractional abundance across the seventeen short-duration clusters at baseline and in response to Reparixin exposure. R and p indicate results from a Spearman correlation test. E) 2D tSNE map with the four behavior classes shown as shadings and the contours of each MDA231 cell condition. F) Heatmap showing the magnitudes of the eight motility parameters for the LY12 T-cell lymphoma cells treated with Actin modulators (Jasplakinolide (Jasp.) and Curcubitacin E (CucE)). G) Bar plots showing the fractional abundance of cells per short-duration motility clusters treated with Actin modulators. H) Contour maps showing the distributions and localizations of cells at baseline (DMSO) and treated with varying concentrations of Jasp and CucE.

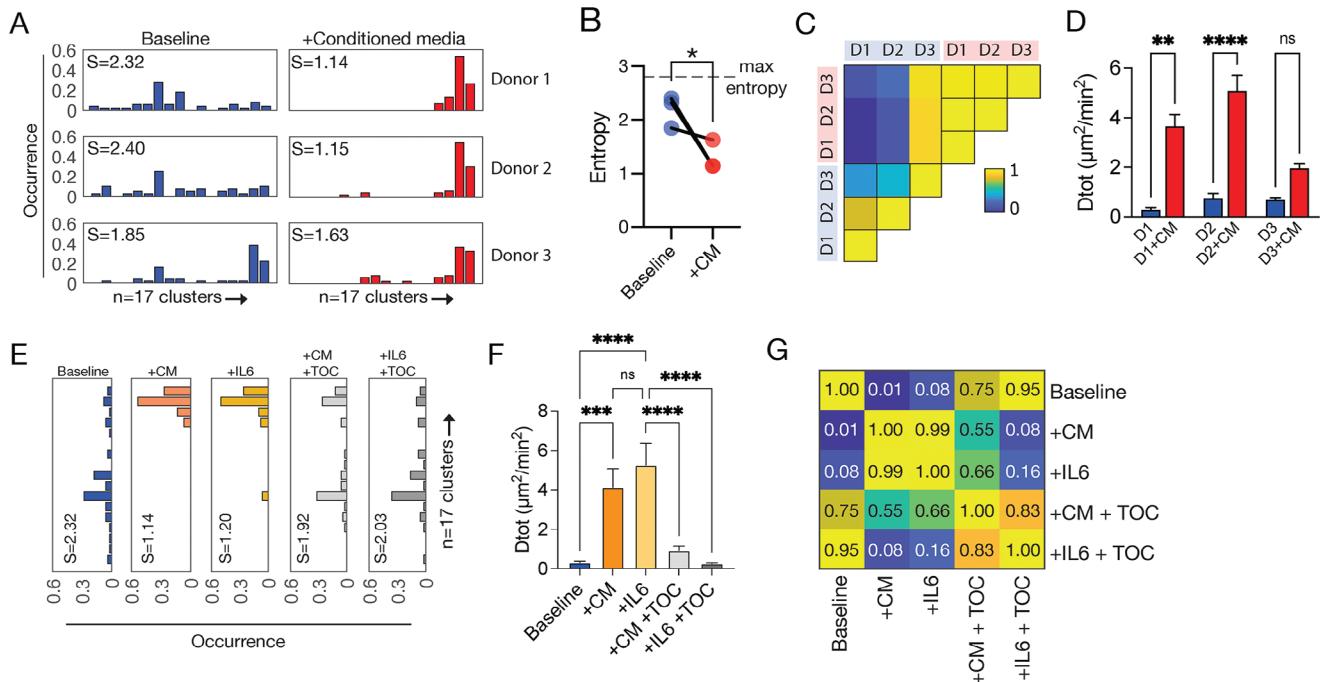


Figure 3. Quantification of cell motility patterns in primary monocytes. A) Bar plots showing the abundance of cells among the 17 cell motility clusters for three donors at baseline and after exposure to conditioned media. B) Scatter plot showing the decrease in cellular heterogeneity after CM-exposure. C) Correlation map of cell motility clusters among donor samples. D) Bar plots showing the magnitude of the total diffusivities at baseline and after CM-exposure. E) Bar plots showing the distributions of cell motility clusters across testing conditions. F) Bar plot showing the total diffusivities across testing conditions. G) Correlation map indicating similarities in the cell motility clusters among testing conditions.

2.9. Test Case 3: Fractional Redistributions Among Cell Motility Clusters Define Monocytes Motility

Having trained a model to now project new cell measurements onto the CaMI motility space, we were curious as to how CaMI can quantitatively assess cell motility properties in primary human cells. To explore this question, we specifically focused on primary human monocytes because these cells are known to rely on motility to orchestrate immune function in response to local microenvironmental cues and signals. Moreover, prior studies have shown the aging and cancer can disrupt monocyte ability to perform these critical functions, suggesting that having a tool to assess how monocytes navigate cellular microenvironments could lead to novel solutions to influence patient health and disease potential. With the trained linear SVM model, we quantified the distributions of short-duration motility patterns for primary monocytes that were not included in the initial CaMI workflow and dataset. Briefly, we collected peripheral blood mononuclear cells (PBMCs) from three healthy donors (D1, D2, D3) and isolated monocyte populations based on CD14+ expression. We then embedded monocytes from each donor into 2 mg mL⁻¹ collagen-I gels, with eight technical replicates (i.e., wells in a 96 well plate) per donor. In half of the wells per donor (4 wells each), we exposed monocytes to conditioned media from breast cancer cells (MDA-MB 231) for \approx 12 h prior to imaging. Using 3D time-lapse microscopy, we imaged and analyzed the monocytes moving in 3D collagen-I gels (2 mg mL⁻¹) for 2.5 h at 5-min intervals then applied the aforementioned linear SVM to classify the motility pattern of each cell.

We observed that each donor exhibited unique baseline profiles of short-duration motility patterns (Figure 3A). Donor 3 exhibited high baseline abundances of cells in the SG4 behavior class (specifically in motility clusters S14, S15, S16, S17), while donors 1 and 2 exhibited higher abundances of cells in SG2 behavior class. Furthermore, when monocytes were exposed to conditioned media (+CM), donors 1 and 2 exhibited similar shifts toward enrichments in behavior class SG4, specifically motility clusters S14, S15, S16, S17, corresponding to high displacement and low persistence motility behaviors (Figure 3A). These shifts in cell motility patterns after CM exposure were also associated with decreased cellular heterogeneity as determined by the Shannon entropy (Figure 3B) and a higher correlation seen among cells exposed to CM relative to their baselines (Figure 3C). Interestingly, donor 3 did not exhibit drastic changes in the distribution among cell motility clusters after CM exposure, with no significant observed in the baseline and CM treated cells based on the total diffusivity (i.e., effective speed) (Figure 3D). Together, the results confirm that monocytes exposed to CM-associated signals increase their movements, with the lack of response in donor 3 suggesting a resistance or insensitivity to CM-associated signals. While the magnitude of the response can be determined using conventional cell motility analysis, the addition of CaMI analysis provided a direct visualization of the responses and demonstrates that the ensemble response itself (increased cell motility) is due to a fractional redistribution among cell motility clusters.

To decipher these responses further, we asked whether specific molecules within the conditioned media were capable of driving

similar shifts in the motility behavior. Since cells from donor 1 exhibited drastic shifts in response to CM exposure, we selected that sample to performed additional experiments. Recent studies on cancer cell conditioned media indicates that interleukin-6 (IL6) is highly expressed. Furthermore, other studies have shown that IL6 exposure increases cell migration in monocytes.^[27] To test the effects of IL6 exposure on monocyte motility patterns, we exposed cells from donor 1 to either CM or 500ng mL⁻¹ of recombinant human IL6 and profiled cell motility patterns. Similar described above, cells exposed to both CM and IL6 exhibited a characteristic shift in the abundance of cells among cell motility clusters to enrich for cells in behavior class SG4 (specifically S14, S15, S16, and S17), with a corresponding decrease in cellular heterogeneity and higher total diffusivities/effective speeds (Figure 3E,F). Lastly, to test whether IL6 was directly modulating the phenotype, we co-treated cells with either conditioned media or IL6 and 100μl IL6-receptor blocking antibody Tocilizumab. Results indicated that monocytes treated with Tocilizumab regained a baseline-like profile denoted by the shifts in the distributions among cell motility clusters (Figure 3E), a decrease in total diffusivity, and an increase in the magnitude of the cellular heterogeneity to baseline levels (Figure 3F). Furthermore, correlation analysis on the distributions of cells across cell motility clusters among conditions showed that cells exposed to CM and IL6 were highly correlated, with baseline monocytes showing strong correlation with cells co-treated with CM or IL6 and Tocilizumab (Figure 3G).

Altogether, we show that using the CaMI framework, we can directly quantify dynamic cell behaviors, and cellular heterogeneities to profile emergent responses of cells using cell motility. Furthermore, results indicate that all the biological insights gained from conventional motility analyses are conserved, with new insights gained regarding heterogeneity, fractional redistributions, and the direct visualization of the cell motility trends at baseline and after perturbations.

2.10. Quantification of Dynamic Cell Behaviors over Long Durations

In many of the conditions used, more than 8 h of continuous cell tracking data was available. We wondered what long-duration motility patterns existed within this motility database, and how these long durations motility patterns related to the short-duration cell motility patterns (2.5 h). We rationalized that this could be informative, especially since current approaches to analyze cells motility relies on the assumption of time invariance. To evaluate these questions, we compiled cells with 8 h of motility data at 5-min intervals and computed the same eight motility parameters described in the short-duration analyses. With this data ($n = 24\,753$ cells) (Figure 4A), we performed k-means clustering analysis to identify 25 clusters (Figure 4B). We verified cluster number using the inertia and silhouette values and coherence staircase plot (Figure S7A,B, Supporting Information). These 25 motility clusters (C1 to C25) also separated into four behavior classes (CG1, CG2, CG3, CG4) based on unsupervised hierarchical clustering using correlation distances (Figure 4C). Each long duration motility pattern and behavior

class displayed distinct patterns in the magnitudes of the eight motility parameters (Figure 4D,E; Figure S7C-E, Supporting Information). Plotting the trajectories of cells within each of the long-duration motility clusters revealed not only distinct patterns of movements (Figure 4F) but also a polarization of highly motile cells defined by MSD100 in motility clusters belonging to behavior classes CG3 and CG4 (Figure S7C, Supporting Information) and higher persistence for cells in behavior classes CG2 and CG3 (Figure S7D, Supporting Information).

We next evaluated the ability to predict the motility pattern of unclassified cells. Using the same approach described for the short-duration motility clusters. Similarly described above, we trained a new linear SVM model on the 8-h cell tracking data, tested the prediction in a 90:10 split dataset, and newly simulated trajectories for 25 000 single cells. Results showed a 97.2% accuracy for the 90:10 split dataset (Figure S8A, Supporting Information) and a 95.6% accuracy for the simulated trajectories (Figure S8B, Supporting Information). With this tool, we can also identify and classify distinct cell long-duration motility patterns for cells not used to train the CaMI platform (clustered long duration trajectories and higher order parameters reported in Tables S4–S6, Supporting Information).

2.11. Connecting Motility Patterns Across Short and Long Durations

In most cell motility analyses, the inherent assumption of time-invariant movement or time independence limits the ability to infer temporal stability of cell behaviors. To infer temporal patterns, we coupled short and long-duration motility clusters within the same cells to determine temporal transitions between short-term motility behaviors, thereby inferring overall long-term temporal patterns of motility behaviors. For a subset of cells having 8 h of movement, we broke each 8-h trajectory into three non-overlapping 2.5-h segments. We determined motility clusters (S1-17 and C1-25) and complimentary behavior classes (SG1-SG4 and CG1-CG4) for all short-duration segments as well as the long-duration trajectories. With this data, we then asked whether a cell classified in a particular motility cluster or behavior class for the first 2.5 h stayed within the same short-term behavioral class for subsequent 2.5-h time segments. Overall, we observed that for the $\approx 25\,000$ cells assessed, cells classified within classes SG1 and SG4 tended to stay within the same behavior class for the three consecutive time segments (Figure 4G), with higher temporal stabilities across the time segments (Figure 4H). Conversely, SG2 and SG3 tended to shift behavior classes over time and had relatively lower temporal stability.

Taken together, we describe the ability to quantify the temporal dynamics of the motility patterns at short and long durations. This quantification of temporal dynamics is critical, as it provides a robust approach to evaluate temporal cell motility patterns without the direct limitations of the time invariance assumption. By separating cell trajectories into shorter time segments, where the time-invariance assumption is valid, we can decipher spatio-temporal patterns that accurately describe single-cell behaviors.

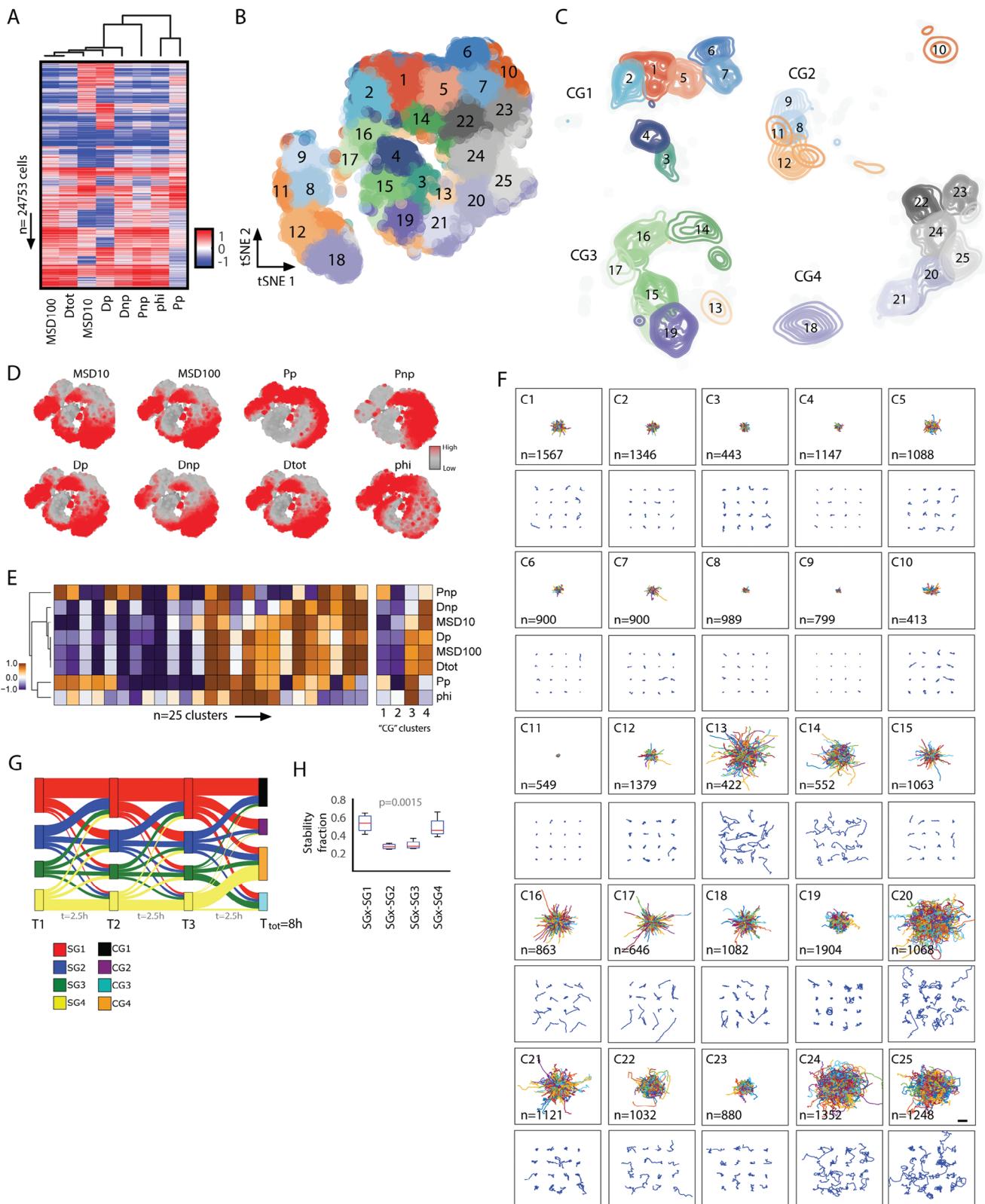


Figure 4. Identifying and classifying cell motility patterns across long durations. A) Heatmap showing eight motility parameters per cell ($n = 24,733$ cells). Each row represents a single cell, and each column represents a motility parameter. B,C) 2D t-SNE scatter plot where each dot is a cell and colors indicate the 25 identified long-duration cell motility clusters (B) and contour plots further separated into four long-duration behavior classes (CG1-4) that show the distributions and localization of cells per motility cluster (C). D) 2D t-SNE plot showing the magnitude of the eight motility parameters per cluster region. E) Heat map showing the magnitude of the eight cell motility parameters for the twenty-five long-duration cell motility clusters (left)

2.12. Using CaMI to Evaluate Functional Heterogeneity Across Conditions

Given the ability to profile dynamic behaviors of single cells using CaMI, we further tested the ability to derive key biological insights that were not directly apparent using conventional motility analyses based on average motility parameters alone. Using a subset of data from long-duration movies (i.e., 8 h), we performed hierarchical clustering based on cross-correlation analysis across the 409 conditions. This analysis yielded four groups, as shown in the cross-correlation heatmap (Figure 5A) and 2D tSNE plots (Figure 5B). As a biological test case, we took data from LY12 cells (T-cell lymphoma cells) moving within 3D collagen-1 gels across five increasing collagen concentrations. Plotting the trajectories of cells within each condition, we qualitatively observed a bi-phasic response in the overall magnitude of the cell displacements, with the lowest displacements observed at 3 mg mL^{-1} collagen (Figure 5C). Plotting this condition in the 2D t-SNE space, we quantitatively observed a bi-phasic or switching effect (cells in 4 mg mL^{-1} are more similar to cells in 1 mg mL^{-1} than to cells in 3 mg mL^{-1}) (Figure 5D). This switching phenotype with changing collagen concentrations is likely due to local changes in molecular crowding effects and effects on collagen alignment and porosity.^[28,29] Furthermore, plotting the abundance of cells within the 25 clusters for each condition, we also observed a bi-phasic response in the heterogeneity (defined based on the Shannon entropy), with a minimization at 3 mg mL^{-1} (Figure 5E).

Given these striking spatial associations with collagen concentration, we asked whether complementary temporal associations were based on short and long durations. To address this, we stacked the three 2.5-h time segments, paying attention to the fraction of cells within each short-duration behavior class (SG1 to SG4) and the frequency at which cells switched classes with increasing concentrations. From this analysis, we also observed a switching phenotype based on temporal associations, with the majority of cells staying in SG4 (yellow) across the three segments for 0.5, 1, and 4 mg mL^{-1} , an increase toward SG1 for 2 mg mL^{-1} , and a shift toward SG1 as the dominant behavior class for 3 mg mL^{-1} (Figure 5F). Furthermore, while this phenotype switch is observed spatially at 3 mg mL^{-1} , from the temporal perspective, this shift is already apparent at 2 mg mL^{-1} and is enhanced at 3 mg mL^{-1} .

Altogether, CaMI enables the quantification of real spatial-temporal dynamics and heterogeneities that define bi-phasic and phenotype-switching responses to increasing collagen concentration.

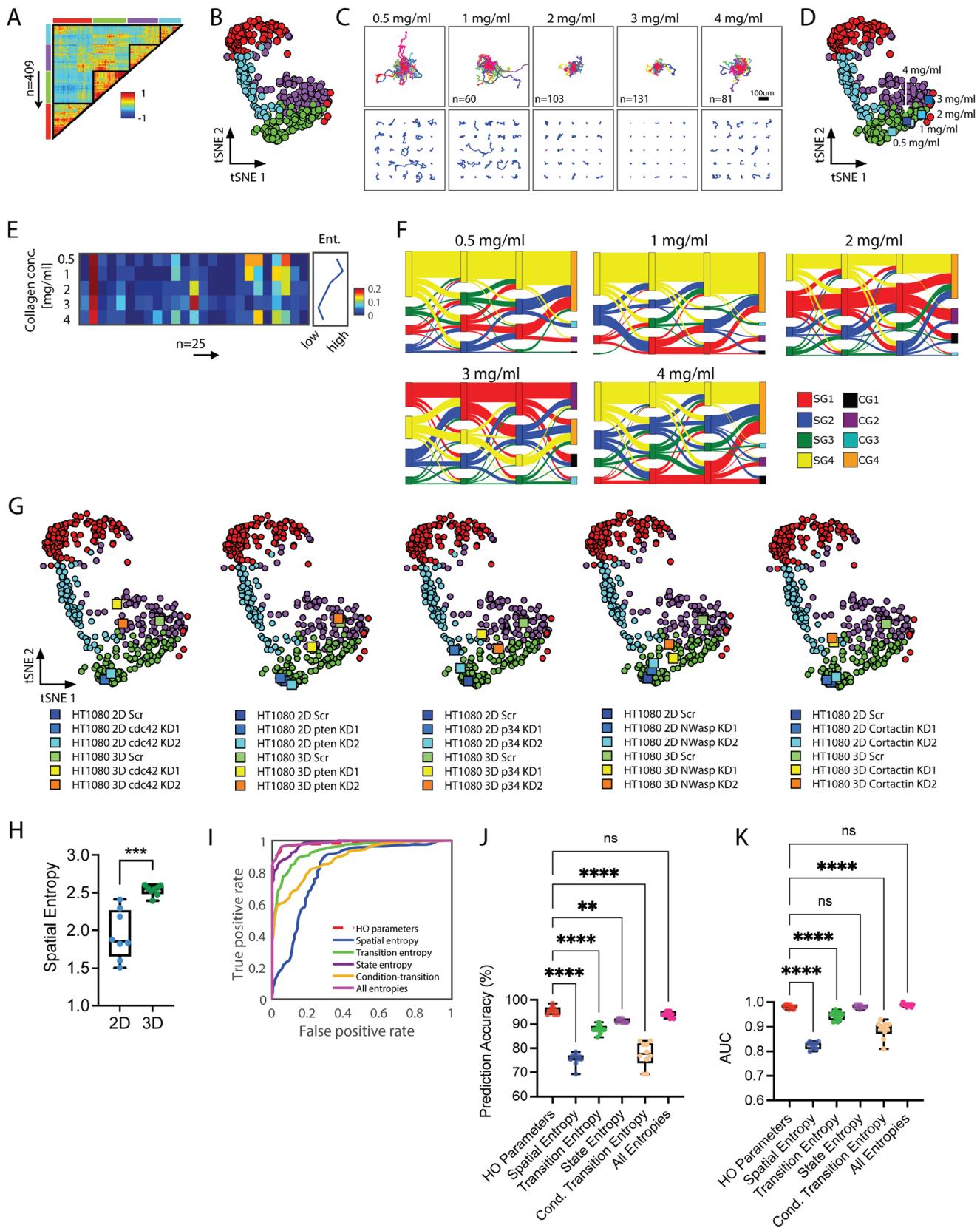
2.13. Predicting Dimensionality of Cellular Conditions Based on Spatial and Temporal Heterogeneities

Given the ability to dynamically map spatio-temporal cell behaviors, we wondered whether we could predict the dimensionality

of cells moving in 2D and 3D microenvironments. Recent studies have demonstrated that cell movements in 2D do not necessarily recapitulate 3D movements,^[19,30,31] which is especially applicable in studies involving the characterization of metastasis and cancer cell invasion. Furthermore, it is unclear why there is such a disconnect in motility behaviors in 2D and 3D microenvironments, where 2D cancer cell migration is not predictive of cancer cell invasion. We evaluated cell motility patterns in 60 matched 2D and 3D conditions to elucidate what defines this difference in 2D versus 3D. These conditions were comprised of ht1080 cells under multiple genetic and drug-induced perturbations, including knockdowns using different shRNAs and concentration-dependent drug treatments. This wide range of perturbations allows direct interrogation of the universe of movement patterns exhibited by cells moving in 2D and 3D microenvironments. Mapping the 2D and 3D conditions in t-SNE space, we observed differential localization across regions (Figure 5G), indicating inherent differences in the motility patterns of the cells, with cells in 3D microenvironments exhibiting higher heterogeneities (Figure 5H). This differential localization within the 2D t-SNE space was observed in knockdown conditions and as a function of drug concentrations in ht1080 cells and other cell types (Figure S9, Supporting Information). Although we did not observe dimension-exclusive motility clusters (meaning that both 2D and 3D cellular conditions had cells classified in each of the 25 long-duration motility clusters), the fraction of cells within each cluster varied greatly between 2D and 3D.

Prompted by these results, we asked whether we could predict the dimensionality of moving cells based on the magnitude of the heterogeneity across the cell populations since cells in 3D tended to exhibit higher phenotypic heterogeneities relative to matched 2D conditions (Figure 5H). To comprehensively define the heterogeneity, we further defined four types of entropies that captured cells' spatial and temporal nature (see Experimental Section). These different types of entropy were a) spatial entropy, which is computed based on the fractional abundance of cells within each short or long-duration motility cluster ($n = 17$ or $n = 25$), b) state entropy, computed based on the fractional abundance of cells within short and long duration behavior classes (SG1-SG4, CG1-CG4), c) transition entropy, which is computed based on the probability of transitioning from one short duration behavior class to the another along subsequent time segments, and d) conditional-transition entropy, estimates the entropy based on the conditional probabilities that define the likelihood of transitioning to a specific short duration behavior class. We then trained a model to predict the dimensionality of cells using a linear SVM for each entropy and the combination of entropies (see Experimental Section). Comparing the prediction accuracies using either the eight motility parameters or the entropies, we found that the combination of entropies was able to predict the dimensionality in >95% of the cases, which was comparable to using the eight motility parameters to predict the dimensionality (Figure 5I-K).

and within each of the four long-duration behavior classes (right). F) Visualization of cell trajectories for cells classified within each cluster, the top panel shows the origin-centered trajectories for all cells within a cluster, and the bottom panel shows 16 randomly selected cells per cluster. Scalebar denotes $100\mu\text{m}$. G) Sankey plot showing the transitions based on the fractional abundance of cells within each short-duration behavior class for three-time segments (2.5 h each) and how they ultimately connect to the four long-duration behavior classes. H) Box-and-whisker plots of the stability fraction of cells within each short-duration behavior class.



Taken altogether, we show the ability of CaMI to predict patterns of cell movements and derive biological insights across cellular conditions, such as how the dimensionality of the microenvironment influences dynamic behaviors. Furthermore, understanding phenotypic heterogeneity (both spatial and temporal) is critical to interpreting single-cell and population-level cell behaviors.

3. Discussion

Cell movements play a critical role in the physiology of organisms, both in homeostasis and disease.^[32] CaMI is a computational framework that takes advantage of the single-cell nature and spatio-temporal dynamics of cell motility data to quantitatively extract distinct motility patterns across multiple conditions. Conventionally, cell motility is quantified using small numbers of cells and reported as bulk averages of cell populations. Moreover, cell motility experiments are often conducted using inconsistent imaging conditions and/or spatiotemporal resolutions, preventing the unification of several motility experiments for broader analyses. CaMI harmonizes motility data from distinct resolution to a consistent temporal resolution, enabling higher order parameter extraction and unsupervised clustering of motility data from a wide range of sources. Here, we have demonstrated the utility of CaMI to quantify cell motility patterns, determine spatial and temporal heterogeneities, define dynamic phenotypic transitions, and predict the dimensionality of cellular conditions. In addition, CaMI provides a direct visualization scheme of single-cell behaviors to enable direct interpretation of motility data while retaining all the biological information gained from conventional bulk analyses. Furthermore, CaMI can classify dynamic cell behaviors from imaging data extracted from either *in vitro* and *in vivo* settings or response to genetic or chemical perturbation, for example, drugs and shRNAs.

While CaMI provides a robust solution to decipher emergent patterns of single-cell behaviors, there are certain contexts where it may be limited. A fundamental limitation of CaMI is that dominant cell motility patterns would be identified as clusters since cell motility clusters are derived based on multivariate statistical approaches of combined motility datasets (i.e., spanning multiple cell types, dimensionalities, and perturbations). As a result, rare, cell-type-dependent, or perturbation-dependent cell motility patterns may not be identified as distinct but will be classified within larger, dominant clusters. While this is the case in most big-data classifications, CaMI can provide reliable and insightful informa-

tion, as demonstrated by the high cluster coherence for cell-type-dependent or perturbation-dependent conditions by retraining the motility clusters using an appropriate dataset. In conditions where cell type or condition-specificity is critical. One way around this limitation is to train the classification model based on a specific cell type or condition. For instance, we have done this for dermal fibroblasts that resulted in eight total motility clusters.^[7] However, the size of the dataset (*i.e.*, the number of cells) must be large enough to be robust and reproducible. Lastly, CaMI is designed as a comparative tool, where cellular conditions are differentiated based on the fractional redistribution among defined cell patterns. Multiple comparative conditions are needed to derive deep biological insights and understand how a given perturbation influences the resultant cell behaviors based on the abundance of cells within each motility cluster or behavior class.

Additionally, the test cases presented in this study came largely from *in vitro* experiments of cells moving on 2D substrates or within 3D hydrogels. While we only looked at (x, y, t) coordinates, the framework can be expanded to look at true 3D movements (x, y, z, t) in addition to 2D movements of cells in a 3D hydrogel. Lastly, the CaMI framework can be used on cells moving *in vivo* based on movies from intra-vital imaging or cells moving in *ex vivo* organotypic cultures (viable sections of thick tissues that can be mounted onto a microscope). In short, as long as the coordinates (x, y, t) or (x, y, z, t) can be generated, the CaMI workflow can be applied and tested, the key consideration is the dataset of cell trajectories that are used to identify the cell motility clusters.

Despite these limitations, we still believe CaMI is a widely applicable tool to benchmark heterogeneity in cell motility behaviors in various settings.^[3,24] With emerging literature on cell motility as a clinical biomarker of overall health, the CaMI framework can be used in clinical settings to benchmark patient cell motility behaviors and associate distinct cell motility states with overall health outcomes. For instance, in evaluating and engineering designer cellular traits for next generation CAR-T cell technologies^[33] or evaluating and modulating the potential for cancer cells to metastasize.^[13]

Altogether, we present a robust framework to identify and classify dynamic single-cell behaviors across multiple conditions. We anticipate that increased implementation of modern single-cell approaches and data science analyses will lead to a better understanding of dynamic cellular behaviors and heterogeneity. While CaMI provides a robust approach to quantify functional cell behaviors, additional work is needed to better understand phenotypic transitions and mechanistic associations of cell behaviors to

Figure 5. Quantifying functional heterogeneity and biological insights using CaMI. A) Cross-correlation heatmap highlighting the grouping of the 409 biological conditions based on similarities in overall cell behaviors. B) 2D t-SNE plot showing the 409 conditions painted based on the identified groupings. Each dot denotes a single biological condition ($n = 409$). C) Single-cell trajectories of LY12 cells as a function of increasing collagen concentrations. The top panel shows origin-centered trajectories for all cells analyzed, and the bottom panel shows 25 randomly selected cells per condition. D) 2D t-SNE plot shows a bi-phasic/switching effect with increasing collagen concentration. LY12 cells moving in 4 mg mL^{-1} gels are more similar to LY12 cells moving in 1 mg mL^{-1} gels relative to cells moving in 3 mg mL^{-1} collagen-I gels. E) Heatmap showing the abundance of cells within each cluster per collagen concentration. The line plot on the right shows a bi-phasic change in the Shannon entropy, denoting the lowest cellular heterogeneity for cells moving in 3 mg mL^{-1} collagen-I gels. F) Sankey plots show the temporal transitions of cells across short-duration classes as a function of increasing concentration. Sankey plots also indicate a switch in temporal kinetics and temporal heterogeneity. G) 2D t-SNE plot showing the location of matched 2D and 3D conditions of ht1080 cells. H) Box plot showing the Shannon entropy of cell moving in 2D microenvironments are significantly lower than for cells moving in 3D microenvironments. I) ROC curve showing prediction accuracy, denoted by the true and false positive rates, for higher order motility parameters and different spatio-temporal entropies. J,K) Box-and-whisker plots show the prediction accuracy (J) and area under the curve (K) for higher-order motility parameters and various spatio-temporal entropies. Data indicates that combining entropies yields the same prediction accuracy as using the higher-order parameters.

their underlying molecular states. Recently, studies have started to multiplex dynamic properties of cells with underlying genetic and epigenetic states of cells^[34–36]. However, more studies are needed to comprehensively assess and establish deterministic links between cell motility patterns and behaviors, cell state transitions, cell fates, and underlying molecular modifications (e.g., cells with knockdowns in particular genes). Ultimately, multiplexed *in situ* quantification of molecular and phenotypic behaviors quantified simultaneously could enable precision health approaches where dynamic cell behaviors connect to disease states. This link will be critical in light of heterogeneous cellular behaviors within cell populations and the associations with clinical outcomes and treatment responses.^[37,38]

4. Experimental Section

Data Curation: Previously published data and newly acquired data were compiled to generate an *in vitro* single-cell motility database for the following analyses.^[7,17–22] All motility data were derived from time-lapse movies collected using brightfield or phase microscopy. Single-cell trajectories were extracted using commercial cell tracking software (Metamorph). Cell motility data were compiled from several different sources, with movies ranging in total duration from 2.5 to 16 h and acquisition intervals from 2 to 10 min. All trajectories were collected in 2 dimensions (x, y); the database contains experiments of cells moving both within 2D substrates or 3D scaffolds, but the analyses here only consider movement in two dimensions for both cases. The entire database comprises of 74 253 single cells across 409 unique biological conditions. Table S1 (Supporting Information) contains the total duration, interval time, pixel size for calibration, and experimental descriptions for each biological condition. To render the individual single-cell trajectories comparable, statistical corrections (described below as pseudo-Monte Carlo) were applied to generate single-cell trajectories with consistent acquisition intervals of 5 min. From these videos, short-duration trajectories of 2.5 h and long-duration trajectories of 8-h duration were generated. To generate these trajectories, longer videos were cut into non-overlapping shorter sections meaning that multiple trajectories in the training data could pertain to the same cell. Seventy-four thousand two hundred fifty-three short-duration, 2.5-h trajectories (Table S2, Supporting Information) and 24 753 long-duration, 8-h trajectories (Table S4, Supporting Information) were harmonized.

Assessing Geometric Concordance among Varied Acquisition Time Intervals: To understand how acquisition time interval impacts trajectory shape, ht1080 fibrosarcoma cells moving within 3D collagen-1 gels were selected because this dataset was initially acquired at the highest temporal resolution of 2 min. Degraded time-resolution trajectories were generated *in-silico* for each original trajectory with acquisition times ranging from 2.5 to 10 min in 0.5-min increments. The concordance was measured between trajectories with varied acquisition times. The concordance (Figure S1A, Supporting Information) was quantified as the summed triangular area formed between points acquired at consecutive timesteps.

Pseudo-Monte-Carlo Technique to Predict x,y Locations of Cells at a Given Time Point: First, the required (or missing) X-Y location from linear interpolation ($x_{\text{interp},i}$, $y_{\text{interp},i}$) was estimated. Then, a small perturbation was employed to the interpolated point according to the equations below:

$$x_{pMC,i} = x_{\text{interp},i} + \delta_{x,i} \quad (1)$$

$$y_{pMC,i} = y_{\text{interp},i} + \delta_{y,i} \quad (2)$$

$$\delta_{x,i} = \frac{(x_{i+1} - x_{i-1})}{pMCF} (|x_{i+1} - x_{i-1}| \geq |y_{i+1} - y_{i-1}|) \quad (3)$$

$$\delta_{y,i} = \frac{(y_{i+1} - y_{i-1})}{pMCF} (|x_{i+1} - x_{i-1}| \leq |y_{i+1} - y_{i-1}|) \quad (4)$$

The perturbation is either applied to the dimension that experiences more displacement in a given interval. For example, if between two time steps the cell moves more horizontally than vertically, the perturbation is applied in the X dimension and no perturbation is applied in the Y dimension. A perturbation is applied in both dimensions when the cell moves perfectly horizontally. *pMC factor* values between 2 and 2.03 (Figure S1C, Supporting Information) were tested and statistically compared resultant higher order parameters using a Kruskal-Wallis test followed with the Dunn test for multiple comparisons testing. A *pMC* value of 2.01 was selected because this yielded no statistically significant difference from the original trajectory for all higher order parameters. A *pMC* factor of 2.0 was also sufficient but there were certain situations where *pMC* factor 2.01 performed better.

Calculation of Higher Order Parameters Using Anisotropic Persistence Random Walk (APRW): To quantify the spatio-temporal motility patterns, bulk cell movements were analyzed based on trajectories using the Anisotropic Persistent Random Walk model (APRW).^[19,23] From this analysis, the parameters were generated that describe the movements of the cells, namely, MSD10, MSD100, Dtot, Dp, Dnp, Pp, Pnp, and ϕ . MSD10 and MSD100 were calculated since the denoted regimes are shorter and longer than the average persistence time across all conditions. The following equations were used to estimate the motility parameters.

$$MSD(\tau) = \langle (x(t+\tau) - x(t))^2 + (y(t+\tau) - y(t))^2 \rangle \quad (5)$$

$$MSD_p(\tau) = S_p^2 P_p \left(\tau - P_p \left(1 - e^{-\tau/P_p} \right) + 2\sigma_p^2 \right) \quad (6)$$

$$MSD_{np}(\tau) = S_{np}^2 P_{np} \left(\tau - P_{np} \left(1 - e^{-\tau/P_{np}} \right) + 2\sigma_{np}^2 \right) \quad (7)$$

$$D_p = \frac{S_p^2 P_p}{4} \quad (8)$$

$$D_{np} = \frac{S_{np}^2 P_{np}}{4} \quad (9)$$

$$D_{\text{tot}} = D_p + D_{np} \quad (10)$$

$$\phi = \frac{D_p}{D_{np}} \quad (11)$$

where S is the cell speed, P is the persistence time, $2\sigma^2$ is the noise (error) in the position of the cell, $\tau = n \Delta t$ and $n = 1, 2, \dots N_{\min} - 1$, Δt is the size of the time step.

Parameters and Descriptions: The eight motility parameters used for all analyses include the magnitudes of cellular displacements and speeds (MSD10, MSD100), the total diffusivity and diffusivities along primary and secondary axes of migration (Dtot, Dp, and Dnp), the persistence times along the primary and secondary axes of migration (Pp, Pnp), and the spatial persistence/anisotropy (ϕ). Time lags of 10 and 100 (for the MSDs) were selected based on the average persistence times across datasets. Ten minutes were on the order of average persistence time, and 100 min was greater than the average persistence time.

Normalized Mean Squared Displacement (MSD) Calculation: We normalized the Mean Squared Displacement (MSD) per time point to their corresponding time interval. For the non-homogenized trajectories, the following equation was used:

$$\frac{MSD(\tau_a)}{\tau_a} \quad (12)$$

where τ_a is the time interval of the actual non-homogenized trajectory. Pseudo-Monte-Carlo was used to extract the normalized MSDs at varying time intervals as follows:

$$\frac{MSD(\tau_i)}{\tau_i} \quad (13)$$

τ_i is the time interval used ($\tau_i = 2, 2.5, 3, 3.5, \dots 10$ min).

Clustering Analysis, Evaluating Cluster Stability and Cluster Coherence: k-means clustering was performed on the correlation distances to determine the emergent motility clusters among short-duration and long-duration trajectories. To determine behavioral classes of trajectories, unsupervised hierarchical clustering was performed on the centroids of k-means identified clusters, using the average linkage method and correlation distance metric. To analyze cluster stability, two factors were used: Inertia and Silhouette Values. Inertia values were based on the elbow method to identify the number of clusters when minimal additional data is gained with increased clusters.^[26] Silhouette values were calculated in MATLAB's silhouette function. To compute cluster coherence, the Euclidean distances were computed between the z-score, log normalized motility parameters for each cell, and z-score, log-normalized motility parameter values of the centroid of each cluster (obtained through k-means clustering).

Simulation of Trajectories: To simulate trajectories, APRW-based simulation of trajectories were used, as detailed in Wu et al.^[23] To test classification accuracies across short- and long-term durations, 75 000 and 25 000 unique cell trajectories were simulated, respectively.

Support Vector Machine (SVM) Based Data Classification and Model Prediction: MATLAB's built-in SVM (ClassificationsSVM) function was used to train on datasets and predict from the trained model. The trained model was a linear SVM using a hinge loss function with 90:10 train-test split. To get the cluster of a single cell, first, the higher-order parameters of the cell were calculated. Then, z-score and log normalization were performed to receive the input for a single cell to the SVM-based classification.

Centroid Based Prediction: To validate SVM-based prediction, centroid-based classification system was used that computes the Euclidean distance between the z-score, log normalized higher order parameters of each cell and z-score, log normalized higher order parameter values of the centroid (obtained through k-means clustering) of each cluster. The cluster was assigned as the cluster demonstrating the least Euclidean distance.

Sankey Dataset Calculation: To plot the temporal Sankey datasets, the clusters were first divided into pertinent groups: SG1, SG2, SG3, and SG4 for 2.5 h datasets and CG1, CG2, CG3, and CG4 for 8 h datasets. The temporal transition of 2.5 h trajectories were linked to their respective 8 h trajectories. Hence, three temporal segments of 2.5 h trajectories were plotted. The stability fraction was computed as the fraction of cells entering and staying in a short durations class from each of the 4 behavior classes for each 2.5-h time segment. For instance, SG1 and SG4 had higher stability fractions since most of the cells tended to stay in SG1 or SG4 if they started in SG1 or SG4. However, there were smaller, similar fractions of cells entering SG2 and SG3 from each of the four short duration behavior classes.

Entropy Calculations: *Spatial Entropy:* To compute spatial entropy, the occurrences of 8 h trajectories of the biological condition were used under consideration. The entropy was computed using the following equation:

$$S_{\text{spatial}} = - \sum_{i=1}^{25} p_i \ln(p_i) \quad (14)$$

The occurrences are probabilities (p_i) of cells in each $n = 25$ cluster ($i = 1, 2, 3, \dots, 25$).

State Entropy: To compute state entropy, the probability of each group SG1, SG2, SG3, SG4 was calculated at each time segment, T1, T2, and T3 (Equation 15) and CG1, CG2, CG3, and CG4 for the last part of the temporal Sankey (Equation 16).

$$S_{\text{state}}|_{T1, T2, T3} = - \sum_{i=1}^4 p_{SGi} \ln(p_{SGi}) \quad (15)$$

$$S_{\text{state}}|_C = - \sum_{i=1}^4 p_{CGi} \ln(p_{CGi}) \quad (16)$$

$S_{\text{state}}|_{T1}, S_{\text{state}}|_{T2}, S_{\text{state}}|_{T3}$ are calculated separately from Equation (15).

Transition Entropy: Transition entropy measures the heterogeneity in the distribution of transition from one-time segment to another.

$$S_{\text{transition}}|_{T1 \rightarrow T2; T2 \rightarrow T3; T3 \rightarrow C} = - \sum_{i=1}^4 \sum_{j=1}^4 p_{ij} \ln(p_{ij}) \quad (17)$$

For the computation of p_{ij} to estimate $S_{\text{transition}}|_{T1 \rightarrow T2}; S_{\text{transition}}|_{T2 \rightarrow T3}$, Equation (18) was used that uses the transition from SGi to SGj (groups of 2.5 h trajectories). For the calculation of p_{ij} to estimate $S_{\text{transition}}|_{T3 \rightarrow C}$, Equation (19) was used, which uses the transition from SGi (groups of 2.5 h trajectories) to CGj (groups of 8 h trajectories).

$$p_{ij} = \begin{bmatrix} p_{SG1 \rightarrow SG1} & p_{SG1 \rightarrow SG2} & p_{SG1 \rightarrow SG3} & p_{SG1 \rightarrow SG4} \\ p_{SG2 \rightarrow SG1} & p_{SG2 \rightarrow SG2} & p_{SG2 \rightarrow SG3} & p_{SG2 \rightarrow SG4} \\ p_{SG3 \rightarrow SG1} & p_{SG3 \rightarrow SG2} & p_{SG3 \rightarrow SG3} & p_{SG3 \rightarrow SG4} \\ p_{SG4 \rightarrow SG1} & p_{SG4 \rightarrow SG2} & p_{SG4 \rightarrow SG3} & p_{SG4 \rightarrow SG4} \end{bmatrix} \quad (18)$$

$$p_{ij} = \begin{bmatrix} p_{SG1 \rightarrow CG1} & p_{SG1 \rightarrow CG2} & p_{SG1 \rightarrow CG3} & p_{SG1 \rightarrow CG4} \\ p_{SG2 \rightarrow CG1} & p_{SG2 \rightarrow CG2} & p_{SG2 \rightarrow CG3} & p_{SG2 \rightarrow CG4} \\ p_{SG3 \rightarrow CG1} & p_{SG3 \rightarrow CG2} & p_{SG3 \rightarrow CG3} & p_{SG3 \rightarrow CG4} \\ p_{SG4 \rightarrow CG1} & p_{SG4 \rightarrow CG2} & p_{SG4 \rightarrow CG3} & p_{SG4 \rightarrow CG4} \end{bmatrix} \quad (19)$$

Conditional Transition Entropy: The heterogeneity in the distribution in motility groups was measured, given a specific group of a time segment. Three types of conditional transition entropies can link the three-time segments of 2.5 h trajectories (T1, T2, and T3). $S_{x \rightarrow i \rightarrow j}; S_{i \rightarrow x \rightarrow j}; S_{i \rightarrow j \rightarrow x}$ (The temporal transitions are defined as follows- T1: SGx, T2: SGi, T3: SGj, OR, T1: SGi, T2: SGx, T3: SGj, OR, T1: SGi, T2: SGj, T3: SGx, where, SGx is a given group in the respective time segment). p_{ij}^x is the probability of transition from SGi to SGj given SGx (at the respective time segment, namely, T1, T2, or T3).

$$S_{\text{cond.trans.}}|_{x \rightarrow i \rightarrow j; i \rightarrow x \rightarrow j; i \rightarrow j \rightarrow x} = - \sum_{i=1}^4 \sum_{j=1}^4 p_{ij}^x \ln(p_{ij}^x) \quad (20)$$

$$p_{ij}^x = \begin{bmatrix} p_{SG1 \rightarrow SG1}^x & p_{SG1 \rightarrow SG2}^x & p_{SG1 \rightarrow SG3}^x & p_{SG1 \rightarrow SG4}^x \\ p_{SG2 \rightarrow SG1}^x & p_{SG2 \rightarrow SG2}^x & p_{SG2 \rightarrow SG3}^x & p_{SG2 \rightarrow SG4}^x \\ p_{SG3 \rightarrow SG1}^x & p_{SG3 \rightarrow SG2}^x & p_{SG3 \rightarrow SG3}^x & p_{SG3 \rightarrow SG4}^x \\ p_{SG4 \rightarrow SG1}^x & p_{SG4 \rightarrow SG2}^x & p_{SG4 \rightarrow SG3}^x & p_{SG4 \rightarrow SG4}^x \end{bmatrix} \quad (21)$$

Predicting 2D versus 3D: To predict cells moving in either 2D or 3D, a linear SVM for each entropy was trained based on 25 iterations of a 90:10 split of the 60 conditions. The final accuracy results are reported as the average values for each 25 iterations.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

The authors acknowledge financial support for this study from the National Institutes of Health U01AG060903 (J.M.P., J.W.), the Johns Hopkins University Older Americans Independence Center of the National Institute on Aging (NIA) under award number P30AG021334 (J.M.P., J.W.), and Start-up funds from the Biomedical Engineering Department and the Whiting School of Engineering at Johns Hopkins University.

Conflict of Interest

D.M. and J.M.P. are co-inventors on a patent application related to this work. All other authors declare no other conflicts of interest.

Author Contributions

D.M. and N.S. contributed equally as co-first authors. P.K. and N.Z. contributed equally as co-second authors. J.M.P. conceived the study; D.M., N.S., and J.M.P. conceived data analysis and designed visualization workflows; N.Z., H.J., W.D., J.M.P., A.J., B.S., L.V., A.A., D.R., A.M., L.C., and J.W. performed experiments, collected data, and contributed reagents; D.M., N.S., N.Z., P.K., C.M., and J.M.P. performed formal analysis; D.M., N.S., N.Z., J.W., and J.M.P. interpreted results; N.S., D.M., and J.M.P. wrote manuscript; all authors contributed to reviewing and editing manuscript; J.M.P. supervised study and secured funding.

Data Availability Statement

The authors declare that all the data supporting the findings in this study including the raw trajectories for both short (Table S3) and long (Table S6) time scales are available on the Phillip lab GitHub repository: <https://github.com/PhillipLab-JHU/CaMI>.

Keywords

cell motility, high-throughput cell phenotyping, single-cell behaviors

Received: January 24, 2024

Revised: June 21, 2024

Published online: August 13, 2024

- [1] J. B. Swann, M. J. Smyth, *J. Clin. Invest.* **2007**, *117*, 1137.
- [2] K. Choe, Y. Hontani, T. Wang, E. Hebert, D. G. Ouzounov, K. Lai, A. Singh, W. Béguelin, A. M. Melnick, C. Xu, *Nat. Immunol.* **2022**, *23*, 330.
- [3] G. Crainiciuc, M. Palomino-Segura, M. Molina-Moreno, J. Sicilia, D. G. Aragones, J. L. Y. Li, R. Madurga, J. M. Adrover, A. Aroca-Crevillén, S. Martin-Salamanca, A. S. del Valle, S. D. Castillo, H. C. E. Welch, O. Soehnlein, M. Graupera-F. Sánchez-Cabo, A. Zarbock, T. E. Smithgall, M. Di Pilato, T. R. Mempel, P. L. Tharaux, S. F. González, A. Ayuso-Sacido, L. G. Ng, G. F. Calvo, I. González-Díaz, F. Díaz-de-María, A. Hidalgo, *Nature* **2022**, *601*, 415.
- [4] D. Wirtz, K. Konstantopoulos, P. C. Searson, *Nat. Rev. Cancer* **2011**, *11*, 512.
- [5] M. Efremova, S. A. Teichmann, *Nat. Methods* **2020**, *17*, 14.
- [6] C. Zhu, S. Preissl, B. Ren, *Nat. Methods* **2020**, *17*, 11.
- [7] J. M. Phillip, N. Zamponi, M. P. Phillip, J. Daya, S. McGovern, W. Williams, K. Tschudi, H. Jayatilaka, P. H. Wu, J. Walston, D. Wirtz, *Commun Biol.* **2021**, *4*, 81.
- [8] X. Ma, O. Dagliyan, K. M. Hahn, G. Danuser, *PLoS Comput. Biol.* **2018**, *14*, e1006321.
- [9] S. V. Stassen, G. G. K. Yip, K. K. Y. Wong, J. W. K. Ho, K. K. Tsia, *Nat. Commun.* **2021**, *12*, 5528.
- [10] J. C. Kimmel, A. Y. Chang, A. S. Brack, W. F. Marshall, *PLoS Comput. Biol.* **2018**, *14*, e1005927.
- [11] W. Du, P. Nair, A. Johnston, P.-H. Wu, D. Wirtz, *Annu. Rev. Biomed. Eng.* **2022**, *24*, 275.
- [12] H. Jayatilaka, J. M. Phillip, *Breast Cancer Manag.* **2019**, *8*, <https://doi.org/10.2217/bmt-2019-0002>.
- [13] H. Jayatilaka, A. Giri, M. Karl, I. Aifuwa, N. J. Trenton, J. M. Phillip, S. Khatau, D. Wirtz, *FASEB J.* **2018**, *32*, 1207.
- [14] J. M. Phillip, P. H. Wu, D. M. Gilkes, W. Williams, S. McGovern, J. Daya, J. Chen, I. Aifuwa, J. S. H. Lee, R. Fan, J. Walston, D. Wirtz, *Nat. Biomed. Eng.* **2017**, *1*, 0093.
- [15] L. Venturutti, M. Teater, A. Zhai, A. Chadburn, L. Babiker, D. Kim, W. Béguelin, T. C. Lee, Y. Kim, C. R. Chin, W. T. Yewdell, B. Raught, J. M. Phillip, Y. Jiang, L. M. Staudt, M. R. Green, J. Chaudhuri, O. Elemento, P. Farinha, A. P. Weng, M. D. Nissen, C. Steidl, R. D. Morin, D. W. Scott, G. G. Privé, A. M. Melnick, *Cell* **2020**, *182*, 297.
- [16] P. H. Wu, A. Giri, S. X. Sun, D. Wirtz, *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 3949.
- [17] J. M. Phillip, P.-H. Wu, D. M. Gilkes, W. Williams, S. McGovern, J. Daya, J. Chen, I. Aifuwa, J. S. H. Lee, R. Fan, J. Walston, D. Wirtz, *Nat. Biomed. Eng.* **2017**, *1*, 0093.
- [18] H. Jayatilaka, P. Tyle, J. J. Chen, M. Kwak, J. Ju, H. J. Kim, J. S. H. Lee, P. H. Wu, D. M. Gilkes, R. Fan, D. Wirtz, *Nat. Commun.* **2017**, *8*, 15584.
- [19] P.-H. Wu, A. Giri, S. X. Sun, D. Wirtz, *Proc. Natl. Acad. Sci.* **2014**, *111*, 3949.
- [20] A. Giri, S. Bajpai, N. Trenton, H. Jayatilaka, G. D. Longmore, D. Wirtz, *FASEB J.* **2013**, *27*, 4089.
- [21] H. Jayatilaka, A. Giri, M. Karl, I. Aifuwa, N. J. Trenton, J. M. Phillip, S. Khatau, D. Wirtz, *FASEB J.* **2018**, *32*, 1207.
- [22] E. R. Jerison, S. R. Quake, *Elife* **2020**, *9*, e53933.
- [23] P. H. Wu, A. Giri, D. Wirtz, *Nat. Protoc.* **2015**, *10*, 517.
- [24] J. M. Phillip, N. Zamponi, M. P. Phillip, J. Daya, S. McGovern, W. Williams, K. Tschudi, H. Jayatilaka, P. H. Wu, J. Walston, D. Wirtz, *Commun. Biol.* **2021**, *4*, 81.
- [25] J. C. Kimmel, A. Y. Chang, A. S. Brack, W. F. Marshall, *PLoS Comput. Biol.* **2018**, *14*, e1005927.
- [26] J. M. Phillip, K. S. Han, W. C. Chen, D. Wirtz, P. H. Wu, *Nat. Protoc.* **2021**, *16*, 754.
- [27] C. Zhang, Y. Li, Y. Wu, L. Wang, X. Wang, J. Du, *J. Biolog. Chem.* **2013**, *288*, 1489.
- [28] S. I. Fraley, P. H. Wu, L. He, Y. Feng, R. Krisnamurthy, G. D. Longmore, D. Wirtz, *Sci. Rep.* **2015**, *5*, 14580.
- [29] D. O. Velez, B. Tsui, T. Goshia, C. L. Chute, A. Han, H. Carter, S. I. Fraley, *Nat. Commun.* **2017**, *8*, 1651.
- [30] S. I. Fraley, Y. Feng, R. Krishnamurthy, D. H. Kim, A. Celedon, G. D. Longmore, D. Wirtz, *Nat. Cell Biol.* **2010**, *12*, 598.
- [31] J. P. Baskaran, A. Weldy, J. Guarin, G. Munoz, P. H. Shpilker, M. Kotlik, N. Subbiah, A. Wishart, Y. Peng, M. A. Miller, L. Cowen, M. J. Oudin, *APL Bioeng.* **2020**, *4*, 26105.
- [32] D. A. Lauffenburger, A. F. Horwitz, *Cell* **1996**, *84*, 359.
- [33] A. C. Johnston, G. M. Alicea, C. C. Lee, P. V. Patel, E. A. Hanna, E. Vaz, A. Forjaz, Z. Wan, P. R. Nair, Y. Lim, T. Chen, W. Du, D. Kim, T. D. Nichakawade, V. W. Rebecca, C. L. Bonifant, R. Fan, A. L. Kiemen, P.-H. Wu, D. Wirtz, *bioRxiv* **2024**, <https://doi.org/10.1101/2023.12.13.571595>.
- [34] G. Crainiciuc, M. Palomino-Segura, M. Molina-Moreno, J. Sicilia, D. G. Aragones, J. L. Y. Li, R. Madurga, J. M. Adrover, A. Aroca-Crevillén, S. Martin-Salamanca, A. S. del Valle, S. D. Castillo, H. C. E. Welch, O. Soehnlein, M. Graupera, F. Sánchez-Cabo, A. Zarbock, T. E. Smithgall, M. Di Pilato, T. R. Mempel, P.-L. Tharaux, S. F. González, A. Ayuso-Sacido, L. G. Ng, G. F. Calvo, I. González-Díaz, F. Díaz-de-María, A. Hidalgo, *Nature* **2022**, *601*, 415.
- [35] S. J. J. Kwok, D. T. Montoro, A. L. Haber, S.-H. Yun, V. Vinarsky, *bioRxiv* **2023**, *12*, e76645.
- [36] L. You, P.-R. Su, M. Betjes, R. G. Rad, T.-C. Chou, C. Beerens, E. van Oosten, F. Leukens, P. Gasecka, M. Muraro, R. van Tol, D. van Steenderen, S. Farooq, J. A. U. Hardillo, R. B. de Jong, D. Brinks, M.-P. Chien, *Nat. Biomed. Eng.* **2022**, *6*, 667.
- [37] P. L. Bedard, A. R. Hansen, M. J. Ratain, L. L. Siu, *Nature* **2013**, *501*, 355.
- [38] I. Dagogo-Jack, A. T. Shaw, *Nat. Rev. Clin. Oncol.* **2018**, *15*, 81.