Profiling emergent behaviors from single-cell motility

Debonil Maity[1,2], Pratik Kamat[2,3], Nahuel Zamponi[4], Hasini Jayatilaka[2], Anjil Giri[2], Wenxuan Du[2,3], Adrian Johnston[2,3], Bartholomew Starich[2,3], Michelle Karl[2,3], Anshika Agrawal[3], Deanna Riley[2], Leandro Venturutti[5], Ari Melnick[4], Leandro Cerchietti[4], Jeremy Walston[2,6], Jude M. Phillip[1,2,3,7,+]

[1.] Department of Biomedical Engineering, Johns Hopkins University, Baltimore, Maryland

[2.] Institute for Nanobiotechnology, Johns Hopkins University, Baltimore, Maryland

[3.] Department of Chemical and Biomolecular Engineering, Johns Hopkins University, Baltimore, Maryland

[4.] Department of Medicine, Division of Hematology and Medical Oncology, Weill Cornell Medicine, New York, New York

[5.] Department of Pathology and Laboratory Medicine, University of British Columbia; Centre for Lymphoid Cancer, British Columbia Cancer Research Institute Vancouver, British Columbia

[6.] Department of Geriatric Medicine and Gerontology, Johns Hopkins School of Medicine, Baltimore, Maryland

[7.] Department of Oncology, Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins School of Medicine, Baltimore, Maryland

[+] Corresponding Author:
Jude M. Phillip, PhD
Johns Hopkins University,
3400 North Charles Street,
Croft Hall 100,
Baltimore, MD, 21218
jphillip@jhu.edu

Keywords: cell motility, high-throughput cell phenotyping, single-cell profiling, cell behaviors

Short title: Identifying and classifying single-cell motility patterns

Teaser: A computational framework to identify and classify single-cell motility behaviors and phenotypic heterogeneity.

Abstract

Cell motility is essential to many biological processes as cells navigate and interact within their local microenvironments. Currently, most methods to quantify cell motility rely on the ability to follow and track individual cells. However, results from these approaches are typically reported as averaged values across cell populations. While these approaches offer biological simplicity, it limits the ability to assess cellular heterogeneity and infer generalizable patterns of single-cell behaviors at baseline or after perturbations. Here, we present CaMI, a computational framework that takes advantage of the single-cell nature of cell motility data to identify and classify distinct spatio-temporal behaviors of single cells. Using CaMI, we demonstrate the ability to robustly classify single-cell motility patterns in a large dataset (n=74,253 cells), quantify spatio-temporal heterogeneities, determine motility patterns in unclassified cells, and provide a visualization scheme for direct interpretation of dynamic cell behaviors. Furthermore, as a biological proof of concept, we investigate the biphasic spatial and temporal responses of T-cell lymphoma cells moving in Collagen-I gels of varying collagen concentrations and predict the dimensionality (2D vs. 3D) of matched cellular conditions based solely on spatio-temporal heterogeneities. Together, we present a multivariate framework to robustly classify emergent patterns of single-cell behaviors, highlighting cellular heterogeneity as a critical feature that establishes the behaviors of cellular populations.

Introduction

Single-cell profiling is critical for properly understanding and interpreting cellular heterogeneity and biological phenotypes in homeostasis and disease[1,2]. Recent advances in single-cell technologies and analyses have enabled the quantitative description of cell types, cell states, and cell fates based on time-anchored snapshots of genetic, epigenetic, or protein signatures. The application of similar single-cell approaches to profile the dynamics and functional behaviors of cells holds much promise. Profiling cell motility and other dynamic properties of cells as they move and interact within local microenvironments can provide critical insights into functional cell behaviors in the context of health and disease[3]. Furthermore, quantifying cell motility informs our understanding of many biological processes, including immune locomotion and surveillance[4,5] and cancer metastasis[6,7]. However, conventional cell motility data captured *in vitro* or *in vivo* are often performed on small numbers of homogeneous cells and tend to be reported as averaged values for populations of cells, *e.g.,* average speeds and average displacements. While in some ways averaging motility parameters offer biological simplicity, it limits the ability to assess cellular heterogeneity and infer generalizable cell motility patterns that define single-cell behaviors per condition[8].

Combining single-cell profiling of cell motility with modern data science approaches and machine learning offers promising solutions to define functional cell behaviors across multiple conditions[8–12]. Furthermore, performing these 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 molecular programs on dynamic cell behaviors[13,14]. Despite this promise, robust quantification of cell motility patterns to inform functional behaviors and heterogeneities of 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) and a lack of large publically available cell motility datasets for comparative interrogation of motility trends and behaviors. To address this, we present the computational framework CaMI, which stands for Cellular Motility Index, to enable the direct integration and harmonization of motility datasets from multiple sources having different acquisition settings. Following data harmonization, CaMI enables the robust identification and classification of single-cell motility behaviors, quantification of phenotypic heterogeneities, and provides a visualization scheme to interpret biological conditions quantitatively.

Using CaMI, we demonstrate the ability to combine non-uniform cell motility datasets to profile 74,253 single-cells spanning 409 biological conditions (**Supplemental table 1**). These biological conditions include different cell types, genetic and drug perturbations for cells in either two-dimensional (2D) or three-dimensional (3D) microenvironments. With this combined dataset (**Supplementary Dataset 1**), we 1) identify and classify distinct cell motility behaviors at single-cell resolution for short (2.5 hours) and long (8 hours) durations, 2) determine phenotypic heterogeneities per condition (*i.e*.,

spatial and temporal), and 3) predict the motility patterns for previously unclassified cells. Furthermore, we show the ability to map temporal transitions among different cell motility behaviors and enable new inferences that are typically limited due to the assumption of time-invariance, which is a critical assumption in most motility analyses. As biological proof of concepts, we applied CaMI to experimental conditions of T- lymphoma cells moving in 3D Collagen-I gels of varying collagen concentrations and predicted the dimensionality of cells (2D vs. 3D) from 60 matched conditions based solely on the magnitude of spatial and temporal heterogeneities. We also show that cellular heterogeneity is a defining feature of cell motility, with experimental perturbations inducing redistributions that shift the abundances of cells among distinct cell motility patterns and behaviors. Finally, we provide a large, annotated cell motility dataset to directly compare motility patterns within user-generated datasets.

Results

Data acquisition and curation

To develop a computational framework to classify universal patterns and behaviors of single cells across multiple conditions, we combined existing and newly acquired cell motility datasets[8,15–20]. We tracked cells using MetaMorph, and x-y trajectories across all experimental conditions had total imaging durations ranging between 2.5 to 16 hours, with image acquisition intervals ranging from 2 to 10 minutes. All cell trajectories presented were derived from time-lapse movies collected using brightfield or phase-contrast imaging.

Image acquisition time determines the shape of cell movement trajectories

There are no universal standards for acquiring cell motility data regarding total acquisition duration or image acquisition intervals. To combine motility datasets across multiple conditions (n=409) from multiple sources, we first needed to develop a method to integrate the data without introducing noise and artifacts into the datasets. As a result, we needed a way to directly combine datasets independent of the total duration and the acquisition intervals. To do this, we first restricted the total imaging duration to 2.5 hours to maximize the total number of cells, then evaluated the effective contribution of the image acquisition intervals (*i.e.,* the time between subsequent images) on the magnitude of cell motility. Using a test case of HT1080 fibrosarcoma cells moving in 3D Collagen-I gels, we *in-silico* varied the time-intervals from the most time-resolved—*i.e*., 2 minutes, to a degraded time-resolution of 10 minutes (**Figure 1A**). Here, varying the time intervals are as simple as leaving out intervening points from the fully time-resolved trajectory. Results indicated that the interval times significantly impacted the shapes of the cell trajectories, which was more prominent in highly motile cells. To quantify this deviation, we computed the concordance among cell trajectories of varying time intervals (**Figure 1A**). We calculated the concordance as the geometric similarities among cell trajectories based on the areas of triangles

formed between intervening points along a given trajectory (**Figure S1A**, **Methods**). Together, these results highlight that acquisition intervals contribute significantly to the magnitude of the perceived motility and that temporal resolution matters, *i.e.*, cell movements based on 2-minute intervals do not equal cell movement at 10-minute intervals.

Harmonizing single-cell motility data across multiple datasets

Given the importance of interval times on the shape of cell movement trajectories, we developed a robust way to determine the intermediate x-y locations along the movement trajectories of cells. Using geometric interpolation, we identified spatially weighted cell locations for symmetric (midpoint) and asymmetric multiples of the acquisition-interval times (**Figure 1B-C**). We observed that while interpolated cell locations from asymmetric multiples of the interval times showed no statistical differences in the concordance relative to actual data points (**Figure 1C**), symmetric multiples of the interval times using interpolation showed significant deviations (**Figure 1B, Figure S1B**). To correct this, we developed a pseudo-Monte Carlo (pMC) approach that weights a small directional propensity to the interpolated locations to estimate the location of cells at a given time point (see **Methods**). With this, we showed that a pMC factor of 2.01 (pMC-2) worked well across all conditions, generating similar concordances to the trajectories generated using actual data points for both the symmetric and asymmetric conditions (**Figure 1B-C**). Plots of the mean squared displacements (MSD) showed that pMC-2 could replicate short and long time-scale behaviors, whereas interpolation showed significant deviations at short time scales (**Figure 1D**). To further confirm this, we computed motility parameters based on the anisotropic persistent random walk (APRW) model, a framework that acknowledges that cells do not always follow random walks on 2D or 3D substrates[17,21]. Results indicated no statistical differences between actual and reconstructed trajectories using pMC-2 for any of the eight parameters assessed. However, interpolation showed differences in some motility parameters, including short-time MSD and the persistence time in the secondary axis of migration (**Figure 1E, Figure S1C**).

Using pMC-2 reconstructed trajectories across all conditions (n=409), we asked whether there was an optimal interval time that worked across all experimental conditions within our dataset, allowing the seamless integration of multiple datasets. Taking the 74,253 single cells, we iteratively varied the interval times between 2 and 10 minutes for every cell, then computed the p-values relative to the real acquisition intervals per cell and condition. Results indicated that an interval of 5-minute worked well across all conditions, as indicated by the least number of conditions having p-values greater than 0.05 (**Figure 1F**) relative to the ground truth of each cell (*i.e.,* highest time resolution). With this result, we converted all conditions to have an interval time of 5 minutes and a total duration of 2.5 hours. We chose 2.5 hours not only to maximize the number of cells within our dataset but also to represent a time-scale comparable to what is achievable in both *in vivo* and *in vitro* systems.

Taken altogether, we show the ability to combine and integrate non-uniform cell motility datasets that span multiple conditions and acquisition intervals.

Classifying functional cell behaviors at single-cell resolution

Using this new harmonized dataset (**Supplementary dataset 1**), we asked whether we could identify distinct patterns of cell movements across all conditions. Using the 74,253 single-cells, we computed eight motility parameters based on the APRW model (see **Methods**), which we previously showed to describe critical characteristic of cell movements[8,15,21]. These parameters include, short and long time-scale mean squared displacements (MSD10, MSD100), persistence times in the primary and secondary axes of migration (Pp, Pnp), total diffusivities and diffusivities in the primary and secondary axes of migration (Dtot, Dp, Dnp), and the anisotropic index that describes the spatial persistence (ϕ) (see **Methods**). Using k-means clustering and calculations of the inertia and silhouette values[22], we determined inherent cell-based groupings that divided into 17 distinct patterns of cell motility (**Figure 2A-B**, **Figure S2A-C**). To confirm this classification, we calculated the coherence factor based on the vector-distances of every cell from each cluster-centroid. Results showed the expected staircase plot, highlighting that cells classified within each cluster exhibited small distances from their assigned cluster centroids, and significant distances relative to the other unassigned cluster-centroids (**Figure 2C**).

Using two-dimensional t-stochastic neighbor embedding (tSNE), we plotted all 74,253 cells and painted each cell based on the designated motility cluster (n=17). This analysis showed that each cell clustered close to other cells classified within the same cluster (S1 to S17). We then grouped clusters using unsupervised hierarchical clustering to describe gross similarities in cell behaviors, resulting in four behavior classes (SG1, SG2, SG3, SG4) (**Figure 2D**). Each cluster and corresponding class showed intrinsic behavioral patterns defined based on the magnitude of the eight motility parameters (**Figure 2E, Figure S2D**). To visually confirm that different clusters denoted differential cell motility behaviors, we plotted the trajectories of cells classified within each cluster, highlighting clear differences in cell behaviors per motility cluster (**Figure 2F**).

Taken together, we demonstrate a multivariate approach to identify and classify distinct single-cell motility patterns across multiple datasets (**Supplementary Table1**).

Predicting cell motility clusters in unclassified cells

Prompted by the robust classification of cells across multiple datasets, we asked whether we could predict the motility cluster of unclassified cells. As part of this analysis, we performed a randomized 90:10 split (training and validation sets) of the combined dataset (n=74,253 single cells). Using 90% of cells, we trained a machine learning algorithm using linear support vector machine (SVM), then applied the trained classification model to the remaining 10% of unclassified cells (see **Methods**). Our analysis

showed a 98.7% accuracy across all clusters (**Figure 2G**), where we determined the true class of cells within each cluster based on the initial cellular classification using all 74,253 cells. To further confirm the classification scheme and show the ability to predict the motility cluster in previously unclassified cells, we simulated 75,000 new trajectories (see **Methods**), then performed the classification using the SVM model trained on the 74,253 cells. This analysis achieved a 96.7% accuracy across all clusters (**Figure 2H**). Here, we determined the true class of each simulated cell by performing an iterative analysis to re-classify all the cells adding one new unclassified cell (simulated cell) each time, (*i.e.* re-training on 74,253 cells + 1 cell, 75,000 times).

Together, we demonstrate the robust prediction of single-cell motility patterns using a trained machine learning model based on the eight key motility parameters that describe dynamic cellular behaviors.

Cell motility patterns define phenotypic shifts and heterogeneity of primary monocytes

To demonstrate the biological utility of our approach, we collected peripheral blood mononuclear cells from three healthy donors and isolated monocyte populations (CD14+) from each donor. Using 3D time-lapse microscopy, we imaged the monocytes moving in three-dimensional collagen gels (2mg/ml) for 3 hours at 5-minute intervals, using 2.5 hours for analysis. In a fraction of the monocytes from each donor, we applied a chemical perturbation by exposing monocytes to conditioned media taken from breast cancer cells. Following exposure, we generated the coordinates of each cell trajectory per movie, then classified cells using the trained SVM model described above. While each donor exhibited unique baseline profile according to the abundance of cells within each motility cluster, there were overt similarities across all donors (**Figure 2I**). Interestingly, cells exposed to conditioned media (+CM) demonstrated a significant shift in the motility patterns relative to their baseline. The exposed monocytes from all three donors showed a similar shift to cells having elevated abundances of motility clusters S12, S13, S14, and S15, corresponding to patterns that describe highly motile cells. Furthermore, these shifts were associated with decreased phenotypic heterogeneity, as measured by the Shannon entropy (**Figure 2J**), and a higher correlation among CM-exposed monocytes as opposed to each donor's baseline (**Figure 2K**, **Figure S2E**). Based on this shift towards patterns of higher motility, we plotted bulk motility parameters for baseline and +CM conditions. Results confirmed that populations of monocytes exposed to CM had enhanced motility based on average speeds, average MSD10 and average total-diffusivity (**Figure 2L**).

Taken together, we show that our single-cell framework can directly quantify dynamic behaviors and phenotypic heterogeneity across cellular conditions. Using this single-cell approach, we retain all the biological insights that can be determined from bulk analysis, while adding new ways to quantify phenotypic heterogeneity and direct visualization of cell motility data.

Quantification of dynamic cell behaviors over long time durations

We had more than eight hours of continuous cell tracking data for a subset of cells within the combined dataset. As a result, we wondered whether additional dynamic patterns could be captured from longer movies, enabling subsequent interrogation of cell behaviors at short and long durations. To evaluate this, we compiled cells having 8-hours of motility data at 5 minute intervals and computed the same eight motility parameters described above. With this data (n=24,753 cells) we performed k-means clustering analysis (**Figure 3A**) and computed the inertia and silhouette values, from which we identified 25 clusters for long durations (**Figure 3B**). The 25 clusters were confirmed based on the staircase plots highlighting the coherence within each cluster (**Figure 3C**), with cells clustering within proximity in two-dimensional tSNE space (**Figure 3D**). These 25 clusters (CG1 to CG25) also separated into four behavior classes (CG1, CG2, CG3, CG4) based on unsupervised hierarchical clustering, displaying distinct patterns for the eight motility parameters (**Figure 3E**, **Figure S3A-D**). We plotted the cells classified per cluster to visualize the motility patterns within each cluster. Results indicated that cell clusters exhibited distinct patterns of cell migration that grossly define gradients of directedness, temporal persistence, and displacement per single cell (**Figure 3F**).

We next evaluated the ability to predict the motility pattern of unclassified cells. Using the same approach noted above, we trained a new linear SVM model on the 8-hour cell tracking data, tested the prediction in a 90:10 split dataset, and newly simulated trajectories for 25,000 single cells. Results indicated a 97.2% accuracy for the 90:10 split dataset (**Figure 3G**), and a 95.6% accuracy for the simulated trajectories (**Figure S3E**). Taken together, we show the ability to also classify distinct patterns of cell motility for long durations of cell movements (**Supplementary Table 2**).

Connecting motility patterns across short and long time durations

In most cell motility analyses the inherent assumption of time-invariant movement or time-independence limits the ability to infer generalizable temporal cell behaviors. To understand the temporal transitions and infer temporal behaviors, we took advantage of having both short and long duration data for a subset of cells. Using this information, we linked the analyses of motility patterns at short durations (2.5 hours that define 17 clusters) with motility patterns at long durations (8 hours that define 25 clusters). For the subset of cells having 8 hours of movement, we cut each trajectory into three 2.5-hour segments and determined motility clusters and complimentary behavior classes (SG1-SG4 and CG1-CG4) within short and long durations (**Supplementary Table 3**). With this data, we asked whether a cell classified in a particular motility cluster or behavior class for the first 2.5 hours stayed within that class for subsequent 2.5-hour time segments. Overall, we observed that for the ~24,000 cells evaluated, cells classified within classes SG1 and SG4 tended to stay within the same

behavior class for the continuous three time-segments (**Figure 3H**). In contrast, cells classified in SG2 and SG3 tended to change behavior classes along subsequent segments. From this, we computed the conditional probabilities of transition for each behavior class (SG1-SG4) (**Figure 3I**), highlighting that SG1 and SG4 had higher temporal stabilities relative to SG2 and SG3 (**Figure 3J**). Furthermore, there was an association between short and long durations, with classes in long durations associating with differential short duration classes; *e.g.,* CG2 comprised mainly of SG1 and SG3 cells, and CG1 consisted of SG1 with a minority of SG2 cells (**Figure 3H**).

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 particularly important, as it provides a new way to overcome the limitations imposed by the assumption of time-invariance. By separating cell trajectories into shorter time segments, where the time-invariance assumption is valid, we can decipher spatio-temporal patterns that accurately describes spatio-temporal behaviors of cells. With all these steps, from data harmonization and integration to spatio-temporal classification we define CaMI to describe dynamic cell behaviors.

Using CaMI to decipher functional heterogeneity across conditions

Given the ability to profile dynamic behaviors of single cells using CaMI, we tested the ability to derive key biological insights that was not apparent using conventional motility analyses based on bulk-averaged parameters. Using a subset of data taken from long durations movies (*i.e.* 8 hours), 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 4A, Figure S4A**) and two-dimensional tSNE plots (**Figure 4B**). As a biological test-case, we took data from LY12 cells (T-cell lymphoma cells) moving in 3D Collagen-1 gels across five different collagen concentrations. Plotting the trajectories of cells within each condition, we qualitatively observed a bi-phasic response in the magnitude of the motility, with the lowest cell motility observed at 3mg/ml collagen (**Figure 4C**). Plotting this condition in the 2D t-SNE space (**Figure 4B**), we quantitatively observed a bi-phasic or switching effect. This biphasic response was indicated by cells moving from the green region (low collagen concentration) to the purple region (high collagen-3mg/ml), then back to green region for 4mg/ml (**Figure 4D**), highlighting a switch in the motility phenotypes based on spatial motility patterns (similar motility for 1mg/ml and 4mg/ml). This switching phenotype with changing collagen concentrations is likely due to local changes in molecular crowding effects based on collagen alignment and porosity[23,24]. Furthermore, plotting the abundance of cells within the 25 clusters for each condition, we also observed a bi-phasic response in the heterogeneity itself (defined based on the Shannon entropy), with a minimization at 3mg/ml (**Figure 4E**).

Given these striking spatial associations with collagen concentration, we asked whether there were complimentary temporal associations based on short and long time durations. To address this, we stacked the three 2.5-hour time segments, paying attention to the fraction of cells within each short duration behavior classes (SG1 to SG4) and the frequency at which cells switched classes with increasing concentrations. From this analysis, we also observed a switching/bi-phasic phenotype based on temporal associations, with the majority of cells staying in SG4 (yellow) across the three segments for 0.5mg/ml, 1mg/ml and 4mg/ml, an increase towards SG1 for 2mg/ml, and a shift towards SG1 as the dominant behavior class for 3mg/ml (**Figure 4F**). Furthermore, while this phenotype-switch is observed spatially at 3mg/ml, from the temporal perspective this shift is already apparent at 2mg/ml and is enhanced at 3mg/ml. We observed these temporal switching dynamics across conditions, not only for LY12 cells. Conditions for other cell types also exhibited similar switching dynamics with drug treatments or knockdowns using different shRNAs. These generalized findings establish the utility of CaMI, and the ability to determine new insights into motility patterns in multiple biological settings (**Figure S4**). In addition, CaMI analysis can quantify dynamic phenotypic transitions to identify precursor shifts in phenotypes before the complete switching (*i.e.* 2mg/ml shows a trend towards switching from SG4 to SG1 before SG1 becoming dominant at 3mg/ml).

Taken together, CaMI enables the quantification of real spatial-temporal dynamics and heterogeneities that define bi-phasic and phenotype-switching responses to increasing collagen concentration. These results highlight key dynamic cell behaviors that are not apparent using conventional bulk motility analysis (**Figure 4G**).

Predicting dimensionality of cellular conditions based on spatial and temporal heterogeneities

Given the ability to dynamically map spatio-temporal cell behaviors, we wondered whether CaMI could predict the dimensionality of cells moving in 2D and 3D microenvironments. Recent studies have demonstrated that cell movements in 2D do not predict 3D movements[17,25,26], 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 movements 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 vs. 3D. These conditions comprised 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, indicating inherent differences in the motility patterns of the cells, with cells in 3D microenvironments exhibiting higher heterogeneities (**Figure 4H**). This

differential localization within the 2D t-SNE space was not only observed in knockdown conditions, but also as a function of drug concentrations in HT1080 cells and other cell types (**Figure S5**). We did not observe dimension-exclusive motility clusters, meaning that 2D and 3D cellular conditions had cells classified in each of the 25 long duration motility clusters. However, the fraction of cells within each cluster varied between 2D and 3D (**Supplemental Table 1**).

Prompted by these results, we asked whether we could predict the dimensionality of moving cells based on the phenotypic heterogeneity of cell populations, as cells in 3D tended to exhibit higher phenotypic heterogeneities relative to matched 2D conditions (**Figure 4I**). To comprehensively define the heterogeneity, we defined four types of entropies that captured spatial and temporal nature of cells (see **Methods**). These different types of entropy denoted: (1) spatial entropy, which is computed based on the fractional abundance in each motility cluster (n=17 or n=25), (2) state entropy, computed based on the fractional abundance of cells within short duration and long duration classes (SG1-SG4, CG1-CG4), (3) transition entropy, which is computed based on the probability of transitioning from one short duration behavior class to the other along subsequent time segments, and (4) conditional-transition entropy, estimates the entropy based on the conditional probabilities that define the likelihood of transitioning to a specific short duration class. We then trained a model to predict the dimensionality of cells using a linear SVM (see **Methods**). Comparing the prediction accuracies using either the eight motility parameters, the individual entropies, or the combination of all four entropies, we found that the combination of entropies was able to predict the dimensionality in more than 95% of the cases. This result was comparable to using the eight motility parameters to predict the dimensionality (**Figure 4J-K**).

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 the dynamic behaviors. Furthermore, understanding phenotypic heterogeneity is critical to interpreting spatio-temporal behaviors.

Discussion

Cell movements play a critical role in the physiology of organisms, both in the context of homeostasis and disease[27]. CaMI is a computational framework that takes advantage of the single-cell nature and spatio-temporal dynamics of cell motility data, enabling the harmonization of datasets across multiple conditions (**Figure 1**). Conventionally, cell motility is quantified using small numbers of cells and reported as bulk averages of cell populations. Here, we have demonstrated the utility of CaMI to quantify cell motility patterns (**Figure 2**), determine spatial and temporal heterogeneities (**Figure 3**), define dynamic phenotypic transitions, and predict the dimensionality of cellular conditions (**Figure 4**). In addition, CaMI provides a clear visualization scheme of single-cell behaviors to enable direct

interpretation of motility data, while retaining all the biological information typically gained from conventional bulk analyses. Furthermore, CaMI can classify dynamic cell behaviors *in vitro* and *in vivo* settings (**Figure S2F**) in response to genetic or chemical perturbation, *e.g*. drugs, shRNAs (**Figure 4H**, **Figure S5**).

While CaMI provides a robust solution to decipher emergent patterns of single-cell behaviors there are certain contexts where it may exhibit limitations. A key limitation of CaMI is that since cell motility clusters are derived based on multivariate statistical approaches of combined motility datasets, (*i.e.* spanning multiple cell types, dimensionalities, and perturbations), only dominant cell motility patterns would be identified as clusters. As a result, rare 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, we are confident that CaMI can provide reliable and insightful information, as demonstrated by the high cluster coherence (*i.e.* minimal distances of each cell to its cluster centroid). In any case, an alternative application of CaMI will be needed in conditions where cell-type or condition-specificity is critical. One way around this is to train the classification model based on a specific cell-type, or condition, as we have previously shown[8]. However, the dataset (*i.e*. number of cells) needs to 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. As such, multiple comparative conditions are needed to derive biological insights and to understand how a given perturbation influences the resistant cell behaviors, based on the abundance of cells within each cluster.

In summary, we present a robust methodology 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 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 and their underlying molecular states. Recently, studies have started to multiplex functional cell behaviors with underlying genetic and epigenetic states of cells[13,14,28]. However, many more studies are needed to comprehensively assess and establish deterministic links between cell motility patterns, cell state transitions, cells 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 particularly important in light of heterogeneous cellular behaviors within cell populations and the associations with clinical outcomes and treatment responses[29,30].

Materials and Methods

Data Curation

Data was generated to classify cell motility patterns based on the x-y coordinates along cell movement trajectories. We used existing motility data from published papers together with newly acquired data. We stored the x and y coordinates, total duration, interval time, pixel size for calibration, and condition descriptions for each dataset. Supplementary dataset 1 contains all curated data. This data matrix was compiled for each cell and used to generate the corrected datasets for analysis (2.5-hours total duration at 5-minute intervals, or 8-hours total durations at 5-minute intervals).

Calculation of Higher Order Parameters using anisotropic persistence random walk (APRW)

To quantify the spatio-temporal motility patterns, we analyzed bulk cell movements based on trajectories using Anisotropic Persistent Random Walk model (APRW)[17,21]. From this analysis we generated the parameters that describe the movements of the cells, namely, MSD10, MSD100, Dtot, Dp, Dnp, Pp, Pnp, and ϕ. We calculated MSD10 and MSD100 since the denote regimes are shorter and longer than the average persistence time across all conditions. We used the following equations to estimate the motility parameters.

$$MSD(\tau) = \left\langle (x(t+\tau) - x(t))^2 + (y(t+\tau) - y(t))^2 \right\rangle \tag{1}$$

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

$$MSD_{np}(\tau) = S_{np}^2 P_{np} \left(\tau - P_{np} \left(1 - e^{-\tau/P_{np}} \right) + 2\sigma_{np}^2 \right) \tag{3}$$

$$D_p = \frac{S_p^2 P_p}{4} \tag{4}$$

$$D_{np} = \frac{S_{np}^2 P_{np}}{4} \tag{5}$$

$$D_{tot} = D_p + D_{np} \tag{6}$$

$$\phi = \frac{D_p}{D_{np}} \tag{7}$$

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 $n = 1, 2, \ldots N_{min}-1$, 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, SP10, and SP100), 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 (used for the MSD's and speeds) were selected based on the average persistence times across datasets, 10 minutes were on the order of average persistence time and 100 minutes being 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, we used the following equation:

$$\frac{MSD(\tau_a)}{\tau_a} \tag{8}$$

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

$$\frac{MSD(\tau_i)}{\tau_i} \tag{9}$$

τ_i is the time interval used (τ_i=2, 2.5, 3, 3.5, … 10 min).

Pseudo-Monte-Carlo technique to predict y-t locations of cells at a given time point

First, we estimate the required (or, missing) X-Y location from linear interpolation $\left(x_{interp,i}, y_{interp,i}\right)$. Then, we employ a small perturbation to the estimated location of the cell obtained through linear interpolation. The perturbation for the location at ith time point is employed as follows:

$$x_{pMC,i} = x_{interp,i} + \delta_{x,i}; \tag{10}$$

$$y_{pMC,i} = y_{interp,i} + \delta_{y,i}; \tag{11}$$

$$\delta_{x,i} = \frac{(x_{i+1} - x_{i-1})}{pMCF}\left(\left|x_{i+1} - x_{i-1}\right| \geq \left|y_{i+1} - y_{i-1}\right|\right); \tag{12}$$

$$\delta_{y,i} = \frac{(y_{i+1} - y_{i-1})}{pMCF}\left(\left|x_{i+1} - x_{i-1}\right| \leq \left|y_{i+1} - y_{i-1}\right|\right) \tag{13}$$

The goodness of pMC method is then checked by varying the pMC factor (pMCF) and comparing with the corresponding Higher Order parameters or through comparing Normalized MSDs.

Assessing concordance among varied time intervals

We used a concordance analysis to geometrically compare different cell trajectories, which calculates the area of triangles formed along subsequent points along a given trajectory (**Figure S1A-B**). For concordance analysis regarding varying time intervals per trajectory, we calculated the triangular area made by points t_1-t_2-t_3, t_3-t_4-t_5 and so on (**Figure S1A**). For concordance analysis regarding interpolation-predicted trajectories, we calculated the trianglular area made by points t_1-t_{i2}-t_3, t_3-t_{i4}-t_5 and so on (**Figure S1B**) and for pMC predicted trajectory, we calculated the area of triangle made by points t_1-t_{i2}-t_{i3}, t_{i3}-t_{i4}-t_{i5} (**Figure S1B**).

Clustering analysis, evaluating cluster stability and cluster coherence

We used either k-means clustering or unsupervised hierarchical clustering in Matlab for clustering analysis. To analyze cluster stability, we used two factors: 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[22]. We calculated silhouette values in MATLAB's silhouette function. To compute cluster coherence, we computed the Euclidean distances between 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, we used APRW based simulation of trajectories as detailed in Wu et al[21].

Support Vector Machine (SVM) based data classification and model prediction

We used MATLAB's SVM function to train on datasets and predict from the trained model. We detail the data partitions for training and validation in the respective figure captions. To get the cluster of a single cell, first, we calculate higher order parameters of the cell. Then we performed z-score, log normalization to receive the input for a single cell to the SVM based classification.

Centroid based Prediction

To validate SVM based prediction, we used centroid based classification system which computes the Euclidean distance between 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. We assigned the cluster as the cluster demonstrating the least Euclidean distance.

Sankey dataset calculation

To plot the temporal Sankey datasets, we first divided the clusters into pertinent groups: SG1, SG2, SG3 and SG4 for 2.5hrs datasets and CG1, CG2, CG3 and CG4 for 8hrs datasets. We then linked the

temporal transition of 2.5hrs trajectories to their respective 8hrs trajectories. Hence, we plotted three temporal segments of 2.5hrs trajectories

Entropy Calculations

Spatial Entropy- To compute spatial entropy we used the occurrences of 8hrs trajectories of the biological condition under consideration. We then computed entropy using the following equation:

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

The occurrences are probabilities (p_i) of cells in each n=25 clusters (i=1,2,3…25).

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

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

$$S_{state}\big|_{C} = -\sum_{i=1}^{4} p_{CGi} \ln(p_{CGi}) \tag{16}$$

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

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

$$S_{transition}\big|_{\substack{T1\to T2;T2\to T3;\\ T3\to C}} = -\sum_{i=1}^{4}\sum_{j=1}^{4} p_{ij} \ln(p_{ij}) \tag{17}$$

For computation of p_{ij} to estimate $S_{transition}\big|_{T1\to T2}; S_{transition}\big|_{T2\to T3}$, we used equation (18) which uses the transition from SGi to SGj (groups of 2.5hrs trajectories). For computation of p_{ij} to estimate $S_{transition}\big|_{T3\to C}$, we used equation (19), which uses the transition from SGi (groups of 2.5hrs trajectories) to CGj (groups of 8hrs trajectories).

$$p_{ij} = \begin{bmatrix} p_{SG1\to SG1} & p_{SG1\to SG2} & p_{SG1\to SG3} & p_{SG1\to SG4} \\ p_{SG2\to SG1} & p_{SG2\to SG2} & p_{SG2\to SG3} & p_{SG2\to SG4} \\ p_{SG3\to SG1} & p_{SG3\to SG2} & p_{SG3\to SG3} & p_{SG3\to SG4} \\ p_{SG4\to SG1} & p_{SG4\to SG2} & p_{SG4\to SG3} & p_{SG4\to SG4} \end{bmatrix} \tag{18}$$

$$p_{ij} = \begin{bmatrix} p_{SG1\to CG1} & p_{SG1\to CG2} & p_{SG1\to CG3} & p_{SG1\to CG4} \\ p_{SG2\to CG1} & p_{SG2\to CG2} & p_{SG2\to CG3} & p_{SG2\to CG4} \\ p_{SG3\to CG1} & p_{SG3\to CG2} & p_{SG3\to CG3} & p_{SG3\to CG4} \\ p_{SG4\to CG1} & p_{SG4\to CG2} & p_{SG4\to CG3} & p_{SG4\to CG4} \end{bmatrix} \tag{19}$$

Conditional Transition Entropy- We measured the heterogeneity in distribution in motility groups, given a specific group of a time segment. Three types of conditional transition entropies can link the three-time segments of 2.5hrs trajectories (T1, T2 and T3). $S_{x\to i\to j}; S_{i\to x\to j}; S_{i\to j\to x}$(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).

$$S_{cond.trans.}\Big|_{\substack{x\to i\to j;i\to x\to j;\\ i\to j\to x}} = -\sum_{i=1}^{4}\sum_{j=1}^{4} p_{ij}^{x}\ln(p_{ij}^{x}) \tag{20}$$

$$p_{ij}^{x} = \begin{bmatrix} p^{x}_{SG1\to SG1} & p^{x}_{SG1\to SG2} & p^{x}_{SG1\to SG3} & p^{x}_{SG1\to SG4} \\ p^{x}_{SG2\to SG1} & p^{x}_{SG2\to SG2} & p^{x}_{SG2\to SG3} & p^{x}_{SG2\to SG4} \\ p^{x}_{SG3\to SG1} & p^{x}_{SG3\to SG2} & p^{x}_{SG3\to SG3} & p^{x}_{SG3\to SG4} \\ p^{x}_{SG4\to SG1} & p^{x}_{SG4\to SG2} & p^{x}_{SG4\to SG3} & p^{x}_{SG4\to SG4} \end{bmatrix} \tag{21}$$

Predicting 2D vs. 3D

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

Data availability

The authors declare that all data supporting the findings in this study are available within the paper and its supplementary information/documents

Code availability

Detailed descriptions of our approach and code utilized is either provided in the supplementary documents or is available through published literature.

Acknowledgement

We acknowledge financial support for this study from the National Institutes of Health U01AG060903 (JMP, JW), and the Johns Hopkins University Older Americans Independence Center of the National Institute on Aging (NIA) under award number P30AG021334 (JMP, JW). We thank Dr. Denis Wirtz for constructive discussions and feedback on this work.

Author contributions

JMP and DM conceived study, analysis and visualization; JMP, HJ, WD, AG, NZ, AJ, BS, MK, AA, DR, LV, AM, LC, JW performed experiments, collected data and contributed reagents; DM, JMP, NZ, PK performed analysis; JMP and DM interpreted results; JMP and DM wrote and edited manuscript; all authors contributed to reviewing and editing manuscript; JMP supervised study and secured funding.

Conflicts of interest

Authors declare no conflict of interest.

Bibliography

1. Efremova, M. & Teichmann, S. A. Computational methods for single-cell omics across modalities. *Nature Methods* (2020). doi:10.1038/s41592-019-0692-4

2. Zhu, C., Preissl, S. & Ren, B. Single-cell multimodal omics: the power of many. *Nature Methods* (2020). doi:10.1038/s41592-019-0691-5
3. Yamada, K. M. & Sixt, M. Mechanisms of 3D cell migration. *Nat. Rev. Mol. Cell Biol.* **20**, 738–752 (2019).
4. Swann, J. B. & Smyth, M. J. Immune surveillance of tumors. *J. Clin. Invest.* **117**, 1137–1146 (2007).
5. Choe, K. *et al.* Intravital three-photon microscopy allows visualization over the entire depth of mouse lymph nodes. *Nat. Immunol.* **23**, 330–340 (2022).
6. Friedl, P. & Wolf, K. Tumour-cell invasion and migration: diversity and escape mechanisms. *Nat. Rev. Cancer* **3**, 362–374 (2003).
7. Wirtz, D., Konstantopoulos, K. & Searson, P. C. P. P. C. The physics of cancer: the role of physical interactions and mechanical forces in metastasis. *Nat. Rev. Cancer* (2011). doi:10.1038/nrc3080
8. Phillip, J. M. *et al.* Fractional redistribution among cell motility states during ageing. *Commun. Biol.* (2021). doi:10.1038/s42003-020-01605-w
9. Ma, X., Dagliyan, O., Hahn, K. M. & Danuser, G. Profiling cellular morphodynamics by spatiotemporal spectrum decomposition. *PLoS Comput. Biol.* **14**, e1006321–e1006321 (2018).
10. Stassen, S. V, Yip, G. G. K., Wong, K. K. Y., Ho, J. W. K. & Tsia, K. K. Generalized and scalable trajectory inference in single-cell omics data with VIA. *Nat. Commun.* **12**, 5528 (2021).
11. Kimmel, J. C., Chang, A. Y., Brack, A. S. & Marshall, W. F. Inferring cell state by quantitative motility analysis reveals a dynamic state system and broken detailed balance. *PLoS Comput. Biol.* (2018). doi:10.1371/journal.pcbi.1005927
12. Du, W., Nair, P., Johnston, A., Wu, P.-H. & Wirtz, D. Cell Trafficking at the Intersection of the Tumor–Immune Compartments. *Annu. Rev. Biomed. Eng.* (2022). doi:10.1146/annurev-bioeng-110320-110749
13. Crainiciuc, G. *et al.* Behavioural immune landscapes of inflammation. *Nature* **601**, 415–421 (2022).
14. Kwok, S. J. J., Montoro, D. T., Haber, A. L., Yun, S.-H. & Vinarsky, V. Single-cell transcriptomics of dynamic cell behaviors. *bioRxiv* (2022). doi:10.1101/2022.01.31.478519
15. Phillip, J. M. *et al.* Biophysical and biomolecular determination of cellular age in humans. *Nat. Biomed. Eng.* **1**, 0093 (2017).
16. Jayatilaka, H. *et al.* Synergistic IL-6 and IL-8 paracrine signalling pathway infers a strategy to inhibit tumour cell migration. *Nat. Commun.* **8**, (2017).
17. Wu, P.-H., Giri, A., Sun, S. X. & Wirtz, D. Three-dimensional cell migration does not follow a random walk. *Proc. Natl. Acad. Sci.* **111**, 3949–3954 (2014).

18. Giri, A. *et al.* The Arp2/3 complex mediates multigeneration dendritic protrusions for efficient 3-dimensional cancer cell migration. *FASEB J.* **27**, 4089–4099 (2013).
19. Jayatilaka, H. *et al.* EB1 and cytoplasmic dynein mediate protrusion dynamics for efficient 3-dimensional cell migration. *FASEB J.* (2018). doi:10.1096/fj.201700444RR
20. Jerison, E. R. & Quake, S. R. Heterogeneous T cell motility behaviors emerge from a coupling between speed and turning in vivo. *Elife* **9**, (2020).
21. Wu, P. H., Giri, A. & Wirtz, D. Statistical analysis of cell migration in 3D using the anisotropic persistent random walk model. *Nat. Protoc.* (2015). doi:10.1038/nprot.2015.030
22. Phillip, J. M., Han, K. S., Chen, W. C., Wirtz, D. & Wu, P. H. A robust unsupervised machine-learning method to quantify the morphological heterogeneity of cells and nuclei. *Nat. Protoc.* **16**, 754–774 (2021).
23. Fraley, S. I. *et al.* Three-dimensional matrix fiber alignment modulates cell migration and MT1-MMP utility by spatially and temporally directing protrusions. *Sci. Rep.* (2015). doi:10.1038/srep14580
24. Velez, D. O. *et al.* 3D collagen architecture induces a conserved migratory and transcriptional response linked to vasculogenic mimicry. *Nat. Commun.* **8**, 1651 (2017).
25. Fraley, S. I. *et al.* A distinctive role for focal adhesion proteins in three-dimensional cell motility. *Nat. Cell Biol.* (2010). doi:10.1038/ncb2062
26. Baskaran, J. P. *et al.* Cell shape, and not 2D migration, predicts extracellular matrix-driven 3D cell invasion in breast cancer. *APL Bioeng.* **4**, 26105 (2020).
27. Lauffenburger, D. A. & Horwitz, A. F. Cell migration: a physically integrated molecular process. *Cell* **84**, 359–369 (1996).
28. You, L. *et al.* Linking the genotypes and phenotypes of cancer cells in heterogenous populations via real-time optical tagging and image analysis. *Nat. Biomed. Eng.* (2022). doi:10.1038/s41551-022-00853-x
29. Bedard, P. L., Hansen, A. R., Ratain, M. J. & Siu, L. L. Tumour heterogeneity in the clinic. *Nature* **501**, 355–364 (2013).
30. Dagogo-Jack, I. & Shaw, A. T. Tumour heterogeneity and resistance to cancer therapies. *Nat. Rev. Clin. Oncol.* **15**, 81–94 (2018).

Figure Captions

Figure 1. Data harmonization across multiple datasets. A. trajectory of a single HT1080 cell showing the effects of varying interval times. **B-C**. trajectory of a single ht1080 cells and the prediction

of cell locations per time point based on interpolation and pMC-2 for symmetric (**B**) and asymmetric (**C**) multiples of the real interval time. **D**. Mean squared displacements (MSD) of primary human dermal fibroblasts (n=873 cells) as a function of time lag. Interpolation and various pMC factors demonstrate short and long time scale behaviors relative to the MSD generated from the actual cell trajectories. **E**. Estimated eight motility parameters based on APRW model. **F**. Heatmap showing optimal time interval across all conditions.

Figure 2. Classifying cell motility patterns across short durations. **A**. 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. **B**. Line plot shows the number of clusters and the inertia/silhouette ratios. **C**. Heatmap of cluster coherence factors showing that cells classified within each cluster exhibit small distances from their cluster centroid. **D**. Two-dimensional tSNE plot showing the 17 motility clusters that further separation into 4 classes, each dot denotes a single cell. **E**. Two-dimensional t-SNE plot showing the magnitude of the eight motility parameters per cluster region, each dot denotes a single cell. **F**. Visualization of cell trajectories for cells classified within each cluster, top panel shows the origin-centered trajectories for all cells within a given cluster, bottom panel shows 16 randomly selected cells per clusters. **G-H**. Confusion matrix showing the classification predictions using linear SVM for cells 90:10 split of data (**G**) and the prediction classifications using linear SVM of 75,000 simulated cell trajectories (**H**). **I**. Bar plot showing the abundance of cells classified within each of the 17 clusters for monocytes from 3 donors (blue) and matched cells exposed to conditioned media (red). **J**. Dot plot showing the decrease in heterogeneity after exposure to conditioned media, each dot denotes a donor. **K**. Cross-correlation heatmap showing that cells exposed to conditioned media were more similar to other exposed cells then their respective baselines. **L**. Box-and-whisker plots showing the magnitude of conventional motility parameters and the increase in motility after exposure to conditioned media.

Figure 3. Classifying cell motility patters across long durations. **A**. Heatmap showing eight motility parameters per cell (n=24,753 cells). Each row represents a single cell and each column represents a motility parameter. **B**. Line plot shows the number of clusters and the inertia/silhouette ratios. **C**. Heatmap of cluster coherence factors showing that cells classified within each cluster exhibit small distances from their cluster centroid. **D**. Two-dimensional t-SNE plot showing the 25 identified clusters and their separation into 4 classes, each dot denotes a single cell. **E**. Two-dimensional t-SNE plot showing the magnitude of the eight motility parameters per cluster region, each dot denotes a single cell. **F**. Visualization of cell trajectories for cells classified within each cluster, top panel shows the origin-centered trajectories for all cells within a cluster, and bottom panel shows 16 randomly selected cells

per clusters. **G**. Confusion matrix showing prediction classification using linear SVM for cells 90:10 split of data. **H**. Sankey plot showing the transitions based on the fractional abundance of cells within each short duration class (n=3 time segments), and how they ultimately connect to the four long duration classes. **I**. Sankey plots show the conditional probabilities' magnitude for cells transitioning across short duration classes. **J**. Heatmap and complimentary box-and-whisker plots show short duration classes' stability fractions.

Figure 4. Quantifying functional heterogeneity and biological insights using CaMI. **A**. Cross-correlation heatmap highlighting the grouping of the 409 cellular conditions based on cell behavior similarities. **B**. Two-dimensional 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**. Two-dimensional t-SNE plot shows a bi-phasic/switching effect with increasing collagen concentration. **E**. Heatmap showing the abundance of cells within each cluster per collagen concentration. Line plot on the right shows the change in the Shannon entropy, denoting cellular heterogeneity. **F**. Sankey plots showing the temporal transitions of cells across short duration classes as a function of increasing concentration. **G**. Heatmap showing the average values of motility parameters with increasing collagen concentration. **H**. Two-dimensional t-SNE plot showing the location of matched 2D and 3D conditions of ht1080 cells. **I**. Box plot showing the Shannon entropy of cell in 2D and 3D environments. **J**. ROC curve showing accuracy of prediction, denoted by the true positive rates and false positive rates, for higher order motility parameters and different types of entropy. **K-L**. Box-and-whisker plots showing the prediction accuracy (**K**) and area under the curve (**L**) for higher order motility parameters and different types of entropy.

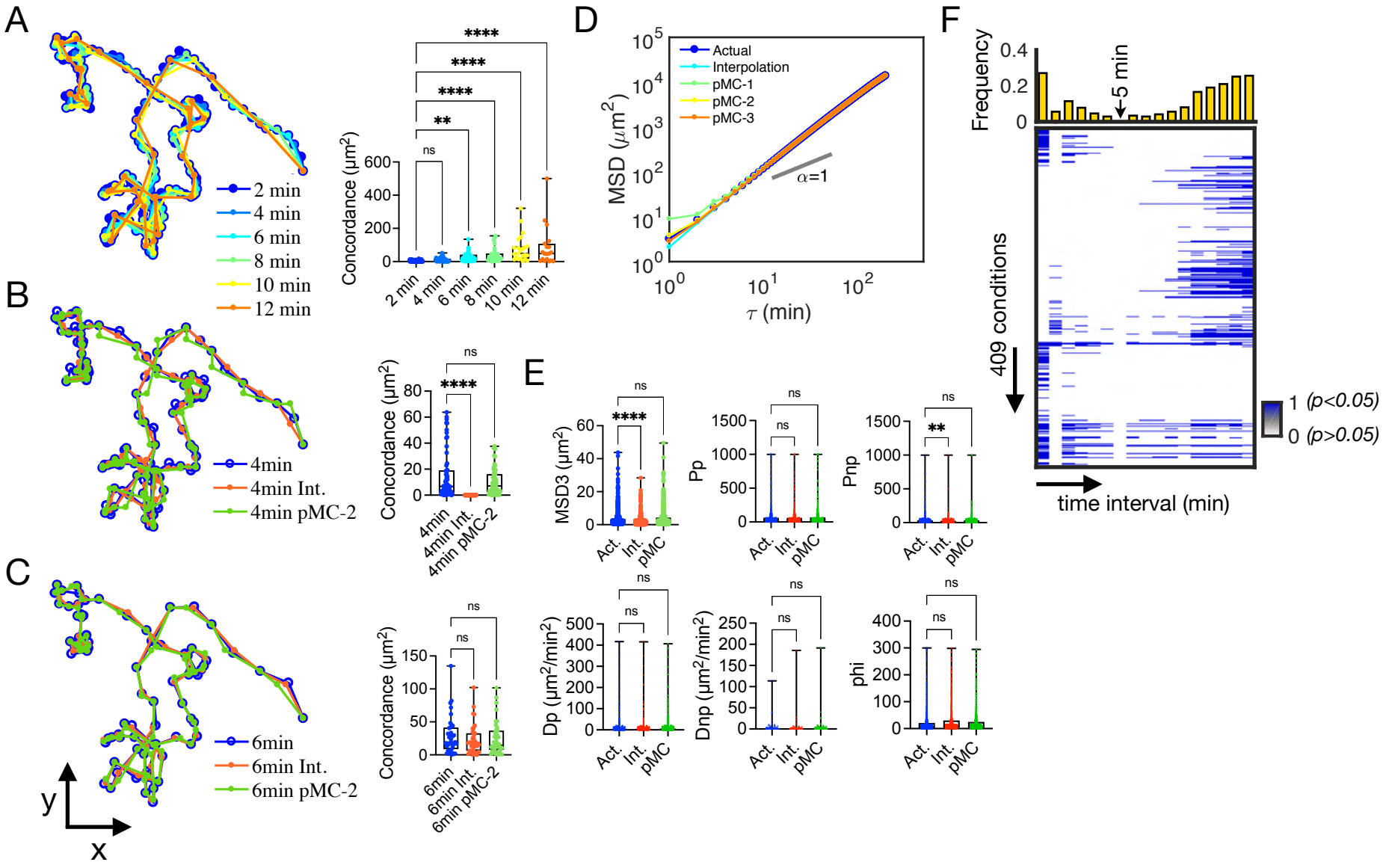
A
2 min
4 min
6 min
8 min
10 min
12 min
Concordance (μm²)
ns
**
****
****
****
2 min
4 min
6 min
8 min
10 min
12 min
B
4min
4min Int.
4min pMC-2
Concordance (μm²)
ns
****
4min
4min Int.
4min pMC-2
C
6min
6min Int.
6min pMC-2
Concordance (μm²)
ns
ns
6min
6min Int.
6min pMC-2
y
x
D
Actual
Interpolation
pMC-1
pMC-2
pMC-3
MSD (μm²)
α=1
τ (min)
E
MSD3 (μm²)
Pp
Pnp
Dp (μm²/min²)
Dnp (μm²/min²)
phi
ns
****
**
Act.
Int.
pMC
F
Frequency
5 min
409 conditions
time interval (min)
1 (p<0.05)
0 (p>0.05)

Figure 1

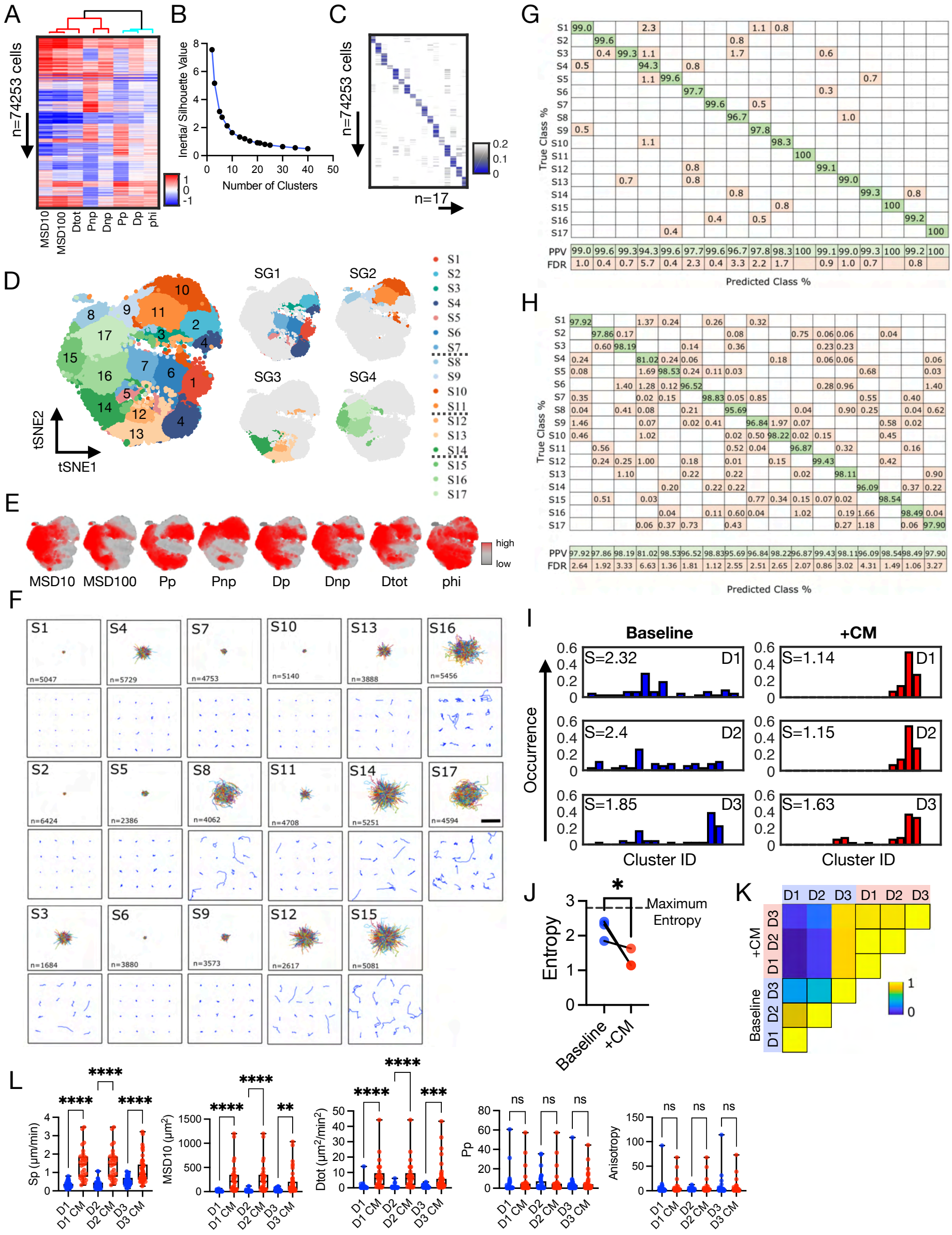
A
n=74253 cells
MSD10
MSD100
Dtot
Pnp
Dnp
Pp
Dp
phi
B
Inertia/ Silhouette Value
Number of Clusters
C
n=74253 cells
n=17
D
SG1
SG2
SG3
SG4
tSNE2
tSNE1
S1
S2
S3
S4
S5
S6
S7
S8
S9
S10
S11
S12
S13
S14
S15
S16
S17
E
high
low
MSD10
MSD100
Pp
Pnp
Dp
Dnp
Dtot
phi
F
n=5047
n=5729
n=4753
n=5140
n=3888
n=5456
n=6424
n=2386
n=4062
n=4708
n=5251
n=4594
n=1684
n=3880
n=3573
n=2617
n=5081
G
True Class %
Predicted Class %
PPV
FDR
H
True Class %
Predicted Class %
PPV
FDR
I
Baseline
+CM
Occurrence
S=2.32
S=2.4
S=1.85
S=1.14
S=1.15
S=1.63
D1
D2
D3
Cluster ID
J
Entropy
Maximum Entropy
Baseline
+CM
K
+CM
Baseline
L
Sp (μm/min)
MSD10 (μm²)
Dtot (μm²/min²)
Pp
Anisotropy
D1
D1 CM
D2
D2 CM
D3
D3 CM
ns

Figure 2

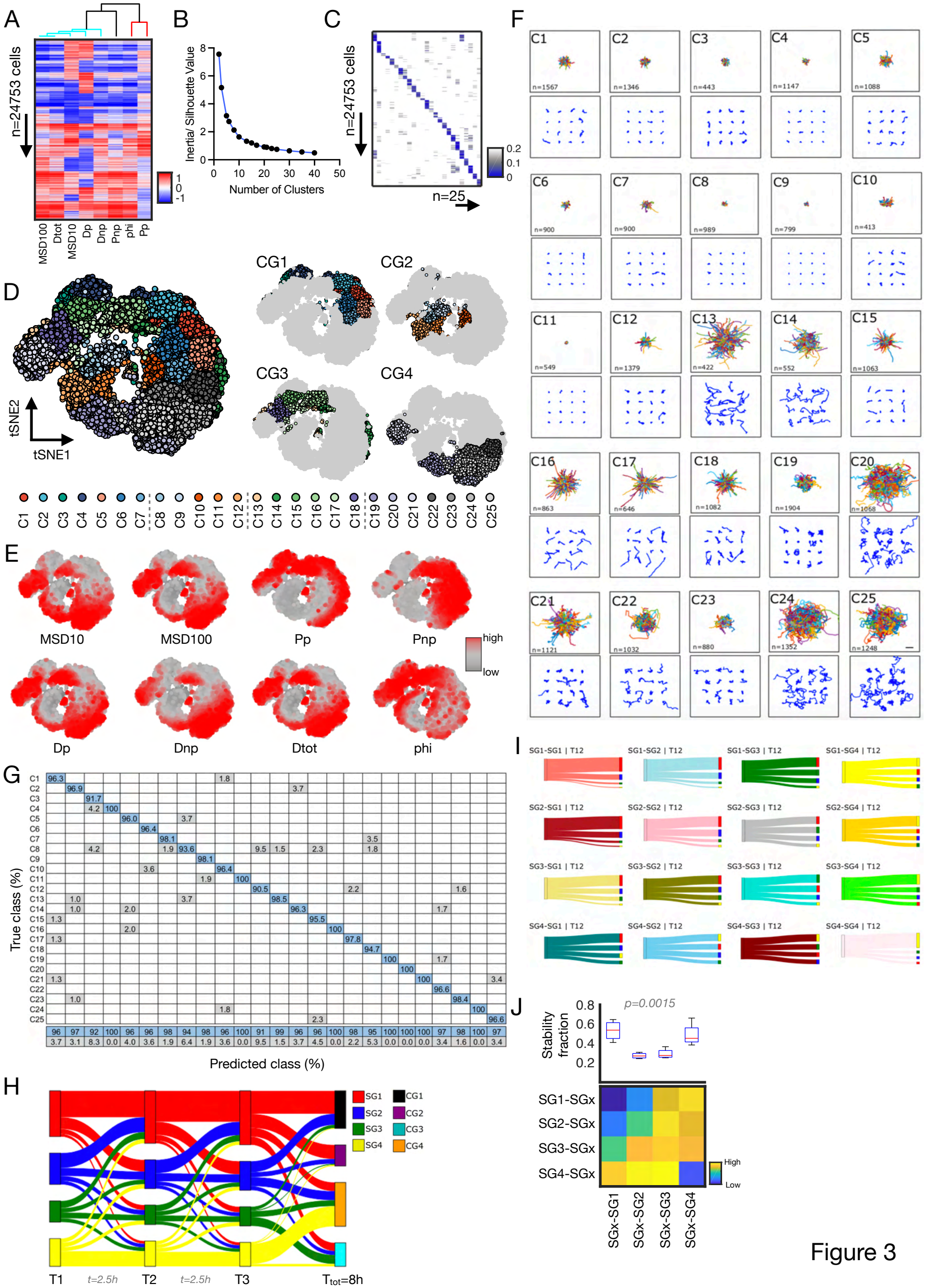
A
n=24753 cells
MSD100
Dtot
MSD10
Dp
Dnp
Pnp
phi
Pp
B
Inertia/ Silhouette Value
Number of Clusters
C
n=24753 cells
n=25
D
tSNE2
tSNE1
CG1
CG2
CG3
CG4
E
MSD10
MSD100
Pp
Pnp
Dp
Dnp
Dtot
phi
high
low
F
G
True class (%)
Predicted class (%)
H
SG1
SG2
SG3
SG4
CG1
CG2
CG3
CG4
T1
t=2.5h
T2
t=2.5h
T3
Ttot=8h
I
SG1-SG1 | T12
SG1-SG2 | T12
SG1-SG3 | T12
SG1-SG4 | T12
SG2-SG1 | T12
SG2-SG2 | T12
SG2-SG3 | T12
SG2-SG4 | T12
SG3-SG1 | T12
SG3-SG2 | T12
SG3-SG3 | T12
SG3-SG4 | T12
SG4-SG1 | T12
SG4-SG2 | T12
SG4-SG3 | T12
SG4-SG4 | T12
J
p=0.0015
Stability fraction
SG1-SGx
SG2-SGx
SG3-SGx
SG4-SGx
SGx-SG1
SGx-SG2
SGx-SG3
SGx-SG4
High
Low

Figure 3

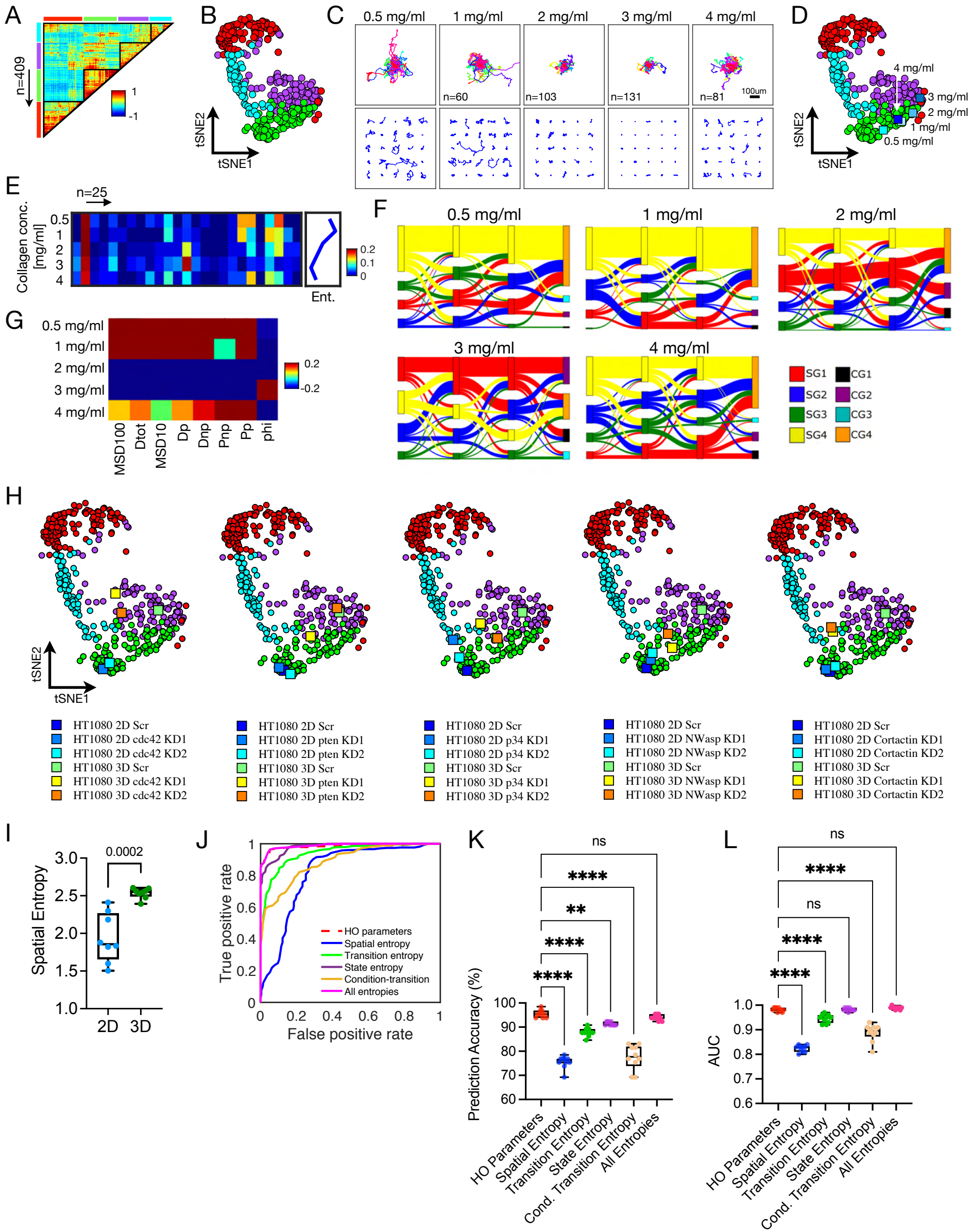
A
n=409
1
-1
B
tSNE2
tSNE1
C
0.5 mg/ml
1 mg/ml
2 mg/ml
3 mg/ml
4 mg/ml
n=60
n=103
n=131
n=81
100um
D
4 mg/ml
3 mg/ml
2 mg/ml
1 mg/ml
0.5 mg/ml
tSNE2
tSNE1
E
n=25
Collagen conc. [mg/ml]
0.5
1
2
3
4
Ent.
0.2
0.1
0
F
0.5 mg/ml
1 mg/ml
2 mg/ml
3 mg/ml
4 mg/ml
SG1
SG2
SG3
SG4
CG1
CG2
CG3
CG4
G
0.5 mg/ml
1 mg/ml
2 mg/ml
3 mg/ml
4 mg/ml
MSD100
Dtot
MSD10
Dp
Dnp
Pnp
Pp
phi
0.2
-0.2
H
tSNE2
tSNE1
HT1080 2D Scr
HT1080 2D cdc42 KD1
HT1080 2D cdc42 KD2
HT1080 3D Scr
HT1080 3D cdc42 KD1
HT1080 3D cdc42 KD2
HT1080 2D Scr
HT1080 2D pten KD1
HT1080 2D pten KD2
HT1080 3D Scr
HT1080 3D pten KD1
HT1080 3D pten KD2
HT1080 2D Scr
HT1080 2D p34 KD1
HT1080 2D p34 KD2
HT1080 3D Scr
HT1080 3D p34 KD1
HT1080 3D p34 KD2
HT1080 2D Scr
HT1080 2D NWasp KD1
HT1080 2D NWasp KD2
HT1080 3D Scr
HT1080 3D NWasp KD1
HT1080 3D NWasp KD2
HT1080 2D Scr
HT1080 2D Cortactin KD1
HT1080 2D Cortactin KD2
HT1080 3D Scr
HT1080 3D Cortactin KD1
HT1080 3D Cortactin KD2
I
Spatial Entropy
0.0002
2D
3D
J
True positive rate
False positive rate
HO parameters
Spatial entropy
Transition entropy
State entropy
Condition-transition
All entropies
K
Prediction Accuracy (%)
ns
****
**
HO Parameters
Spatial Entropy
Transition Entropy
State Entropy
Cond. Transition Entropy
All Entropies
L
AUC
ns
****
HO Parameters
Spatial Entropy
Transition Entropy
State Entropy
Cond. Transition Entropy
All Entropies

Figure 4