Single-cell morphodynamical trajectories enable prediction of gene expression accompanying cell state change

Authors: Jeremy Copperman[3*], Ian C. Mclean[1], Sean M. Gross, Jalim Singh[2], Young Hwan Chang[1,2], Daniel M. Zuckerman[1,2*], and Laura M. Heiser[1,2,3*]

Affiliations:

[1] Department of Biomedical Engineering, Oregon Health and Science University, Portland OR 97239, U.S.A.

[2] Knight Cancer Institute, Oregon Health and Science University, Portland OR 97239, U.S.A

[3] Cancer Early Detection Advanced Research Center, Oregon Health and Science University, Portland OR 97239, U.S.A.

* corresponding authors (copperma@ohsu.edu, zuckermd@ohsu.edu, heiserl@ohsu.edu)

Abstract

Extracellular signals induce changes to molecular programs that modulate multiple cellular phenotypes, including proliferation, motility, and differentiation status. The connection between dynamically adapting phenotypic states and the molecular programs that define them is not well understood. Here we develop data-driven models of single-cell phenotypic responses to extracellular stimuli by linking gene transcription levels to “morphodynamics” – changes in cell morphology and motility observable in time-lapse image data. We adopt a dynamics-first view of cell state by grouping single-cell trajectories into states with shared morphodynamic responses. The single-cell trajectories enable development of a first-of-its-kind computational approach to map live-cell dynamics to snapshot gene transcript levels, which we term MMIST, Molecular and Morphodynamics-Integrated Single-cell Trajectories. The key conceptual advance of MMIST is that cell behavior can be quantified based on dynamically defined states and that extracellular signals alter the overall distribution of cell states by altering rates of switching between states. We find a cell state landscape that is bound by epithelial and mesenchymal endpoints, with distinct sequences of epithelial to mesenchymal transition (EMT) and mesenchymal to epithelial transition (MET) intermediates. The analysis yields predictions for gene expression changes consistent with curated EMT gene sets and provides a prediction of thousands of RNA transcripts through

extracellular signal-induced EMT and MET with near-continuous time resolution. The MMIST framework leverages true single-cell dynamical behavior to generate molecular-level omics inferences and is broadly applicable to other biological domains, time-lapse imaging approaches and molecular snapshot data.

Summary

Epithelial cells change behavior and state in response to signals, which is necessary for the function of healthy tissue, while aberrant responses can drive diseases like cancer. To decode and potentially steer these responses, there is a need to link live-cell behavior to molecular programs, but high-throughput molecular measurement is generally destructive or requires fixation. Here we present a novel method which connects single-cell morphology and motility over time to bulk molecular readouts. Our model predicts gene expression from the observation of label-free live-cell imaging, as a step toward understanding and ultimately controlling cell state change.

Introduction

Uncovering how cells process microenvironmental signals to activate molecular programs that lead to changes in cell state is critical for understanding mechanisms of both normal and disease physiology. Cell state is determined by molecular and cellular composition, including genome and chromatin structure[1,2], proteomic[3] and transcriptomic levels[4], mitochondrial function[5], and metabolic activity[6]. Cell state is intrinsically mutable[7] and is influenced by various physical[8–10], and chemical[11,12] cues.

Single-cell omic analyses have provided an unprecedented catalog of cell states across both normal and diseased tissues[13,14] while spatially-resolved sequencing[15] and highly multiplexed imaging[16–18] have revealed insights into their spatial organization; however, these approaches lack single-cell time-ordered information, limiting the ability to draw mechanistic insights. Live-cell imaging, on the other hand, readily captures cellular dynamics over timescales of seconds to days, but is limited to a small number of molecular read-outs[7,19–21]. Further, analysis of live-cell data typically relies on single timepoint “snapshots” of cell morphology or fluorescently-labeled reporters[22–25]. To overcome these limitations, we recently developed a morphodynamical trajectory embedding method that leverages hidden information from time-ordered live-cell

trajectories, enabling improved prediction of future behavior as compared to single-snapshot-based predictions[26].

It is increasingly appreciated that mechanistic understanding of both normal and diseased biological systems will require consideration of cell state dynamics. Several recent methods impose a dynamical model upon static single-cell measurements to describe gene expression dynamics[27–29], including pseudo-time estimation[30,31] and RNA velocity[32,33]. In contrast, here we develop our computational framework based upon the direct observation of single-cell dynamics obtained from live-cell imaging.[34–37] In practice, we resolve a cell state landscape over hundreds of "microstates", where transitions among microstates are described in a discrete-time Markov model framework[38,39]. Our data-driven modeling approach extends other efforts based on live-cell imaging and trajectory analysis[40,41] by characterizing cell state changes quantitatively observed in live-cell imaging data, yielding distinct states that can be linked to molecular programs observed in companion profiling data.

We explore ligand-induced cell state changes in MCF10A cells via Molecular and Morphodynamics-Integrated Single-cell Trajectories (MMIST), a novel computational methodology integrating live-cell imaging-observed dynamics and gene expression profiling. Broad efforts have been undertaken to map cell morphology and motility to gene expression states[22,42–46], including generative machine learning approaches[47,48]. The central and novel element of MMIST is a mapping between morphodynamical "states" defined using live-cell features measured over time, and companion bulk RNAseq data. The mapping exploits a linear population-matching approach[34–37] based upon the premise that live-cell imaging and bulk molecular profiling share commonly identifiable cell states with shared average transcript levels specific to each state. In this framework, treatment conditions induce changes in state-switching rates leading to differences in state populations, which are directly measured from live-cell imaging. Given this shared set of states across ligand-treatment conditions, state-specific transcription levels are readily computed in a linear algebra framework.

We developed our approach by focusing on the well-characterized human mammary epithelial MCF10A cell line[49,50], a non-transformed cell line that recapitulates key features of epithelial biology, including migration[51,52] and organoid formation[53,54]. It is also easily manipulated in a variety of assays including live-cell imaging[55], knock-down[50], and chemical perturbation[56] and therefore is commonly used for cell biology studies. Prior studies have used MCF10A cells to

probe epithelial responses to growth factors and cytokines[57] and to uncover molecular programs associated with EMT[58–64].

We further focus on cellular and molecular responses to Transforming Growth Factor Beta (TGFB) as an illustrative example and demonstrate the quantitative linkage of EMT-associated live-cell phenotypic responses with EMT molecular programs that we validated in an independent dataset[65]. In total, our novel data-driven modeling approach captures cell state change along sequences of cell state intermediates via live-cell and gene expression phenotypes and enables linkage of imaging and molecular data to uncover molecular correlates of distinct morphodynamic cell states.

Results

Experimental data to facilitate multimodal integration of morphodynamical and gene expression measurements of cell state change

Our method is designed to infer molecular programs associated with distinct cell states by linking morphodynamic measurements acquired in live-cell imaging data to companion snapshot molecular data. We analyze a recently published LINCS MCF10A ligand perturbation dataset[57] that consists of paired live-cell imaging and bulk transcriptomic measurements of MCF10A cells after treatment with multiple ligands, including Epidermal Growth Factor (EGF), Transforming Growth Factor Beta (TGFB), and Oncostatin M (OSM), which are the focus of our analysis. We also leverage live-cell image data and transcriptomic measurements of MCF10A cells genetically engineered to express nuclear and cell cycle reporters[66].

Our computational framework, illustrated in **Figure 1**, leverages companion live-cell image stacks and bulk RNAseq as input, and utilizes statistical physics approaches to yield maps of cell states and their transition sequences[67,68]. Here we outline the major steps. (a) First, we analyze the live-cell image data to identify cell nuclei by training a virtual nuclear reporter[69] on paired phase contrast and nuclear reporter images, then virtually stain nuclei in the entire dataset (**Supplementary Figure 1**). We initially “featurize” individual cells to quantify cell shape and texture and perform local environment featurization using geometric boundaries based on the nuclear centers. We have validated the qualitative performance of the MMIST pipeline using standard cytoplasmic masks obtained using cellpose software on a subset of the data (**Supplementary Figure 2**). We track individual cells across images with Bayesian belief

propagation[70] and compute motility as cell displacement between frames. Local alignment of cell motility is featurized via a neighborhood averaging approach (**Supplementary Figure 3**). Assessment of segmentation and tracking performance is provided in **Supplementary Data Table 1**. (b) Dynamical cell features are constructed based on trajectory snippets (all possible single-cell sub-trajectories of a particular length in a sliding window manner, e.g., frames 1-2-3, 2-3-4, 3-4-5, …) utilizing our morphodynamical trajectory embedding methodology[26]. Thus, the features used for analysis are morphodynamical trajectory snippets, quantified as time-ordered lists of morphological features. (c) Morphodynamical trajectory snippets are used to build a data-driven dynamical model of cell states. (d) Cell states observed in the image data are mapped to gene transcript levels using linear decomposition. This yields temporal sequences of morphodynamical cell state changes and their associated gene expression levels.

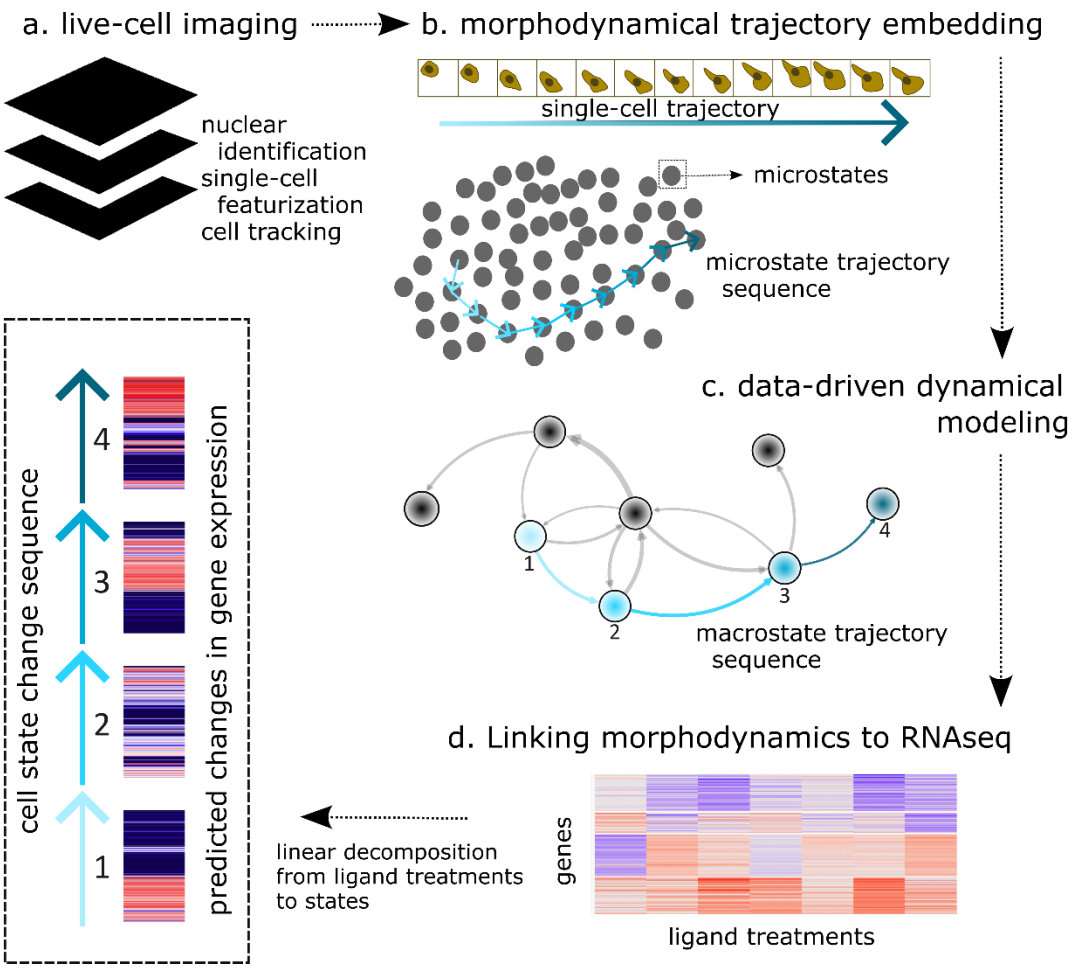


Figure 1: MMIST approach to link live-cell imaging to molecular read-outs.
a) Live-cell imaging of MCF10A cells after treatment with a panel of microenvironmental ligands. Nuclei are identified using a convolutional neural network, and single-cells are featurized and tracked through time. b) Single-cell features are concatenated along single-cell trajectories to construct the morphodynamical trajectory embedding. c) Dynamical models learn cell states and cell state change sequences in the morphodynamical landscape. d) Cell state populations are used as a linear decomposition of bulk gene expression measurements to predict the gene expression programs underlying cell state change.

Single-cell trajectories define morphodynamical cell states

Morphodynamical states are the foundation of our analysis, and we define these states based on groupings of cells with shared dynamical progression—i.e., those that cluster together based on a similar time progression of shape, texture, and motility features. To achieve this, we employ feature vectors that are a time-concatenation of image-based features[26]; this is termed the 'morphodynamical trajectory space.' In this space, we place hundreds of "microstate" centers via clustering. We then count transitions among microstates to build a data-driven transition matrix Markov model of cell state progression[38,39]. Next, microstates are grouped into coarser

“macrostates” using a spectral clustering procedure[71,72]. We refer to these macrostates as morphodynamic states, or simply states. The eigenfunctions of the dynamical model represent dynamical motifs, which we visualize using UMAP[73] dimensionality reduction to facilitate interpretation of cell states (**Figure 2a**). The relationships between different microstates are captured by ”flow”, the average transition direction computed from the Markov model, and denoted by the small arrows in the UMAP. We clustered the microstates and identified 14 discrete morphodynamic states, each with characteristic features (**Figure 2a, lower panels**). with representative single-cell images indicating that the formation of multicellular clusters being a driver of this partitioning. We examined the relationship between morphodynamic states and ligand treatment and found that ligands induce unique cell state changes and transition flows between macrostates, indicating induction of different cell state populations over time in response to perturbation (**Figure 2b**). For example, OSM drives transition flow towards state 10, consisting of highly motile multicellular clusters, whereas TGFB drives transition flow towards state 5, consisting of more isolated cells with extended lamellopodia-like cytoskeletal features. The complete set of ligand-dependent cell state populations are shown in **Supplementary Figure 4c**, and population distributions and transition flows are shown in **Supplementary Figure 5**.

The derived states resolve differences in morphodynamical properties, including cell-cell contact fraction, local alignment of cell-cell motility, motility speed, and cell-cycle phase; with differing contact fraction and motility alignment organizing the major differences across the landscape (**Figure 2c-f**). We quantified these morphological properties for each cell state individually to highlight state-to-state differences (**Figure 2g-j**). We can further query the states in relation to one another, referenced to states 5 and 10 at the extremes of the morphodynamical cell state space. State 5 is characterized by mesenchymal-associated features such as lower local alignment of cell motility, more extended cytoskeletal features, greater cell spreading, and an extended G1 cell cycle duration (**Figure 2j**); this state population increases under TGFB containing treatments. In contrast, state 10 is characterized by many epithelial-associated features, including increased multicellular clustering and collective motility, which are increased after treatments that include OSM; these represent an altered epithelial phenotype that maintains epithelial-associated characteristics.[57] Between these two states, we observe intermediate states with short cell cycle duration (**Figure 2f,j**), increased motility (**Figure 2e,i**), and the fewest cell-cell contacts (**Figure 2c,g**). Under EGF treatment,[49] cells transition between these intermediate states (**Figure 2b**). Thus, based upon morphodynamical features, the derived cell state space matches the well-described biological framework of epithelial and mesenchymal cell states[74],

including extended G1 duration in the mesenchymal state[75–78]. Below, we discuss these observations in the context of EMT.

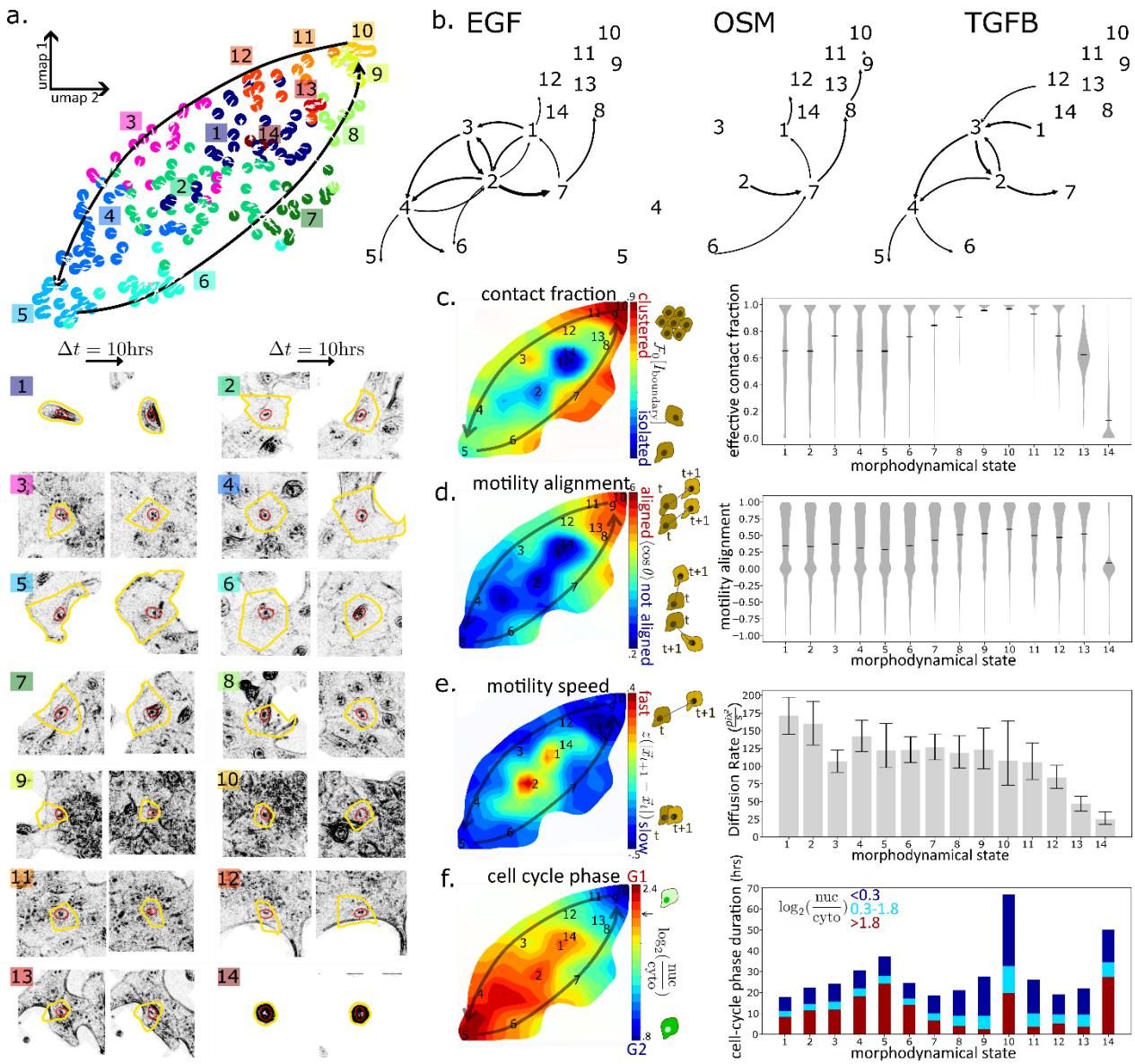


Figure 2: Data-driven models define morphodynamical cell states and state transition dynamics.

a.) The dynamical embedding landscape is visualized via UMAP from 200 microstates (dots) constructed from morphodynamical trajectories (trajectory snippet length = 10H), and average flows from each state (gray arrows), colored and labeled by cell state groupings, i.e., numbered cell "states"; large black arrows are guides to the eye for significant macroscopic flow paths. Lower panels show images from first and last frames of representative trajectory snippets (10H trajectory length) from each state with nuclear segmentations (red contours) and associated Voronoi segmentation (yellow contours). b.) Cell state flow (at t=24H) by ligand treatment. c.-f.) Cell morphology, motility, and cell cycle features by morphodynamical cell state. Panels (g) and (h) show violin-plot distributions of single-cell values, (i) shows average behavior with uncertainty based on single-cell variation, and (j) shows modeled cell-cycle phase durations averaged over single-cell behavior.

Morphodynamic trajectories reveal transcriptional dynamics via state mapping across modalities

Motivated by the observation that the morphodynamic cell states recapitulate aspects of EMT, we next sought to identify the underlying molecular programs associated with cell state transitions. The process relies on having both morphodynamical observations and molecular measurements for an identical set of experimental treatments to enable their linkage. The primary assumption – which can be considered a hypothesis being tested – is that an observed morphodynamical state corresponds to the same RNA levels regardless of the ligand treatment. Consider the example of linking RNA levels to two distinct states (motile and non-motile), where the cell state frequencies are modulated by ligand treatment. If ligand A induces an increase in the motile cell state population as compared to ligand B and *also* higher RNA levels for gene X, then we infer that motility is linked to expression of gene X. This qualitative idea can be made exact in a simple linear algebra framework by decomposing each measured average transcript level as a linear sum over morphodynamical state populations (**Supplementary Figure 4c**) and gene expression profiles.

We validated the linear population matching approach by assessing its capability to predict withheld gene expression levels in ligand combination conditions. The method requires the same number of states as training conditions (ligand treatments), so we performed a new clustering into 10 morphodynamical cell states; specifically, we withheld the OSM+EGF, EGF+TGFB, and the triple combination OSM+EGF+TGFB RNA-seq data from the training set and then predicted gene-expression profiles for these three conditions. The morphodynamical cell state populations from the live-cell imaging in the withheld test set treatments, combined with morphodynamical cell state decomposed gene expression levels from the training set, enable prediction of the RNA-seq in the test set, i.e., expression level for every gene with globally similar patterns of gene expression in measured and reconstructed sets (**Figure 3a**). We assessed the predictive capability of the model at multiple trajectory lengths by computing the Pearson correlation between measured and predicted differential gene expression for 13,516 genes (**Supplementary Data Table 2**). Predictions are maximized at a trajectory length of 10 hours ($r > 0.7$, p-value<0.001, upper-tailed test). The predictions out-perform >99% of null model-predictions based on randomized state populations (**Figure 3b**).

Performance exceeding the random null model demonstrates that the defined states exhibit treatment-independent association with the inferred expression levels. These findings provide support for the validity of our approach to link morphodynamical states observed in image data to companion molecular measurements.

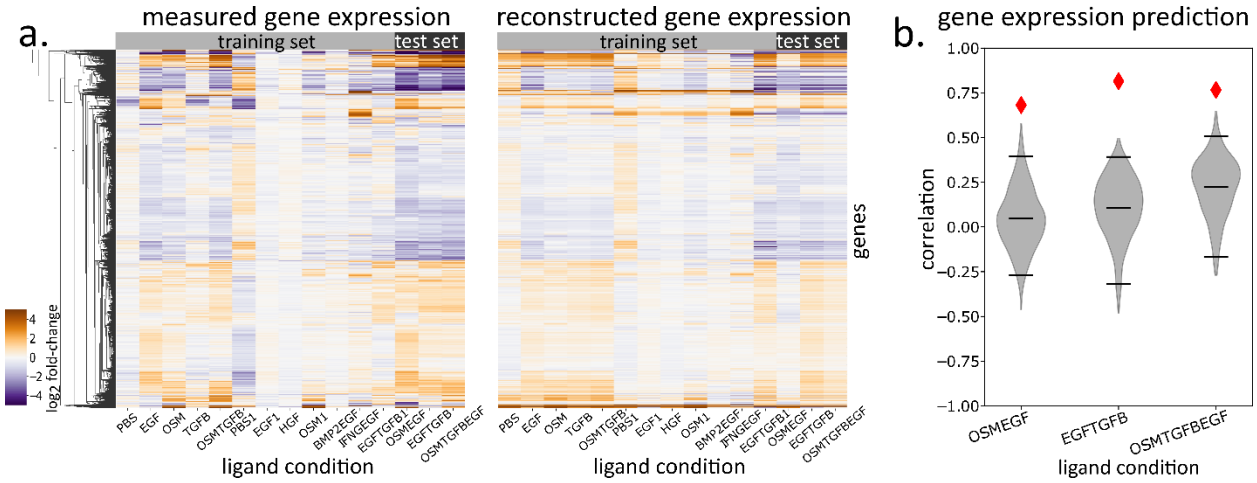


Figure 3: Morphodynamical cell states predict global gene expression patterns. a.) Validation of model gene expression predictions: measured and model-reconstructed gene expression at 24hrs for every experimental condition, including training set (light gray) and test set conditions. b.) Correlation between measured and model-predicted gene expression (red diamonds), and null estimates using random state populations (gray violin plots). Horizontal lines are the mean, 5th, and 95th percentile of the null distribution.

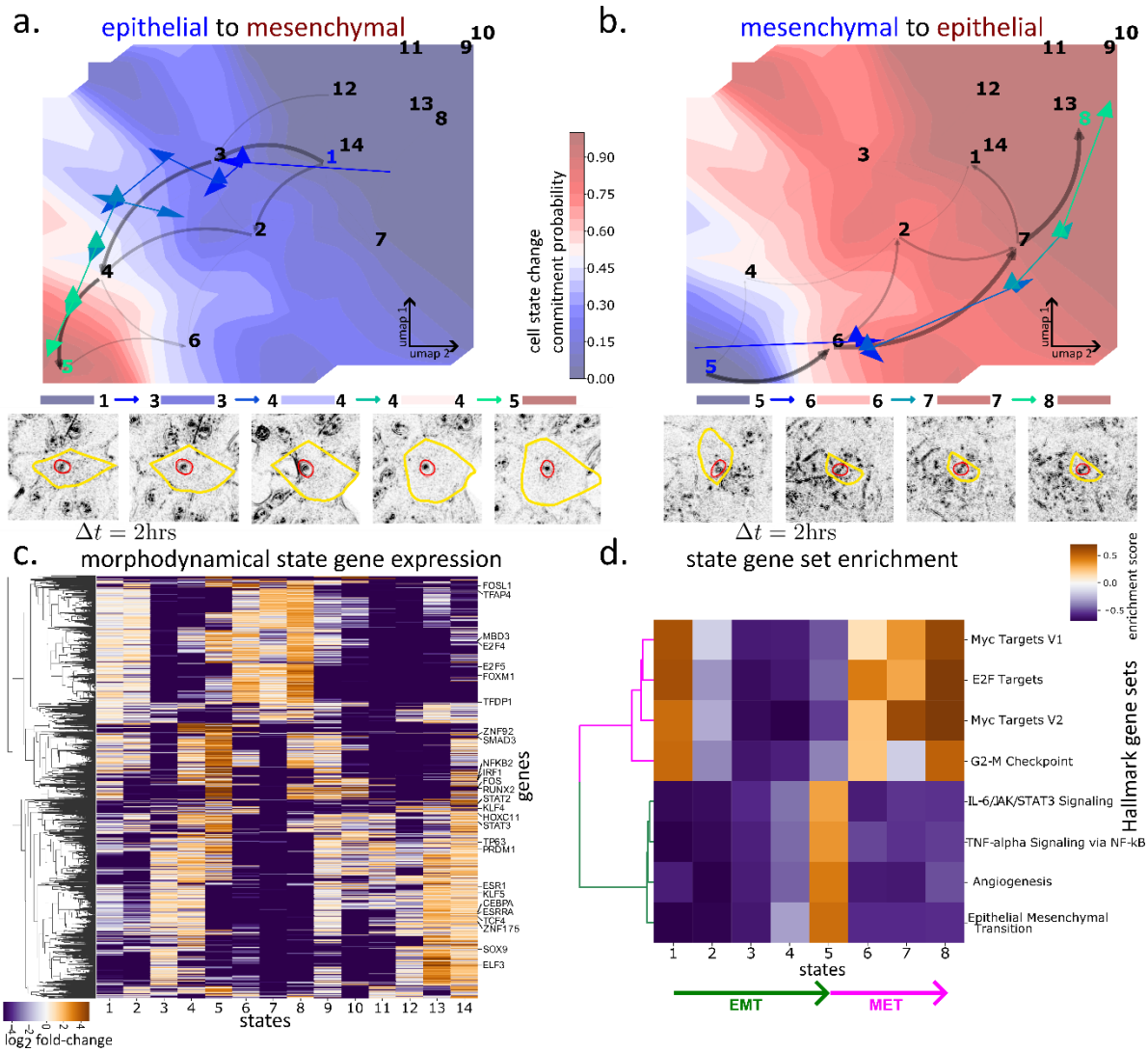
MMIST identifies ligand-induced EMT and MET morphodynamical cell state change sequences

As an illustrative use case of MMIST, we next focused on morphological features associated with canonical epithelial and mesenchymal cell states. Analysis of the morphodynamical cell states revealed features associated with epithelial and mesenchymal states, including changes in cell-cell motility alignment and cell clustering as described above (**Figure 2c-j**). These findings are consistent with the observation that microenvironmental signals can strongly modulate differentiation state of MCF10A cells—for example to form epithelial-differentiated multicellular acinar structures in 3D cultures[79] or to promote a mesenchymal phenotype under TGFB treatment[60,61,64,65]. We used our framework to examine the relationship between these states and the influence of microenvironmental signals in mediating transitions between them, which builds on prior studies of epithelial-mesenchymal transition (EMT) and mesenchymal-epithelial transition

(MET)[27,41,60,65]. To study EMT in our framework, we assigned state 1 as the most highly populated state at the initial trajectory time window (10 hours), while state 5 was assigned as mesenchymal due to its morphological features and enrichment in the TGFB condition. We set the most highly populated state at the initial time window as the initial state to facilitate identification of the most common ligand-induced state transitions ending in the mesenchymal-like state 5. For MET, we assign state 8 as the final state because it is enriched over time in the control EGF treatment as cells reach confluence (**Supplementary Figure 6**).

We observe robust but distinct state transition sequences for the EMT and MET transitions, consistent with a highly driven nonequilibrium system[80,81]. For EMT the dominant sequence of states is (1,2,3,4,5) while for MET, the primary sequence is (5,6,7,8), shown in **Figure 4a,b**. Transitions back to state 1 are common in most treatments (**Figure 2b** and **Supplementary Figure 5**). The dominant sequences of state changes are robust across ligand treatments, though the probability of specific state-to-state transitions varies. For instance, OSM treatment drives most cells towards dense and collectively migrating epithelial-like clusters (state 10), but for the rare cells that do reach state 5 from state 1, the dominant sequence of states remains the same (**Supplementary Figure 7**).

MMIST revealed unique expression patterns associated with each morphodynamical cell state (**Figure 4c**). We performed gene set enrichment over the Hallmark gene sets[82] on the derived morphodynamical state gene expression profiles. The morphodynamical state-decomposed gene expression along the EMT state change sequence shows a transition from a proliferative program enriched for Hallmark Myc Targets V1 and V2, E2F targets, and G2-M transition, to a mesenchymal program enriched for IL4/JAK/STAT3, TNFA via NFKB, Angiogenesis, and Epithelial to Mesenchymal Transition (**Figure 4d**). This switch from a proliferative program to a mesenchymal gene expression program augments our observation that cell-cycle phase durations co-varied with mesenchymal-like features observed in the live-cell data (**Figure 2j**).



Near-continuous gene expression time evolution prediction during TGFB-driven EMT

MMIST yields near-continuous time evolution of morphodynamical cell state populations via a Markov model calibrated by counting transitions between microstates extracted from single-cell trajectories. For example, EGF+TGFB leads to an increase in mesenchymal-like states 4, 5, and 6, whereas these states are decreased after EGF-only treatment (**Figure 5b**). Model predictions were assessed by comparison to state population proportions measured in live-cell imaging as a function of time after EGF+TGFB treatment. The model recapitulates the morphodynamical state population trends observed in the live-cell imaging data, supporting the validity of our Markov

assumptions (**Figure 5b**), notably, decreases in the populations of cell states 7 and 8, and increases in states 4, 5, and 6.

Our computational framework also enables a prediction of gene transcript levels at the same near-continuous time intervals (dt=30min) as those measured in the live-cell image data. Conceptually, we make these predictions by first associating each morphodynamical state with a full gene expression profile (including all genes measured in RNAseq) and then predicting the bulk gene expression over time by computing the weighted sum of the gene expression profiles for states observed in each treatment condition (**Figure 5c**), where genes are ordered the same as in **Figure 4c** so that the temporal enrichment of a gene expression profile similar to state 5 is visible. Under EGF+TGFB, our model predicts a continuous shift in multiple gene programs, including decreases in proliferation-associated programs and increases in mesenchymal-associated programs (**Figure 5d**). These changes are quantified based on average expression over the set of genes contained in the Hallmark gene sets[82] with statistically significant enrichment in the morphodynamical state decomposition analysis (**Figure 4d**)

MMIST can also be used to predict future, unmeasured shifts in cell state populations. For example, the model predicts large shifts in state populations between 0-48H, which we observed experimentally; however, it also predicts continued subtle shifts in state populations beyond the 48H duration of the experiment (**Figure 5e**). We next assessed the ability of our model to predict unseen changes in gene expression programs. Here, we trained our model with RNA-seq data collected at 24H post-treatment, then used it to predict gene expression profiles at 48H based on the predicted morphodynamical state populations shown in **Figure 5e**. We assessed our predictions by computing the correlation between experimentally measured and predicted expression profiles, after normalizing to t=0H. The correlation between predicted t=48H gene expression profile and the withheld t=48H RNAseq data is ~0.5 (**Figure 5f** and **Supplementary Figure 4**). In contrast, t=24H and t=48H experimentally measured RNAseq profiles show correlations of ~0.25, indicating that MMIST predictions can capture molecular programs associated with morphodynamic state change.

The transcriptional programs associated with TGFB-driven EMT have been previously investigated in MCF10A cells, and datasets generated through these efforts provide a useful tool for independent validation of our model[60,61,64]. To evaluate the EMT-associated signature extracted via our morphodynamical analysis, we compare our results to a recently published,

independent, time-resolved gene expression dataset of MCF10A cells treated with EGF+TGFB then harvested for molecular profiling at multiple timepoints including 24, 48, 72, and 96 hours post-treatment, ("PAMAF" data)[65]; this dataset lacked companion live-cell image data. We first assess the biological significance of the model-assigned morphodynamic states based on gene expression levels, finding positive correlation between PAMAF measurements and mesenchymal morphodynamical cell states 4 and 5 after EGF+TGFB treatment (**Figure 5g**). Consistent with this, epithelial states 6, 7, and 8 are among the least correlated. Together, these findings provide support for the robustness of MMIST to identify meaningful biological signals that can be validated in independent data sets.

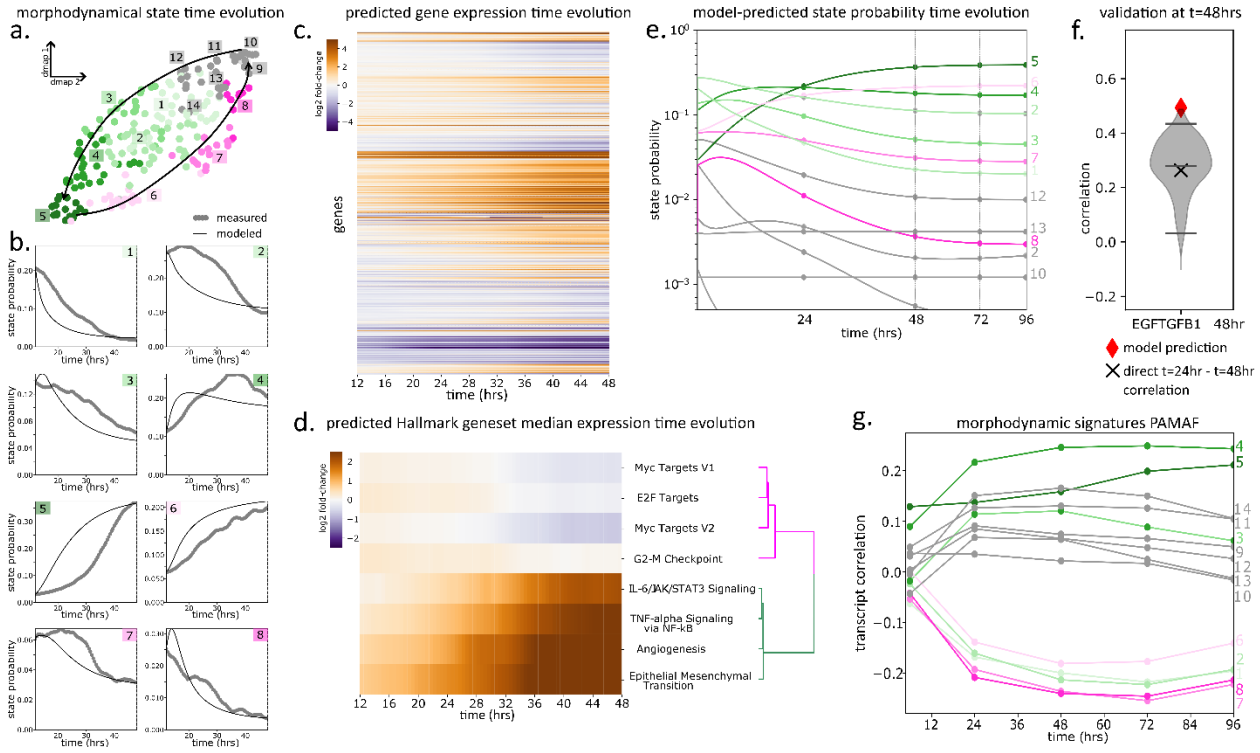


Figure 5: Morphodynamical model predicts EGF+TGFB-induced EMT gene expression time evolution.
When predicted RNA levels of morphodynamic states are integrated with Markov model dynamics, an array of dynamical omics predictions can be made, shown here for the EGF+TGFB condition. a) Morphodynamical states, which are numbered 1-12 and color-coded (mesenchymal: green, epithelial: purple). Color labels for the states are consistent throughout figure. b) State probability time evolution, measured (grey dots) and model-derived (black lines), with y-axis limits set for each plot so small changes in state populations are visible. c) Prediction of gene expression over time at 30-minute intervals using morphodynamical state prediction and live-cell imaging measured state probabilities, with rows ordered identically to Figure 4c,and d) summarized to Hallmark gene sets. e) Model-predicted state probability time evolution over 96 hours, trained from live-cell imaging over 48 hours. f) Correlation between measured and model-predicted gene expression at t=48H (red diamond) based on training data from t=24H, relative to null models with random state probabilities (gray distribution). Also shown: correlation between t=24H and t=48H gene expression (black X). g) Correlation between predicted morphodynamical state gene signatures and PAMAF measurements out to 4 days.

Discussion and Conclusion

Single-cell sequencing and spatial omics methods have provided detailed molecular profiling of cellular heterogeneity in single time-point snapshots[15]. However, there are not yet experimental methods that yield time-resolved molecular profiles with a similar level of detail. RNA velocity and other algorithms attempt to infer dynamics from fixed measurements[27–32], but they lack a direct

mapping to observed single-cell dynamics. Here we have presented a step in the direction of linking live-cell dynamics to deep molecular profiling, capturing sequences of morphodynamical cell state changes mapped to comprehensive gene expression profiles. Our approach provides a direct map from live-cell derived single-cell dynamics to gene expression for a small number of morphodynamical cell states defined through assessment of perturbation responses.

We utilized a paradigm of cell behavior in which individual cells transition between different morphodynamic states with treatment-specific dynamics and state frequencies. Thus, we employ a trajectory space that is common to all observed experimental treatments, where ligand perturbation alters the rates of cell state changes. This report demonstrates the value of this paradigm, as it enables mapping of complex, spatiotemporal phenotypes to gene transcript levels. One limitation of the present model is that it is restricted to the range of behaviors observed for a particular cell type (MCF10A) under the treatments examined and does not represent a comprehensive assessment of all possible cell states. Thus, the derived (coarse-grained) dynamical models are incomplete. As live-cell information increases, for instance via the incorporation of multiplexed live-cell reporters and deep-learning based image featurization[44,45,83–85], integration with fixed single-cell and spatial omics profiling at endpoint may require a separation of shared information across cell populations from unique information to each single-cell[86].

From a physical theoretical point of view, the transition mechanism of a dynamical process is defined via the set of trajectories connecting two states of interest[67,87–89], for instance epithelial and mesenchymal cell states. The single-cell trajectory set that connects these basins contains the set of intermediate transition states[40]. Here, we have captured sequences of EMT and MET intermediates, consistent with the emerging view of epithelial and mesenchymal states as a continuum[13,27]. It is an open question of whether characterization of transition intermediates will yield insight into cell state control, which could inform the control of EMT-driven processes during development or disease progression, such as tumor invasion[90,91]. Future studies could extend our findings by employing inhibitor or gene knockout approaches to functionally assess EMT transition intermediates predicted to be critical for cell state control.

Cell state biomarkers can predict sensitivity to targeted drugs[92,93], and are expressed in a spatially organized manner in both healthy and diseased tissues[94,95]. Morphodynamical cell state definitions can expand upon known biomarker-based cell states, providing a prediction of the dynamical responses to biological manipulation. We expect that the linking of morphodynamics

to gene expression changes, in spatial context, will lead to a deeper understanding and control of cell state change in complex tissue and tissue-like environments.

Characterization of the transition mechanism via live-cell image-based trajectories, such as we have presented, is not a mechanistic explanation at the molecular level. Time-ordered single-cell trajectories of the quantity of molecular species, such as gene transcripts, imply but do not prove causality. We speculate that utilizing molecularly detailed single-cell trajectory data to constrain mechanistic models could provide predictions of causal molecular relationships that could be experimentally validated. Our data-driven approach, as presented, does not yield a prediction for unmeasured perturbations, for instance response to different ligands or drugs. We speculate that mechanistic models[96–99], trained using the type of detailed trajectory data at the molecular level we have presented here, may enable prediction of cell behavior in unseen contexts.

Live-cell phenotypic response to ligand perturbation is well-described by our single-cell morphodynamical trajectory-based data-driven modeling approach and enabled a mapping between live-cell phenotype and time-dependent gene expression changes. Our models yielded a validated prediction of near-continuous gene expression levels during ligand-driven EMT/MET in MCF10A cell culture. MMIST can be applied generally to characterize cell state changes in fundamental biology and, potentially, in various disease settings.

Methods

MCF10A Cell Culture Companion RNAseq and live-cell imaging data used in this paper are described in full in Gross, et al[57]. In brief, MCF10A cells were cultured in growth media composed of DMEM/F12 (Invitrogen #11330-032), 5% horse serum (Sigma #H1138), 20 ng/ml EGF (R&D Systems #236-EG), 10 µg/ml insulin (Sigma #I9278), 100 ng/ml cholera toxin (Sigma #C8052), 0.5 µg/ml hydrocortisone (Sigma #H-4001), and 1% Pen/Strep (Invitrogen #15070-063). For all ligand response experiments, cells were seeded in growth media in collagen-coated well plates and allowed to attach for 6-hours. Cells were then washed with PBS, and growth media was replaced with growth-factor free media lacking EGF and insulin. After an 18-hour incubation, cells were treated with ligands in fresh growth-factor free media. Seven ligand conditions were tested at concentrations determined to elicit maximal cell responses[57](10 ng/ml EGF (R&D Systems #236-EG), 40 ng/ml HGF (R&D Systems #294-HG), 10 ng/ml OSM (R&D Systems #8475-OM), 20 ng/ml BMP2 (R&D Systems #355-BM) + 10 ng/ml EGF, 20 ng/ml IFNy (R&D Systems #258-IF) + 10 ng/ml EGF, 10 ng/ml TGFβ (R&D Systems #240-B) + 10 ng/ml EGF). Cells were then

subjected to live-cell imaging (imaged every 30 minutes for 48 hours), followed by bulk RNAseq. Here these samples are specified by appending a “1” to the treatment condition (e.g. EGF1) in order to distinguish these treatments from the cell cycle reporter studies.

Cell cycle reporter studies To assess cell-cycle responses to ligand treatments, MCF10A cells were genetically modified to stably express the HDHB cell-cycle reporter[66] and a red nuclear reporter. These MCF10A cells were a gift from Gordon B. Mills and from the same aliquot as the LINCS MCF10A dataset described above[57]. The methodology used to generate the reporter cell line has been described previously[100]. Reporter cells were treated with ligand (EGF 10 ng/ml (R&D Systems #236-EG), OSM 10 ng/ml (R&D Systems #8475-OM), TGFB 10 ng/ml (R&D Systems #240-B), EGF 10 ng/ml + OSM 10 ng/ml, TGFB 10 ng/ml + EGF 10 ng/ml, OSM 10 ng/ml + TGFB 10 ng/ml, TGFB 10 ng/ml + EGF 10 ng/ml + OSM 10 ng/ml)and then imaged every 15 minutes for 48 hours with an Incucyte S3 microscope (1020x1280, 1.49 μm/pixel). Three channels were collected – phase contrast, red (nuclear) and green (cell-cycle) – for four fields of view per well. The initial frame coincided with the addition of the ligands and fresh imaging media. [5757]

RNA-seq analysis Detailed description of sample preparation, processing, and alignment can be found in Gross et al[57]. For each ligand treatment, we performed a differential expression analysis from time 0H controls on the RNA-seq gene-level summaries with the R package DESeq2 (1.24.0), with shrunken log2 fold change estimates calculated using the apeglm method. We applied a minimum expression filter such that log2(TPM)>0.5 in at least 3 measurements over treatments and replicates (with TPM transcripts per million), yielding 13,516 genes with measured differential expression from control used in our analysis.

Image preprocessing Foreground (cells) and background pixel classification was performed using manually trained random forest classifiers using the ilastik software[101]. Images were z-normalized (mean subtracted and normalized by standard deviation). In cell images, absolute values of these z-normalized pixel values are shown (white to black). Image stacks were registered translationally using the pystackreg implementation of subpixel registration[102].

Nuclear segmentation A convolutional neural network was trained to predict the nuclear reporter intensity from the matched phase contrast images for imaging data of WT MCF10A cells with no nuclear reporter. In the EGF, OSM, and TGFB conditions, 4 image stacks (12 total) were used to train the FNET 3D reporter prediction convolutional neural network CNN from the Allen Cell

Science Institute[69], with time as the third dimension. This trained CNN was then used to predict nuclear reporter channel from the bright-field image over all image stacks in datasets. See Supplementary Figure 1 for representative nuclear reporter prediction and comparison to ground truth. Nuclear segmentations were generated by performing a local thresholding of the image within 51 pixel-sized windows at 1 standard deviation of intensity. Segmentations were filtered for a minimum size of 25 pixels and a maximum size of 900 pixels, see **Supplementary Table 1** for segmentation performance. To capture features including the local environment around a single nucleus, the image was partitioned into Voronoi cells around each nuclear center, with background classified pixels removed.

Cell featurization Single-cell featurization was performed on the Voronoi-partitioning of the image by nuclear center. Cell features are described in detail in Copperman et al.[26] and repeated here for convenience. Morphology features were obtained as follows: segmented cells were extracted, and mask-centered into zero-padded equal sized arrays larger than the linear dimension of the biggest cell (in each treatment). Principal components of each cell were aligned, and then single-cell features were calculated. Zernike moments (49 features) and Haralick texture features (13 features) were calculated in the Mahotas[103] image analysis package. The sum average Haralick texture feature was discarded due to normalization concerns. Rotation-invariant shape features (15 features) were calculated as the absolute value of the Fourier transform of the distance to the boundary as a function of the radial angle around cell center[104], with the set of shape features normalized to 1. The cell environment was featurized in a related fashion. First, an indicator function was assigned to the cell boundary with value 0 if the boundary was in contact with the background mask, and value 1 if in contact with the cell foreground mask. The absolute value of the Fourier transform of this indicator as a function of radial angle around cell-center then featurized the local cell environment (15 features), with the sum of cell environment features normalized to 1. Note the first component of the cell environment features is practically the fraction of the cell boundary in cell-cell contact. The high-dimensional cell feature space was dimensionally reduced using principal component analysis (PCA), retaining the largest 11 eigen-components of the feature covariance matrix (spanning all treatments and image stacks) which captured >99% of the variability.

Motility features Cell motility was characterized in a single-cell manner, referenced both to the image frame and relative to neighboring cells. Single-cell displacement $\overrightarrow{\Delta x}$ between tracked frames was z-normalized, and cells which could not be tracked backward for a frame had

unrecorded displacements and were not used in our analysis. The local motility alignment of a single-cell to the local neighborhood of contacting cells (sharing a Voronoi boundary) was measured by extracting the cosine of the angle between the single-cell and direct neighbors via $\hat{p}_1 \cdot \hat{p}_2$ with $\hat{p} = \overrightarrow{\Delta x}/|\overrightarrow{\Delta x}|$. Local contact inhibition of locomotion was measured via the higher-order vector formed by $(\hat{p}_1 - \hat{p}_2) \cdot \hat{r}_{12}$ with $\vec{r}_{12}$ the separation vector between cells[105]. Neighborhood averages were taken via the Voronoi partition, averaged over neighbors and weighted by the relative length of the boundary to each neighbor, see Supplementary Figure 3.

Batch normalization The live-cell imaging data analyzed were collected from two separate imaging studies. Single-cell featurization can depend in subtle ways upon the imaging treatment and sample batch. To normalize these effects we utilized a batch normalization procedure at the single-cell feature level. For each morphology feature, we utilized a histogram matching procedure between negative control (PBS) treatments. We then fit a linear model to the histogram-matched distributions, and applied this linear model between sample batches, see Supplementary Figure 8.

Cell tracking To follow single-cells through time to extract the set of single-cell trajectories for morphodynamical trajectory embedding, we utilized a Bayesian likelihood-based approach implemented in the btrack software package[70] using default parameters. This Bayesian approach was applied for each frame over a 12 frame window, and then successful tracks over each pair of successive frames were extracted. See **Supplementary Table 1** for manual validation of tracking performance.

Morphodynamical trajectory embedding To maximize the single-cell information, we extended single-timepoint morphology and motility features over single-cell trajectories using a delay-embedding approach, described in Copperman et al.[26] In brief, single-cell features including motility features, but excluding cell-cycle features, were concatenated along the trajectory length to form morphodynamical feature trajectories. We tested multiple trajectory lengths and selected a trajectory length of 10 hours where the best prediction of withheld treatment combination RNA-seq was obtained, see **Supplementary Table 2**. We utilized a dynamical embedding approach described below to cluster trajectories and visualize this space, and did not perform any further dimensionality reduction upon the trajectory concatenated morphological feature PCAs and motility feature trajectories prior to dynamical model building.

Data-driven dynamical Markov state model To capture dynamical properties within the morphodynamical space, we constructed a transition matrix Markov model within the trajectory embedding space. The embedded space was binned into "microbins" using k-means clustering with $k = 200$ clusters. Results using 50, 100, 200, and 400 clusters are qualitatively similar. In this discrete space, a transition matrix T between bins was estimated from the set of transition counts C_{ij} from microbin *i* to *j* as $T_{ij} = C_{ij}/C_i$ with $C_i = \sum_j C_{ij}$. This accounting was agnostic to cell birth and death processes, yet we observe our model well reproduces morphodynamical state evolution, see Figure 5b.

Dynamical features To evaluate live-cell behavior via characterization of shared dynamics, we have applied a dynamical featurization approach via the data-driven transition matrix model. Using a transition matrix model constructed from all possible single-cell trajectory steps in the in the microbinned trajectory feature space, we construct the Hermitian extension $H = \frac{1}{2}[(T + T') + i(T - T')]$ with T' the transpose of the transition matrix T, this approach numerically stabilizes the eigendecomposition and provides all real eigenvalues for unambiguous ordering of eigencomponents[106]. We retain 15 dominant eigencomponents (see **Supplementary Figure 10**), and concatenate real and imaginary parts of eigenvectors to construct a 30-dimensional characterization of each microbin center. To visualize the dynamical trajectory space, we apply UMAP dimensionality reduction of the microstate eigenvector components to 2 components. Average flows in the UMAP space are calculated via calculating microstate dependent average displacements via the transition matrix $< x_i >= \sum_j (x_i - x_j) T_{ij}$ and averaging over 10 nearest microstate neighbors for smoothness. We note that UMAP flows were used only for visualization, not featurization.

Morphodynamical cell states As a tool for reducing complexity and extracting biological meaning in the morphodynamical embedding space, we defined a set of macrostates by clustering together microstates using dynamical similarity. We utilize the eigencomponents of H (Hermitian extension of the transition matrix T, see Dynamical Embedding) and perform k-means in the kinetic motifs. We utilize a lower cutoff of 0.015 for the total fraction of cell trajectories assigned to each state; if a microstate has too few trajectories assigned, then it is combined with its nearest neighbor by Euclidean distance in the space of dynamical motifs. k-means clusters are increased until the requested number of states with minimum fraction assignment is obtained. We then evaluated the capability of the derived macrostates to describe the state-change dynamics

by evaluating the sum of timescales captured in the microstate transition matrix model, related to the VAMP score[107]. We observe a rapid increase in score increasing to 10 states and continued increase beyond 15 states, see Supplementary Figure 9. Note that the macrostates, like the features themselves, were not designed or optimized for the task of predicting RNA levels.

Cell state change pathways To extract the sequences of morphodynamical cell states under EMT/MET, we adopted a transition path approach to calculate commitors and state change sequences utilizing our data-driven Markov model[67]. Transition matrices were constructed between morphodynamical cell states (macrostates), and flux analysis was carried out using the PyEMMA analysis package[68]; all pathways carrying flux between sets of initial and final states were evaluated to find dominant state change sequences. Committor probabilities (for reaching the final state before returning to the initial state) were highly dependent upon culture treatment, but cell state change sequences were quite robust to culture treatment, see **Supplementary Figure 7.**

Cell-cycle reporter analysis and dynamical modeling To capture cell-cycle dynamics from the HDHB reporter images, we adopted a similar data-driven modeling approach as we took in defining the morphodynamical cell states. Reporter levels in the nuclear and cytoplasmic compartments were extracted, and the ratio of these reporter levels was used as a self-normalizing readout of cell-cycle state, where exclusion of HDHB from the nucleus is known to correlate with G2 cell-cycle state, with maximal nuclear correlation occurring abruptly at mitosis and decreasing gradually from G1 to S, and with minimal nuclear signal at G2[100]. To divide reporter ratio values into cell-cycle stages, we utilized our Markov state modeling and dynamical embedding procedure, first building a microbin model with 50 bins evenly spaced throughout the range of reporter ratio values, then dividing these into 4 macrostates via k-means clustering in dynamical motifs, see **Supplementary Figure 10**. We then calculated mean first passage times using PyEMMA between cell-cycle stages as a readout of cell-cycle stage lifetimes in each of the morphodynamical cell states.

Bulk RNAseq reconstruction To capture the biological drivers of morphodynamical cell state changes, we mapped our morphodynamical cell states (defined above) to RNA-seq-based gene expression profiles. We adopted a linear decomposition approach based on the single assumption that the identical set of morphodynamic states occurs in all experimental conditions so that each state exhibits the same average gene profile in all conditions. In this framework, if cells in treatment A are subdivided into a set of states s with known state populations p_s^A such that

$\sum_s p_s^A = 1$, and the state and treatment dependent average gene levels are known, a bulk measurement of the ith gene can be reconstructed exactly as $< g_i^A >= \sum_s p_s^A g_i^{s,A}$, where $g_i^{s,A}$ is the average expression level for gene *i* in state *s* under experimental condition A. We approximate this exact expression by making the assumption that cells in state s under each treatment have identical *average* gene expression, i.e., that $g_i^{s,A} = g_i^s$ regardless of A, for every s. The utility of this approximation can be evaluated via our results, and is equivalent to letting the states form a non-negative matrix factorization of the bulk expression. Under this assumption, we have a linear system of equations connecting state populations and state gene expression levels $\{< g_i^A >= \sum_s p_s^A g_i^s, < g_i^B >= \sum_s p_s^B g_i^s, \ldots.\}$, one equation for each treatment A,B,C,… based on treatment-specific cell state populations p_s^A, p_s^B, … directly measured via live-cell imaging and morphodynamical analysis, and with paired bulk RNA-seq measurements $< g_i^A >$. If there are as many measurements as states, this linear equation can be inverted for the gene expression profiles in each state, g_i^s. If there are less states than treatments and the solution is over-determined, we obtain the solution over all possible combinations of treatments and average over the results. In practice, true solution of the linear system would yield negative gene levels, so we do a least squares minimization with the constraint of positive gene levels. We use fold-changes rather than absolute gene levels to preserve the batch and replicate normalization, this normalization does not affect the system of equations as it enters on both sides of the equality. To validate our state decomposition of measured bulk RNA-seq pipeline, we split our data into training sets and validation sets. State gene expression levels are trained from the training set gene levels only, and gene expression for withheld test set conditions are then predicted via the measured morphodynamical cell state populations. Null model predictions are constructed from random state populations combined with previously estimated state-specific gene levels (from true populations) as a measure of how unique the measured state populations are at predicting the test set gene expression.

Gene set enrichment To interpret morphodynamical cell state gene expression profiles, we performed gene set enrichment analysis via the pyGSEA package[108]. We utilized the preranked algorithm, sorting genes via the predicted gene expression levels in each morphodynamical cell state. We ran gene set enrichment using the Hallmark gene sets[82], which broadly capture well-studied biological processes and cell signaling activity.

Acknowledgements

We thank Joe Gray for contributions to project conceptualization, Mark Dane for technical guidance and assistance accessing LINCS data, David Aristoff and Gideon Simpson for mathematically oriented discussions, and John Russo and Luke Ternes for input regarding computational implementation. J.C. was supported by the Damon Runyon Cancer Research Foundation Quantitative Biology Fellowship DRQ-09-20. Y.H.C is supported in part by the National Cancer Institute (U54CA209988, U2CCA233280, U01 CA224012). D.M.Z. acknowledges support from the National Science Foundation (MCB 2119837). L.M.H. acknowledges support from NIH research grants U54CA209988 and U54HG008100, and the Anna Fuller Foundation. The authors acknowledge Lauren Kronebusch for help with manuscript editing.

Competing Interest

The authors declare no competing interests.

Data and Code Availability

All codes and scripts to perform the analysis in this work can be found at the project github repository.

https://github.com/jcopperm/celltraj

LINCS MCF10A Molecular Deep Dive data is available in some formats from the synapse database[57] and additional data is available upon request.

References

1. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, (2012).
2. Stergachis, A. B. *et al.* XDevelopmental fate and cellular maturity encoded in human regulatory DNA landscapes. *Cell* **154**, (2013).
3. Brunner, A. *et al.* Ultra-high sensitivity mass spectrometry quantifies single-cell proteome changes upon perturbation. *Mol Syst Biol* **18**, (2022).
4. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* **6**, (2009).
5. Lareau, C. A. *et al.* Massively parallel single-cell mitochondrial DNA genotyping and chromatin profiling. *Nat Biotechnol* **39**, (2021).
6. Zenobi, R. Single-cell metabolomics: Analytical and biological perspectives. *Science* vol. 342 Preprint at https://doi.org/10.1126/science.1243259 (2013).
7. Davies, A. E. *et al.* Article Systems-Level Properties of EGFR-RAS-ERK Signaling Amplify Local Signals to Generate Dynamic Gene Expression Heterogeneity. *Cell Syst* **11**, 161-175.e5 (2020).
8. Boudreau, N. & Bissell, M. J. Extracellular matrix signaling: Integration of form and function in normal and malignant cells. *Curr Opin Cell Biol* **10**, (1998).
9. Discher, D. E., Janmey, P. & Wang, Y. L. Tissue cells feel and respond to the stiffness of their substrate. *Science* vol. 310 Preprint at https://doi.org/10.1126/science.1116995 (2005).
10. Cerchiari, A. E. *et al.* A strategy for tissue self-organization that is robust to cellular heterogeneity and plasticity. *Proc Natl Acad Sci U S A* **112**, (2015).
11. Colombo, M., Raposo, G. & Théry, C. Biogenesis, secretion, and intercellular interactions of exosomes and other extracellular vesicles. *Annual review of cell and developmental biology* vol. 30 Preprint at https://doi.org/10.1146/annurev-cellbio-101512-122326 (2014).
12. Heldin, C. H., Lu, B., Evans, R. & Gutkind, J. S. Signals and receptors. *Cold Spring Harb Perspect Biol* **8**, (2016).
13. McFaline-Figueroa, J. L. *et al.* A pooled single-cell genetic screen identifies regulatory checkpoints in the continuum of the epithelial-to-mesenchymal transition. *Nature Genetics 2019 51:9* **51**, 1389–1398 (2019).
14. Deshmukh, A. P. *et al.* Identification of EMT signaling cross-talk and gene regulatory networks by single-cell RNA sequencing. *Proceedings of the National Academy of Sciences* **118**, e2102050118 (2021).
15. Vandereyken, K., Sifrim, A., Thienpont, B. & Thierry Voet, &. Methods and applications for single-cell and spatial multi-omics. *Nature Reviews Genetics* | **24**, 494–515 (2023).
16. Giesen, C. *et al.* Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nature Methods 2014 11:4* **11**, 417–422 (2014).

17. Lin, J. R., Fallahi-Sichani, M. & Sorger, P. K. Highly multiplexed imaging of single cells using a high-throughput cyclic immunofluorescence method. *Nature Communications 2015 6:1* **6**, 1–7 (2015).
18. Tsujikawa, T. *et al.* Quantitative Multiplex Immunohistochemistry Reveals Myeloid-Inflamed Tumor-Immune Complexity Associated with Poor Prognosis. *Cell Rep* **19**, 203–217 (2017).
19. C. Greenwald, E., Mehta, S. & Zhang, J. Genetically Encoded Fluorescent Biosensors Illuminate the Spatiotemporal Regulation of Signaling Networks. *Chem Rev* **118**, 11707–11794 (2018).
20. Linghu, C. *et al.* Spatial Multiplexing of Fluorescent Reporters for Imaging Signaling Network Dynamics. *Cell* **183**, 1682-1698.e24 (2020).
21. Yang, J. M. *et al.* Deciphering cell signaling networks with massively multiplexed biosensor barcoding. *Cell* **184**, 6193-6206.e14 (2021).
22. Neumann, B. *et al.* High-throughput RNAi screening by time-lapse imaging of live human cells. *Nat Methods* **3**, 385–390 (2006).
23. Held, M. *et al.* CellCognition: Time-resolved phenotype annotation in high-throughput live cell imaging. *Nat Methods* **7**, 747–754 (2010).
24. Nketia, T. A., Sailem, H., Rohde, G., Machiraju, R. & Rittscher, J. Analysis of live cell images: Methods, tools and opportunities. *Methods* vol. 115 65–79 Preprint at https://doi.org/10.1016/j.ymeth.2017.02.007 (2017).
25. Stirling, D. R. *et al.* CellProfiler 4: improvements in speed, utility and usability. *BMC Bioinformatics* **22**, 1–11 (2021).
26. Copperman, J., Gross, S. M., Chang, Y. H., Heiser, L. M. & Zuckerman, D. M. Morphodynamical cell state description via live-cell imaging trajectory embedding. *Commun Biol* **6**, (2023).
27. Karacosta, L. G. *et al.* Mapping lung cancer epithelial-mesenchymal transition states and trajectories with single-cell resolution. *Nat Commun* **10**, 1–15 (2019).
28. Cho, H., Kuo, Y. H. & Rockne, R. C. Comparison of cell state models derived from single-cell RNA sequencing data: graph versus multi-dimensional space. *Mathematical Biosciences and Engineering* **19**, (2022).
29. Tong, A. *et al.* Learning transcriptional and regulatory dynamics driving cancer cell plasticity using neural ODE-based optimal transport. *bioRxiv* (2023).
30. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* **32**, (2014).
31. Bendall, S. C. *et al.* Single-cell trajectory detection uncovers progression and regulatory coordination in human b cell development. *Cell* **157**, 714–725 (2014).
32. La Manno, G. *et al.* RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
33. Bergen, V., Lange, M., Peidli, S., Wolf, F. A. & Theis, F. J. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat Biotechnol* **38**, 1408–1414 (2020).
34. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods* **12**, (2015).
35. Wang, Z. *et al.* Transcriptome Deconvolution of Heterogeneous Tumor Samples with Immune Infiltration. *iScience* **9**, (2018).
36. Wang, X., Park, J., Susztak, K., Zhang, N. R. & Li, M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat Commun* **10**, (2019).

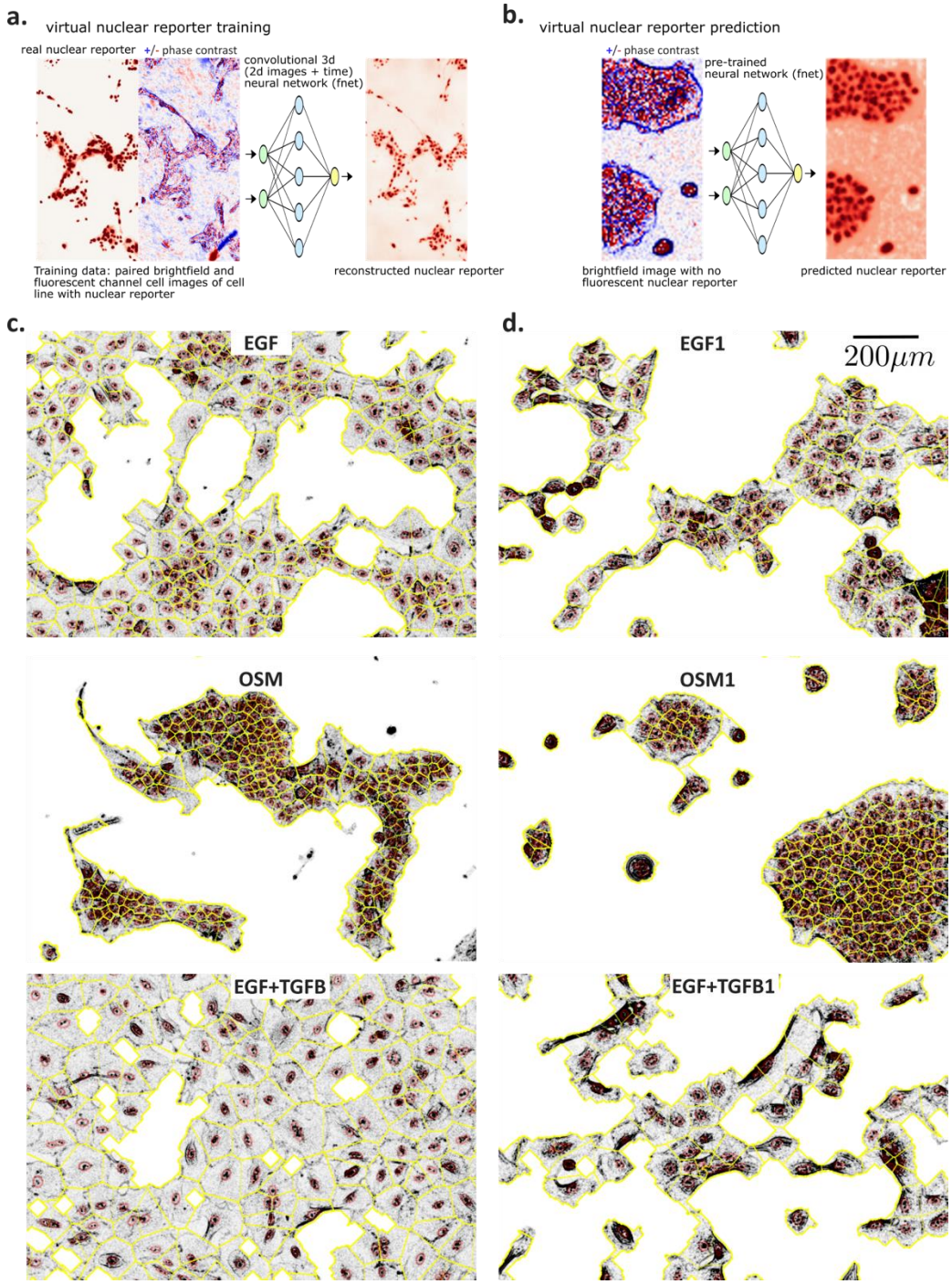
37. Danziger, S. A. *et al.* AdApTS: Automated deconvolution augmentation of profiles for tissue specific cells. *PLoS One* **14**, (2019).
38. Noé, F. & Fischer, S. Transition networks for modeling the kinetics of conformational change in macromolecules. *Curr Opin Struct Biol* **18**, 154–162 (2008).
39. Husic, B. E. & Pande, V. S. Markov State Models: From an Art to a Science. *Journal of the American Chemical Society* vol. 140 Preprint at https://doi.org/10.1021/jacs.7b12191 (2018).
40. Wang, W. *et al.* Live-cell imaging and analysis reveal cell phenotypic transition dynamics inherently missing in snapshot data. *Sci Adv* **6**, eaba9319 (2020).
41. Wang, W., Poe, D., Yang, Y., Hyatt, T. & Xing, J. Epithelial-to-mesenchymal transition proceeds through directional destabilization of multidimensional attractor. *Elife* **11**, (2022).
42. Moffat, J. *et al.* A Lentiviral RNAi Library for Human and Mouse Genes Applied to an Arrayed Viral High-Content Screen. *Cell* **124**, (2006).
43. Yin, Z. *et al.* A screen for morphological complexity identifies regulators of switch-like transitions between discrete cell shapes. *Nat Cell Biol* **15**, (2013).
44. Wu, Z. *et al.* DynaMorph: self-supervised learning of morphodynamic states of live cells. *Mol Biol Cell* **33**, (2022).
45. Chow, Y. L., Singh, S., Carpenter, A. E. & Way, G. P. Predicting drug polypharmacology from cell morphology readouts using variational autoencoder latent space arithmetic. *PLoS Comput Biol* **18**, (2022).
46. Tegtmeyer, M. *et al.* High-dimensional phenotyping to define the genetic basis of cellular morphology. *Nat Commun* **15**, (2024).
47. Yang, K. D. *et al.* Multi-domain translation between single-cell imaging and sequencing data using autoencoders. *Nat Commun* **12**, (2021).
48. Kobayashi-Kirschvink, K. J. *et al.* Prediction of single-cell RNA expression profiles in live cells by Raman microscopy with Raman2RNA. *Nat Biotechnol* (2024) doi:10.1038/s41587-023-02082-2.
49. Soule, H. D. *et al.* Isolation and Characterization of a Spontaneously Immortalized Human Breast Epithelial Cell Line, MCF-10. *Cancer Res* **50**, (1990).
50. Witt, A. E. *et al.* Functional proteomics approach to investigate the biological activities of cDNAs implicated in breast cancer. *J Proteome Res* **5**, (2006).
51. Melani, M., Simpson, K. J., Brugge, J. S. & Montell, D. Regulation of Cell Adhesion and Collective Cell Migration by Hindsight and Its Human Homolog RREB1. *Current Biology* **18**, (2008).
52. Seton-Rogers, S. E. *et al.* Cooperation of the ErbB2 receptor and transforming growth factor β in induction of migration and invasion in mammary epithelial cells. *Proc Natl Acad Sci U S A* **101**, (2004).
53. Debnath, J. *et al.* The role of apoptosis in creating and maintaining luminal space within normal and oncogene-expressing mammary acini. *Cell* **111**, (2002).
54. Debnath, J., Muthuswamy, S. K. & Brugge, J. S. Morphogenesis and oncogenesis of MCF-10A mammary epithelial acini grown in three-dimensional basement membrane cultures. *Methods* vol. 30 Preprint at https://doi.org/10.1016/S1046-2023(03)00032-X (2003).
55. Sampattavanich, S. *et al.* Encoding Growth Factor Identity in the Temporal Dynamics of FOXO3 under the Combinatorial Control of ERK and AKT Kinases. *Cell Syst* **6**, (2018).

56. Caldera, M. *et al.* Mapping the perturbome network of cellular perturbations. *Nat Commun* **10**, (2019).
57. Gross, S. M. *et al.* A multi-omic analysis of MCF10A cells provides a resource for integrative assessment of ligand-mediated molecular and phenotypic responses. *Commun Biol* **5**, (2022).
58. Espinosa-Neira, R., Mejia-Rangel, J., Cortes-Reynosa, P. & Salazar, E. P. Linoleic acid induces an EMT-like process in mammary epithelial cells MCF10A. *International Journal of Biochemistry and Cell Biology* **43**, (2011).
59. Galindo-Hernandez, O., Serna-Marquez, N., Castillo-Sanchez, R. & Salazar, E. P. Extracellular vesicles from MDA-MB-231 breast cancer cells stimulated with linoleic acid promote an EMT-like process in MCF10A cells. *Prostaglandins Leukot Essent Fatty Acids* **91**, (2014).
60. Zhang, J. *et al.* TGF-β-induced epithelial-to-mesenchymal transition proceeds through stepwise activation of multiple feedback loops. *Sci Signal* **7**, (2014).
61. Mori, S. *et al.* Enhanced expression of integrin αvβ3 induced by TGF-β is required for the enhancing effect of fibroblast growth factor 1 (FGF1) in TGF-β-induced epithelial-mesenchymal transition (EMT) in mammary epithelial cells. *PLoS One* **10**, (2015).
62. Rodriguez-Monterrosas, C., Díaz-Aragon, R., Leal-Orta, E., Cortes-Reynosa, P. & Perez Salazar, E. Insulin induces an EMT-like process in mammary epithelial cells MCF10A. *J Cell Biochem* **119**, (2018).
63. Olea-Flores, M. *et al.* Leptin promotes expression of EMT-related transcription factors and invasion in a src and FAK-dependent pathway in MCF10a mammary epithelial cells. *Cells* **8**, (2019).
64. Antón-García, P. *et al.* TGFβ1-Induced EMT in the MCF10A Mammary Epithelial Cell Line Model Is Executed Independently of SNAIL1 and ZEB1 but Relies on JUNB-Coordinated Transcriptional Regulation. *Cancers (Basel)* **15**, (2023).
65. Paul, I. *et al.* Parallelized multidimensional analytic framework applied to mammary epithelial cells uncovers regulatory principles in EMT. *Nat Commun* **14**, (2023).
66. Spencer, S. L. *et al.* XThe proliferation-quiescence decision is controlled by a bifurcation in CDK2 activity at mitotic exit. *Cell* **155**, (2013).
67. E, W. & Vanden-Eijnden, E. Transition-path theory and path-finding algorithms for the study of rare events. *Annu Rev Phys Chem* **61**, (2010).
68. Scherer, M. K. *et al.* PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *J Chem Theory Comput* **11**, (2015).
69. Ounkomol, C., Seshamani, S., Maleckar, M. M., Collman, F. & Johnson, G. R. Label-free prediction of three-dimensional fluorescence images from transmitted-light microscopy. *Nat Methods* **15**, 917–920 (2018).
70. Ulicna, K., Vallardi, G., Charras, G. & Lowe, A. R. Automated Deep Lineage Tree Analysis Using a Bayesian Single Cell Tracking Approach. *Front Comput Sci* **3**, (2021).
71. Röblitz, S. & Weber, M. Fuzzy spectral clustering by PCCA+: Application to Markov state models and data classification. *Adv Data Anal Classif* **7**, (2013).
72. Reuter, B., Weber, M., Fackeldey, K., Röblitz, S. & Garcia, M. E. Generalized Markov State Modeling Method for Nonequilibrium Biomolecular Dynamics: Exemplified on Amyloid β Conformational Dynamics Driven by an Oscillating Electric Field. *J Chem Theory Comput* **14**, 3579–3594 (2018).

73. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv* (2018).
74. Dongre, A. & Weinberg, R. A. New insights into the mechanisms of epithelial–mesenchymal transition and implications for cancer. *Nature Reviews Molecular Cell Biology* vol. 20 Preprint at https://doi.org/10.1038/s41580-018-0080-4 (2019).
75. Donovan, J. & Slingerland, J. Transforming growth factor-β and breast cancer: Cell cycle arrest by transforming growth factor-β and its disruption in cancer. *Breast Cancer Research* vol. 2 Preprint at https://doi.org/10.1186/bcr43 (2000).
76. Mejlvang, J. *et al.* Direct repression of cyclin D1 by SIP1 attenuates cell cycle progression in cells undergoing an epithelial mesenchymal transition. *Mol Biol Cell* **18**, (2007).
77. Lovisa, S. *et al.* Epithelial-to-mesenchymal transition induces cell cycle arrest and parenchymal damage in renal fibrosis. *Nat Med* **21**, (2015).
78. Kohrman, A. Q. & Matus, D. Q. Divide or Conquer: Cell Cycle Regulation of Invasive Behavior. *Trends in Cell Biology* vol. 27 Preprint at https://doi.org/10.1016/j.tcb.2016.08.003 (2017).
79. Petersen, O. W., Rønnov-Jessen, L., Howlett, A. R. & Bissell, M. J. Interaction with basement membrane serves to rapidly distinguish growth and differentiation pattern of normal and malignant human breast epithelial cells. *Proceedings of the National Academy of Sciences* **89**, 9064–9068 (1992).
80. Battle, C. *et al.* Broken detailed balance at mesoscopic scales in active biological systems. *Science (1979)* **352**, 604–607 (2016).
81. 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* **14**, e1005927 (2018).
82. Liberzon, A. *et al.* The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Syst* **1**, (2015).
83. Zaritsky, A. *et al.* Interpretable deep learning uncovers cellular properties in label-free live cell images that are predictive of highly metastatic melanoma. *Cell Syst* **12**, 733-747.e6 (2021).
84. Nguyen, P. *et al.* Unsupervised discovery of dynamic cell phenotypic states from transmitted light movies. *PLoS Comput Biol* **17**, (2021).
85. Ternes, L. *et al.* A multi-encoder variational autoencoder controls multiple transformational features in single-cell image analysis. *Communications Biology 2022 5:1* **5**, 1–10 (2022).
86. Schau, G., Burlingame, E. & Chang, Y. H. DISSECT: DISentangle SharablE ConTent for Multimodal Integration and Crosswise-mapping. in *Proceedings of the IEEE Conference on Decision and Control* vols 2020-December 5092–5097 (Institute of Electrical and Electronics Engineers Inc., 2020).
87. Pratt, L. R. A statistical method for identifying transition states in high dimensional problems. *J Chem Phys* **85**, (1986).
88. Bolhuis, P. G., Chandler, D., Dellago, C. & Geissler, P. L. Transition Path Sampling: Throwing ropes over rough mountain passes, in the dark. *Annu Rev Phys Chem* **53**, 291–318 (2002).
89. Chong, L. T., Saglam, A. S. & Zuckerman, D. M. Path-sampling strategies for simulating rare events in biomolecular systems. *Current Opinion in Structural Biology* vol. 43 88–94 Preprint at https://doi.org/10.1016/j.sbi.2016.11.019 (2017).

90. Yang, J. & Weinberg, R. A. Epithelial-Mesenchymal Transition: At the Crossroads of Development and Tumor Metastasis. *Developmental Cell* vol. 14 Preprint at https://doi.org/10.1016/j.devcel.2008.05.009 (2008).
91. Grasset, E. M. *et al.* Triple-negative breast cancer metastasis involves complex epithelial-mesenchymal transition dynamics and requires vimentin. *Sci Transl Med* **14**, (2022).
92. Risom, T. *et al.* Differentiation-state plasticity is a targetable resistance mechanism in basal-like breast cancer. *Nat Commun* **9**, 3815 (2018).
93. Gambardella, G. *et al.* A single-cell analysis of breast cancer cell lines to study tumour heterogeneity and drug response. *Nat Commun* **13**, (2022).
94. Tatarova, Z. *et al.* A multiplex implantable microdevice assay identifies synergistic combinations of cancer immunotherapies and conventional drugs. *Nat Biotechnol* **40**, (2022).
95. Pong, A., Mah, C. K., Yeo, G. W. & Lewis, N. E. Computational cell–cell interaction technologies drive mechanistic and biomarker discovery in the tumor microenvironment. *Curr Opin Biotechnol* (2024).
96. Sauro, H. M. & Kholodenko, B. N. Quantitative analysis of signaling networks. *Prog Biophys Mol Biol* **86**, (2004).
97. Alon, U. Network motifs: Theory and experimental approaches. *Nature Reviews Genetics* vol. 8 Preprint at https://doi.org/10.1038/nrg2102 (2007).
98. Aldridge, B. B., Burke, J. M., Lauffenburger, D. A. & Sorger, P. K. Physicochemical modelling of cell signalling pathways. *Nature Cell Biology* vol. 8 Preprint at https://doi.org/10.1038/ncb1497 (2006).
99. Birtwistle, M. R. *et al.* Ligand-dependent responses of the ErbB signaling network: Experimental and modeling analyses. *Mol Syst Biol* **3**, (2007).
100. Gross, S. M. *et al.* Analysis and modeling of cancer drug responses using cell cycle phase-specific rate effects. *Nat Commun* **14**, (2023).
101. Sommer, C., Straehle, C., Kothe, U. & Hamprecht, F. A. Ilastik: Interactive learning and segmentation toolkit. in *Proceedings - International Symposium on Biomedical Imaging* 230–233 (2011). doi:10.1109/ISBI.2011.5872394.
102. Thévenaz, P., Ruttimann, U. E. & Unser, M. A pyramid approach to subpixel registration based on intensity. *IEEE Transactions on Image Processing* **7**, 27–41 (1998).
103. Coelho, L. P. Mahotas: Open source software for scriptable computer vision. *J Open Res Softw* **1**, e3 (2013).
104. Alizadeh, E., Xu, W., Castle, J., Foss, J. & Prasad, A. TISMorph: A tool to quantify texture, irregularity and spreading of single cells. *PLoS One* **14**, e0217346 (2019).
105. Bertrand, T. *et al.* Clustering and ordering in cell assemblies with generic asymmetric aligning interactions. *ArXiv* (2020).
106. Atev, S. E. Using Asymmetry in the Spectral Clustering of Trajectories. (University of Minnesota, 2011).
107. Nüske, F., Keller, B. G., Pérez-Hernández, G., Mey, A. S. J. S. & Noé, F. Variational approach to molecular kinetics. *J Chem Theory Comput* **10**, 1739–1752 (2014).
108. Fang, Z., Liu, X. & Peltz, G. GSEApy: a comprehensive package for performing gene set enrichment analysis in Python. *Bioinformatics* **39**, (2023).
109. Stringer, C., Wang, T., Michaelos, M. & Pachitariu, M. Cellpose: a generalist algorithm for cellular segmentation. *Nat Methods* **18**, 100–106 (2021).

Supplementary Data Tables and Figures

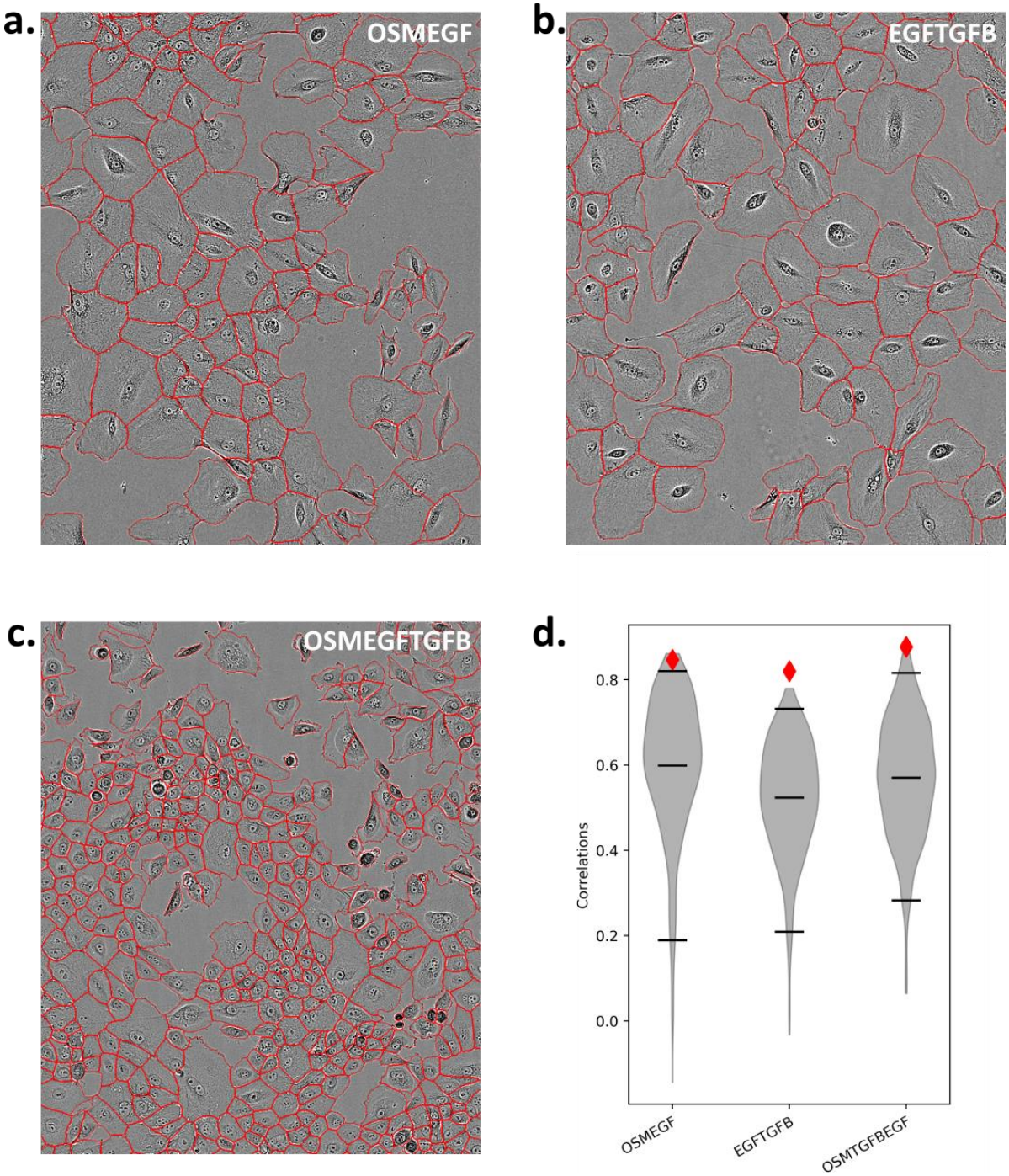


Supplementary Figure 1: Virtual nuclear staining and nuclear center-based segmentation
a) Paired nuclear reporter (red) and z-normalized phase contrast (+ red - blue) training data input, and reconstructed nuclear reporter output b) Nuclear reporter prediction from out of sample phase contrast images (OSM condition). c) Examples of nuclear segmentations (red) and associated Voronoi boundaries (yellow) overlaid upon phase contrast images (z-normalized absolute value,gray)

Supplementary Data Table 1. Segmentation and tracking manual validation. 100 cells per treatment were randomly selected, and evaluated manually to assess segmentation and tracking via the true positive rate (TPR—segmentations or tracks manually validated to be accurate, divided by the total number of cells assessed). Cell identification was assessed by comparing manual counts in a frame at the final timepoint to those obtained from virtually stained nuclear masks.. Segmentation performance from dataset 1 (e.g. EGF1, HGF1) is decreased because for these data the cells did not express a nuclear reporter, and the nucleus was detected via the virtual staining approach only. Tracking performance is decreased as well, due to the decreased segmentation performance and increased time between frames (30 minutes as compared to 15 minutes).

ligand	PBS	EGF	OSM	TGFB	TGFB +EGF	OSM +EGF	TGFB +OSM	TGFB +OSM +EGF
count/manual	99%	99%	98%	96%	99%	98%	96%	99%
seg TPR	92%	97%	97%	95%	96%	96%	95%	95%
% seg error	1%	1%	1%	1%	2%	2%	0%	1%
%ambiguous seg	7%	2%	2%	4%	2%	2%	5%	4%
% tracked	97%	91%	93%	92%	95%	93%	88%	90%
tracking TPR	100%	100%	98%	100%	99%	99%	100%	100%
% tracking error	0%	0%	2%	0%	0%	0%	0%	0%
%ambiguous tracks	0%	0%	0%	0%	1%	1%	0%	0%

ligand	PBS1	EGF1	HGF1	OSM1	IFNG+ EGF1	BMP2+ EGF1	EGF+ TGFB1
count/manual	79%	94%	74%	90%	99%	102%	113%
seg TPR	80%	62%	74%	86%	68%	78%	75%
% seg error	11%	25%	11%	4%	24%	16%	9%
%ambiguous seg	9%	10%	14%	10%	8%	6%	16%
% tracked	78%	52%	78%	69%	55%	66%	49%
tracking TPR	94%	94%	99%	88%	91%	95%	96%
% tracking error	4%	6%	1%	7%	5%	5%	4%
%ambiguous tracks	2%	0%	0%	5%	4%	0%	0%

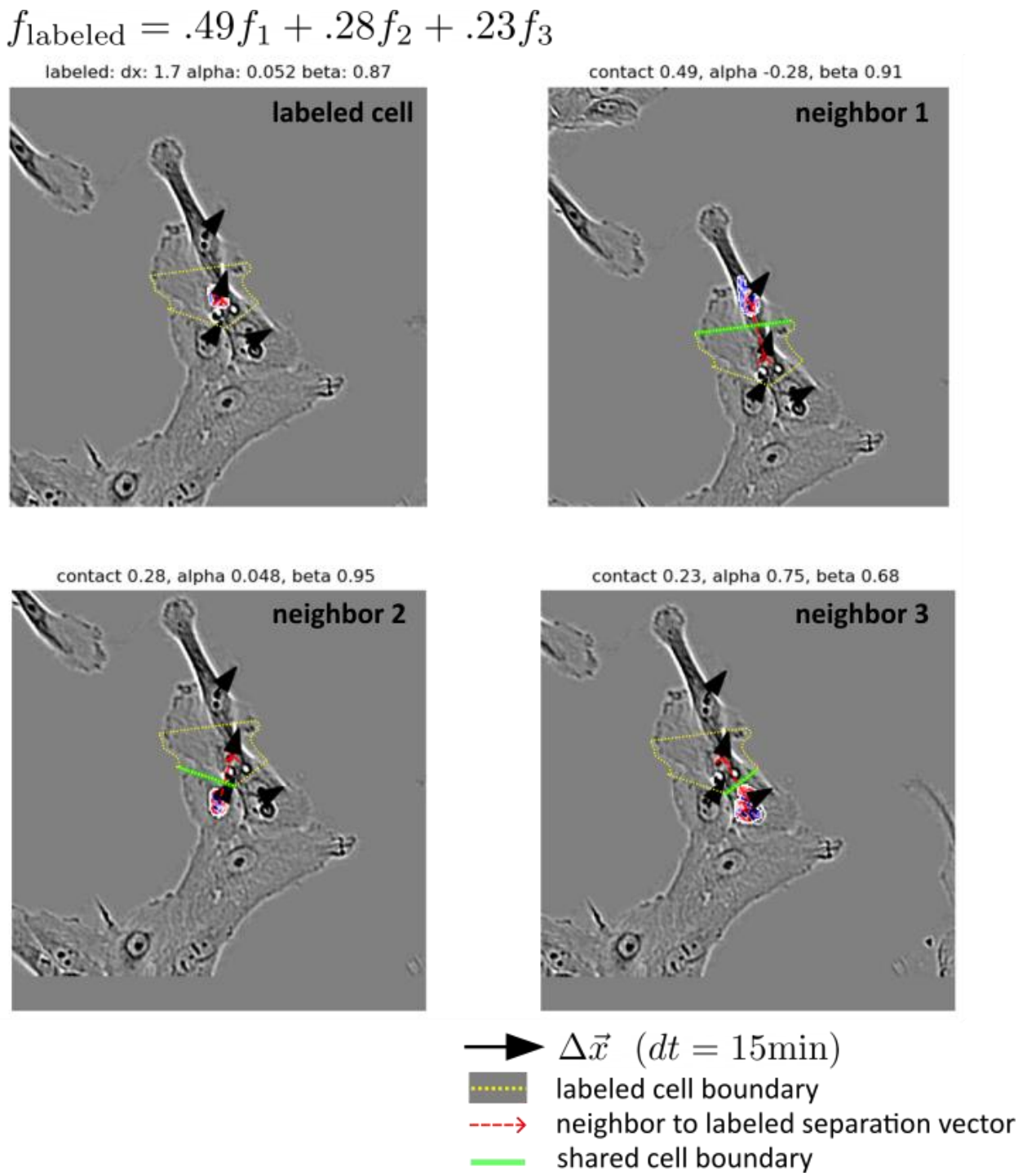


Supplementary Figure 2: MMIST pipeline applied to a data subset with cytoplasmic masks

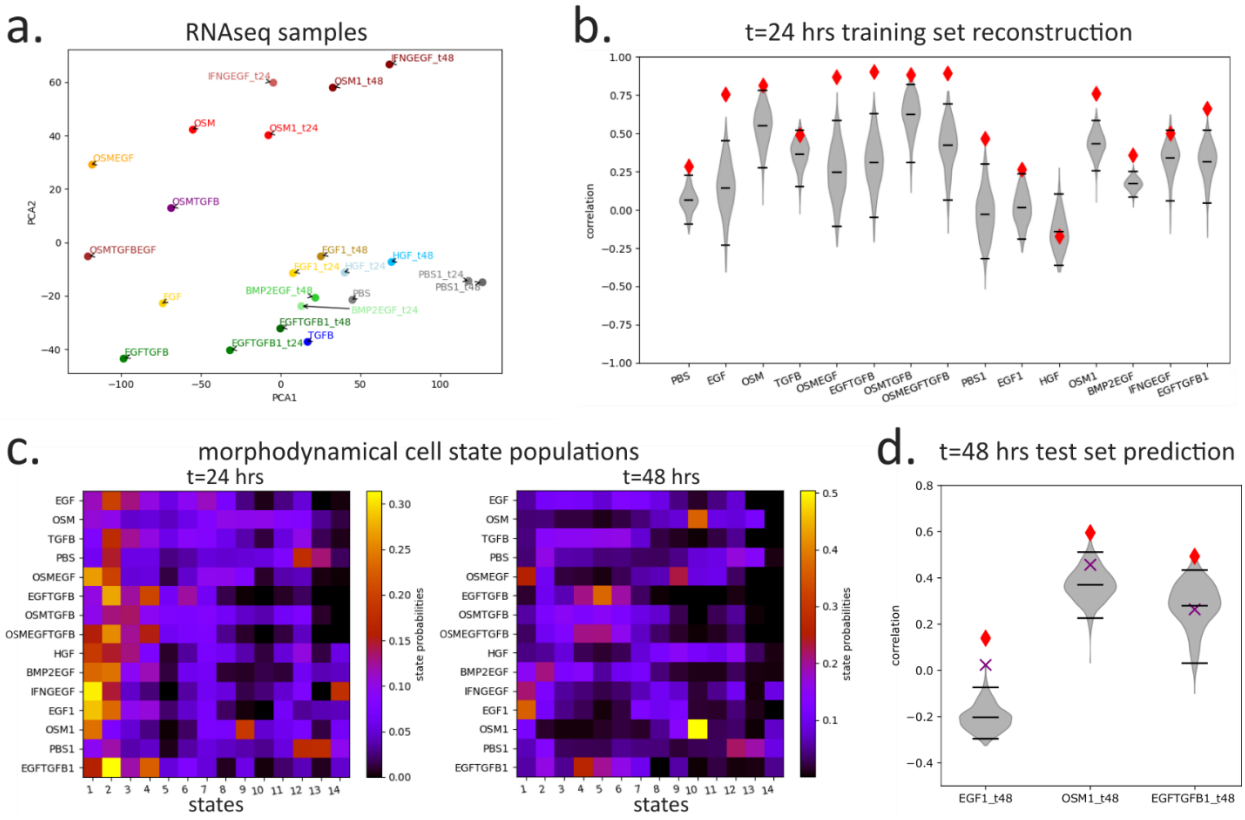
Cytoplasmic masks were obtained using the cellpose[109] software and a subset of the data (only imaging data from the cell-cycle reporter studies. a)-c) Representative images overlaid with cytoplasmic masks from three conditions. Identical tracking and featurization was performed, and the MMIST pipeline was applied to predict RNA-seq at 24 hrs. Predictions were validated via leave-one out cross-validation. d) Model prediction correlation to with-held RNA-seq data from three conditions (red diamonds) and null model (gray violin plots) with the 5th percentile, mean, and 95th percentile of the null distribution indicated (black lines).

test treatment	traj length (hrs)	pred. to exp. corr	diff. from null $\rho - \rho_{\text{null}}$	upreg true pos. rate	downreg true pos. rate
OSM+EGF	0	.26	-0.38	.51	.80
EGF+TGFB	0	.76	0.09	.63	.89
OSM+EGF+TGFB	0	.77	0.04	.77	.95
OSM+EGF	1	.51	0.02	.55	**.99**
EGF+TGFB	1	.76	0.22	.68	.93
OSM+EGF+TGFB	1	.75	0.14	.77	.93
OSM+EGF	4	.50	-0.22	.51	.98
EGF+TGFB	4	.74	0.05	.68	.80
OSM+EGF+TGFB	4	.73	-0.04	.77	**.96**
OSM+EGF	8	.55	-0.15	.52	**.99**
EGF+TGFB	8	.69	0	.68	**.87**
OSM+EGF+TGFB	8	.66	-0.08	.77	.94
OSM+EGF	10	**.70**	**0.21**	**.72**	.81
EGF+TGFB	10	**.81**	**0.27**	**.72**	.84
OSM+EGF+TGFB	10	.77	**0.12**	**.85**	.81
OSM+EGF	12	.71	0.2	.70	.77
EGF+TGFB	12	.74	0.14	.69	.85
OSM+EGF+TGFB	12	.64	0.01	.80	.80
OSM+EGF	16	.66	-0.03	.65	.92
EGF+TGFB	16	.79	0.15	.70	.85
OSM+EGF+TGFB	16	**.79**	0.09	.80	.88

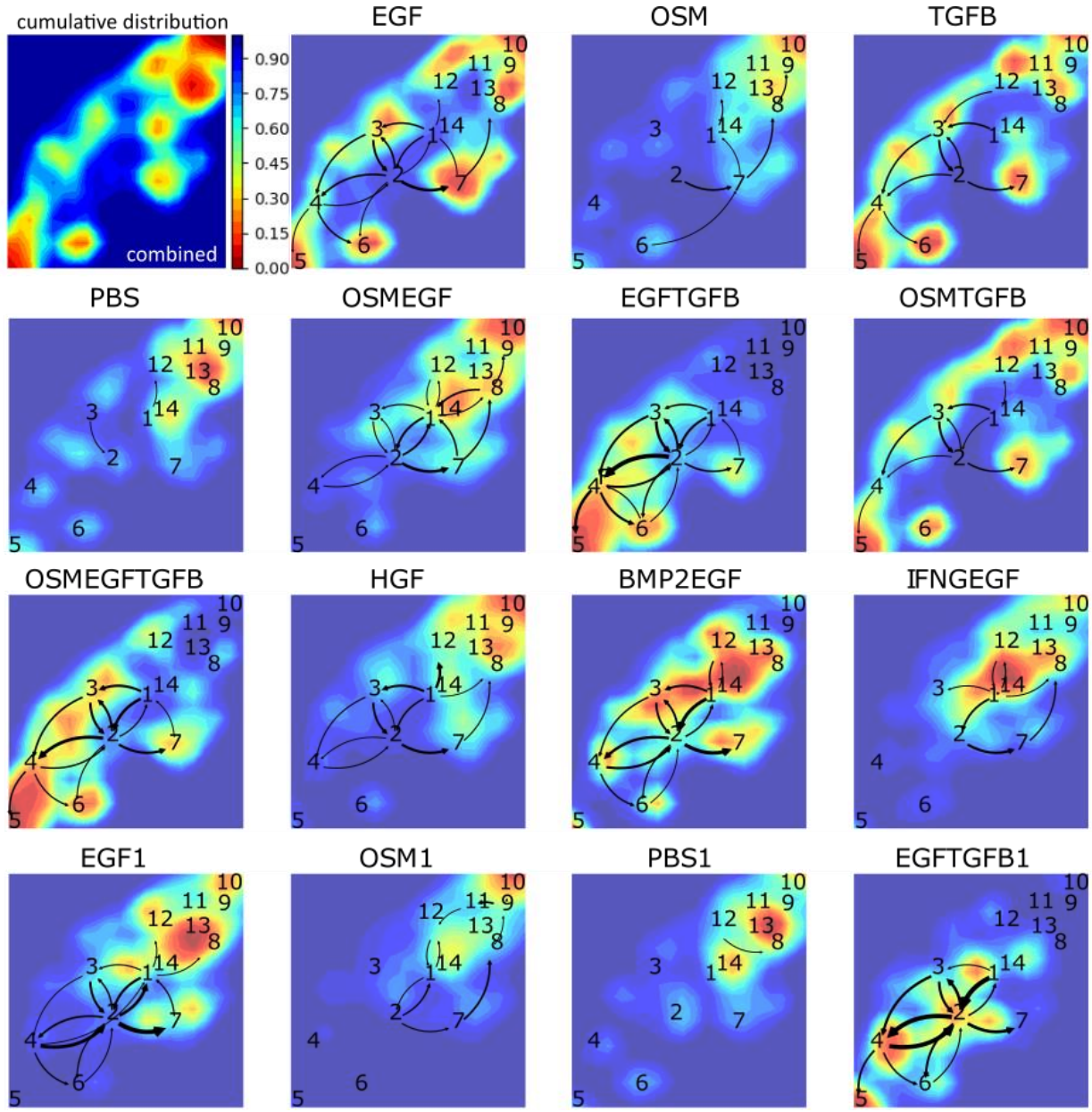
Supplementary Data Table 2. Test set gene expression validation with trajectory length.



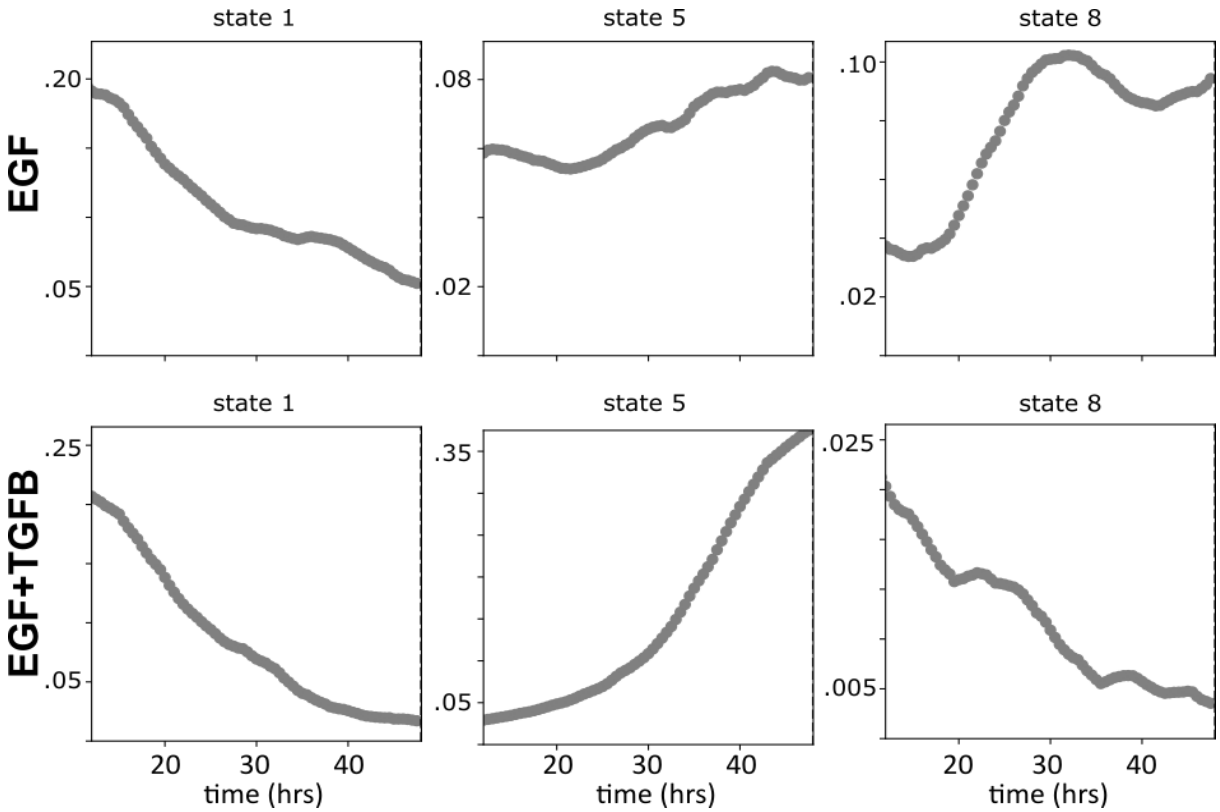
Supplementary Figure 3: Single-cell and neighborhood motility feature
Single-cell featurization of motility for the labeled cell in the upper right (nuclei in color and Voronoi segmentation in yellow) taken as the magnitude of the displacement from previous frame. The single-cell motility in the context of its local neighborhood taken as the neighbor-weighted average of the 3 boundary cells (upper right, lower left, lower right).



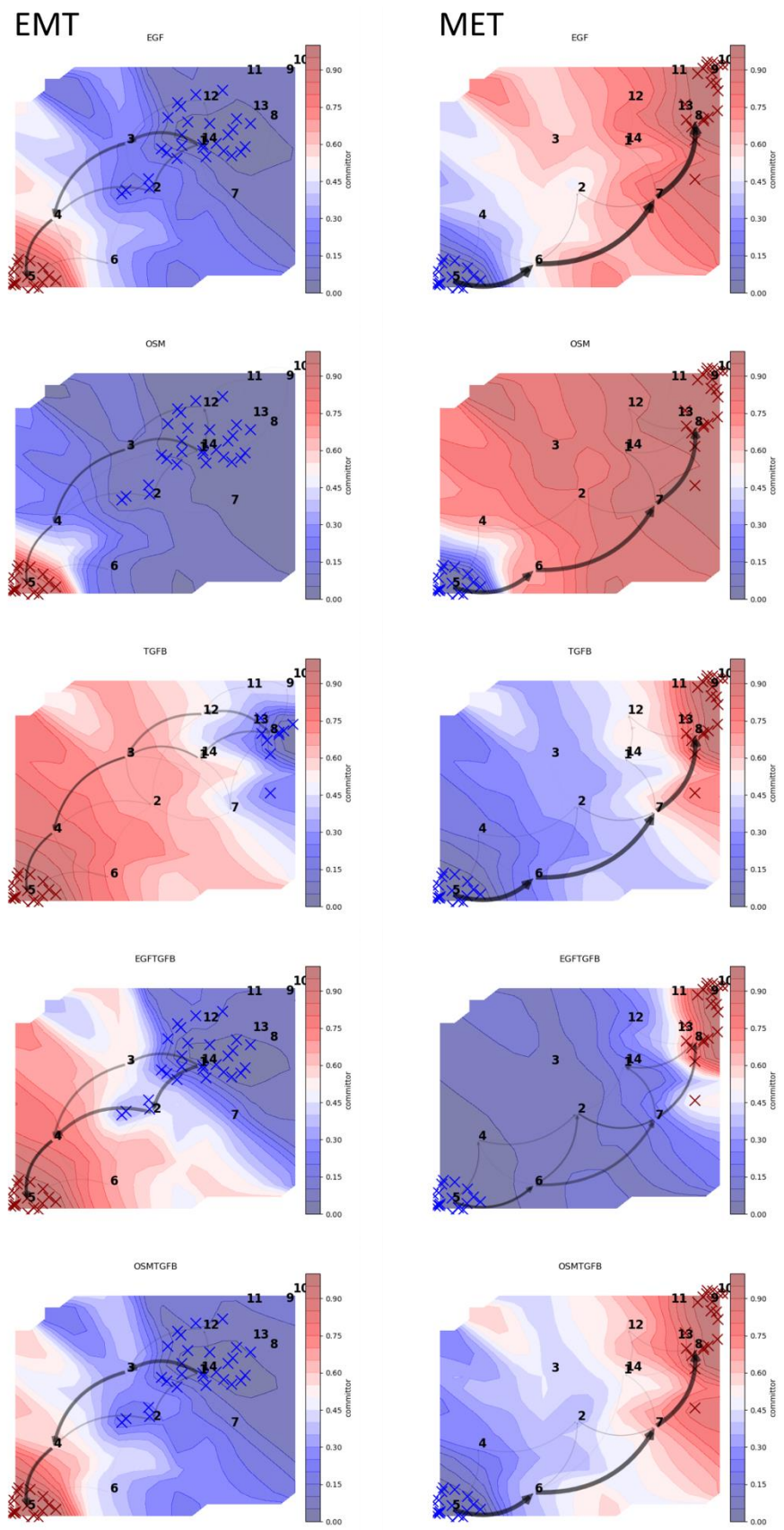
Supplementary Figure 4: Bulk RNAseq decomposition and time dependence
a) PCA1/2 projection of the RNAseq differential expression, showing sorting by ligand treatment and timepoint. b) Correlation between training set reconstruction and real experimental differential expression. c) Morphodynamical state populations at t=24, 48 hrs d) Test set prediction of RNAseq at 48 hours using t=24 hrs trained morphodynamical state gene expression profiles and measured live-cell state populations at t=48hrs.



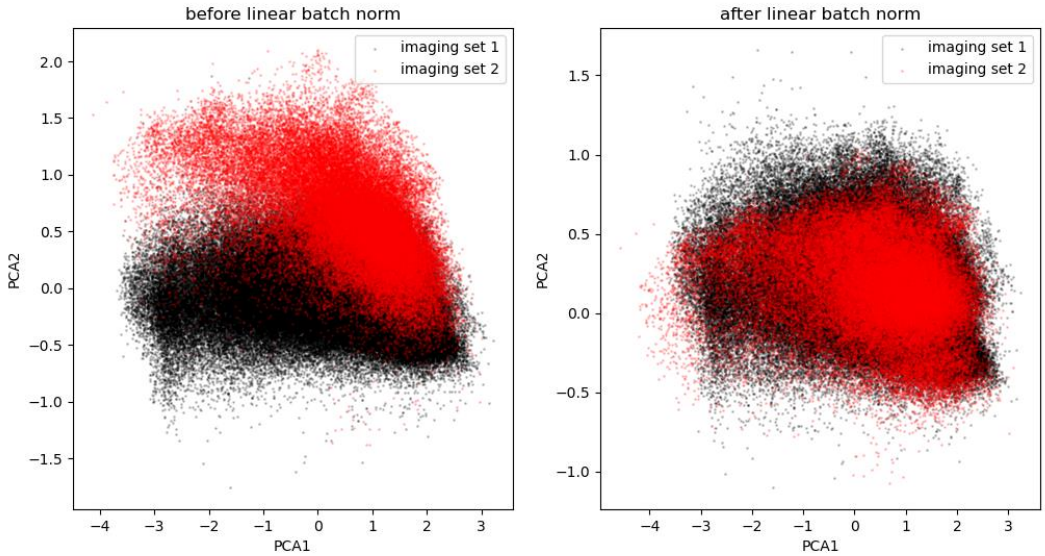
Supplementary Figure 5: Ligand-dependent populations and cell state flows
Cumulative populations in the UMAP embedding space (blue to red), and state-state transition flows at t=24hrs, in each ligand treatment.



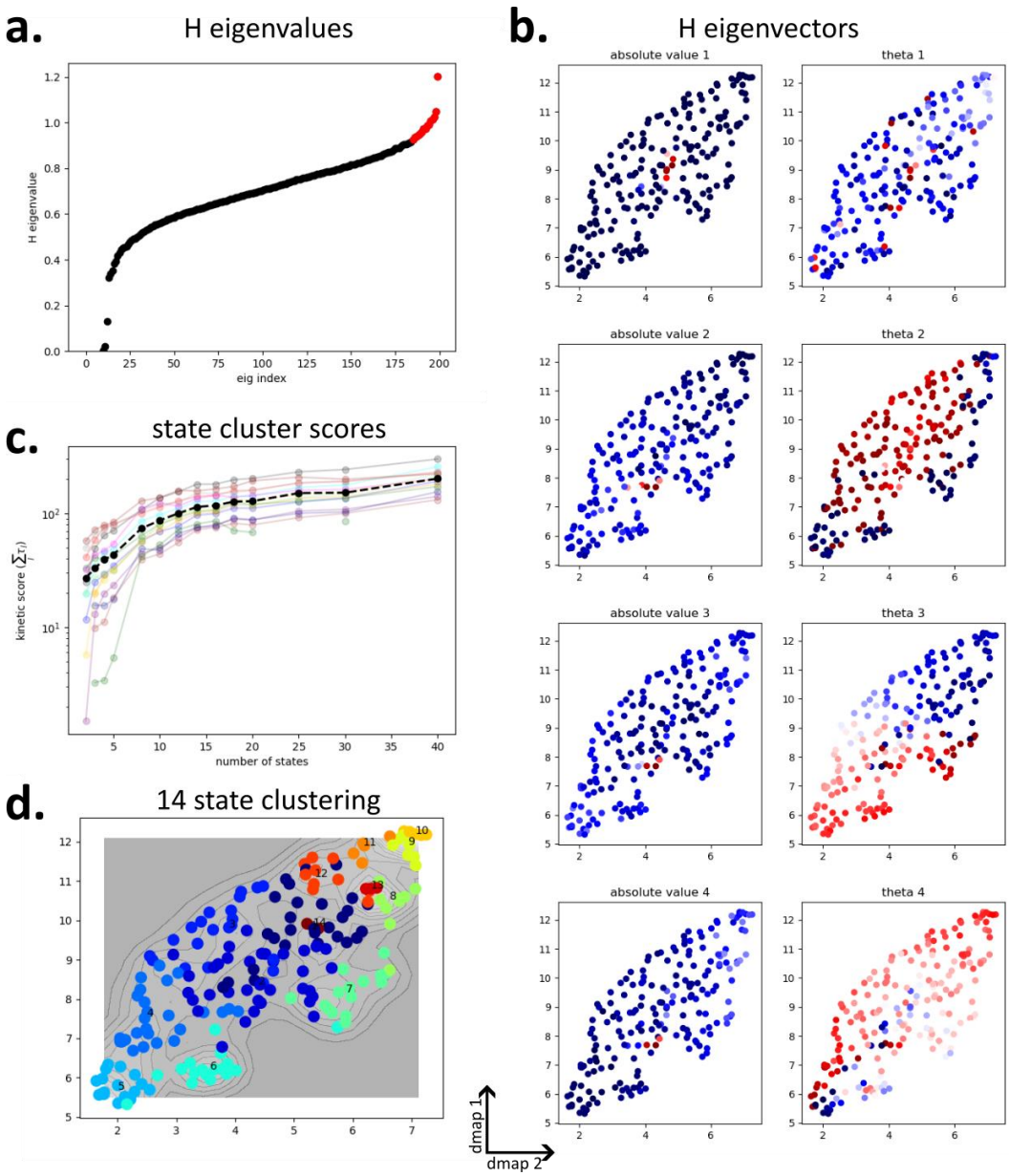
Supplementary Figure 6: EMT/MET initial and final state probabilities
Live-cell imaging inferred initial and final EMT/MET state populations as a function of time. EMT initial state 1 depopulates over time (rightmost plots), while EMT final state 5 increases over time in EGF+TGFB (lower middle), while MET final state 8 increases over time in EGF conditions (upper right).



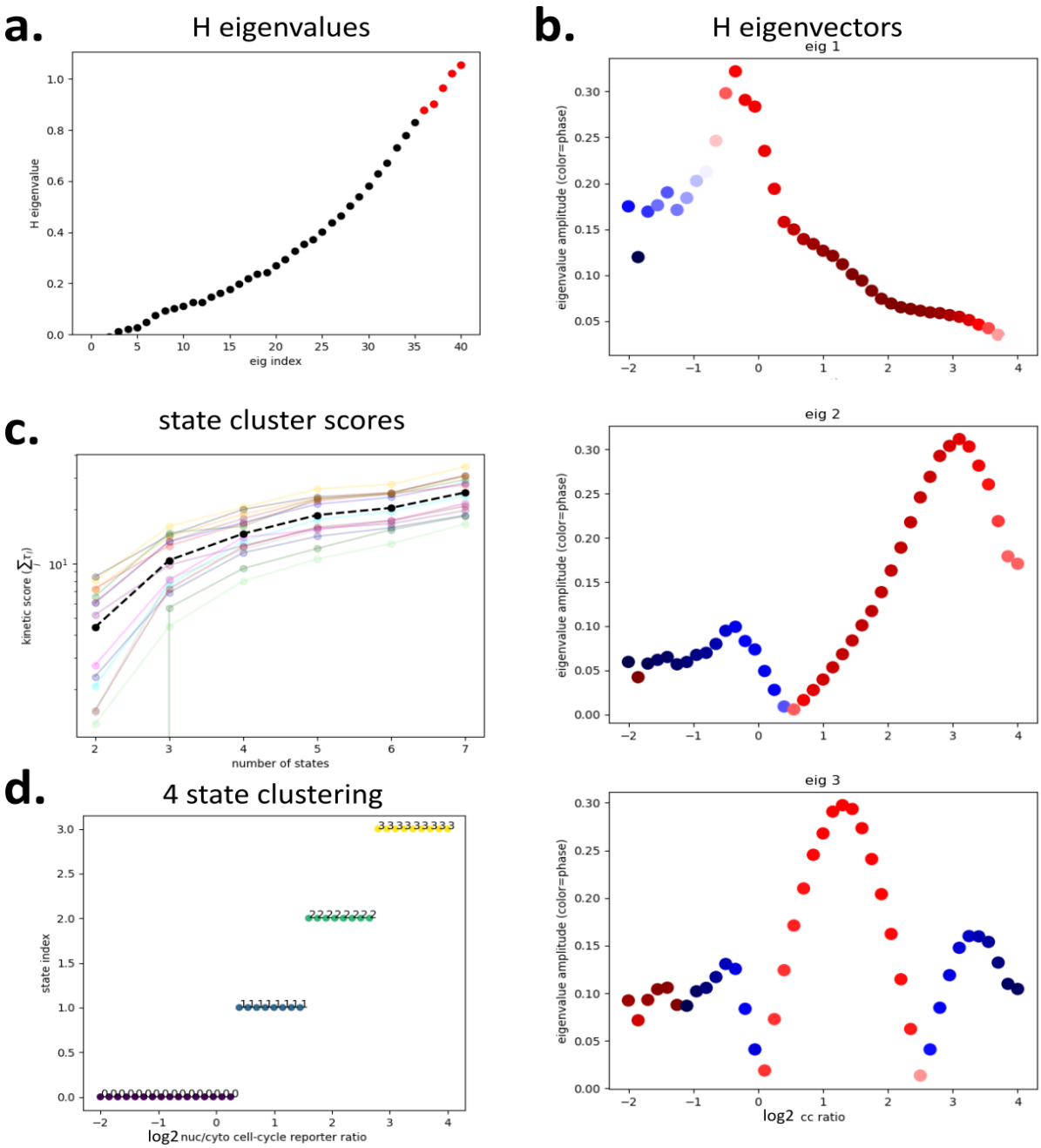
Supplementary Figure 7: EMT/MET morphodynamical cell state change sequences by ligand treatment Possible EMT cell state change sequences (initial state 1, final state 5) left, and MET cell state change sequences (initial state 5, final state 8) on the right (black arrows, thickness proportional to transition flux), with final state commitment probability (blue to red) calculated from the 200 k-means state centers and averaged over the UMAP surface.



Supplementary Figure 8: Feature batch normalization
Scatterplots of the first two PCA components for imaging experiments 1 (black) and 2 (red), in overlapping treatments, analyzed in this work. Each dot is a cell before (left) and after (right) applying our batch normalization procedure.



Supplementary Figure 9: Dynamical clustering of morphodynamical trajectories
(a) Eigenvalues of the Hermitian extension $H = \frac{1}{2}[(T + T') + i(T - T')]$ with T the transition matrix. (b) 2D UMAP of the eigenvectors of H (each point is a microstate of the transition matrix) colored by absolute value and Euler angle of the complex value. (c) State clustering dynamical information quantified by the sum of the timescales from the eigenvalues of T. The sum of timescales increases rapidly towards 15 states and begins to saturate. (d) K-means clustering into 14 states.



Supplementary Figure 10: Dynamical clustering of cell-cycle states

(a) Eigenvalues of the Hermitian extension $H = \frac{1}{2}[(T + T') + i(T - T')]$ with T the transition matrix from dividing log2 of the cell-cycle reporter levels into 51 microstates. (b) Eigenvectors of H (each point is a microstate of the transition matrix) colored by Euler angle of the complex value. (c) State clustering dynamical information quantified by the sum of the timescales from the eigenvalues of T. The sum of timescales increases rapidly towards 4 states and begins to saturate. (d) K-means clustering into 4 states.