npj | systems biology and applications

Article

Published in partnership with the Systems Biology Institute

https://doi.org/10.1038/s41540-024-00389-7

Low-frequency ERK and Akt activity dynamics are predictive of stochastic cell division events

Jamie J. R. Bennett [1], Alan D. Stern [2], Xiang Zhang[3], Marc R. Birtwistle [3] ✉ & Gaurav Pandey [1] ✉

Understanding the dynamics of intracellular signaling pathways, such as ERK1/2 (ERK) and Akt1/2 (Akt), in the context of cell fate decisions is important for advancing our knowledge of cellular processes and diseases, particularly cancer. While previous studies have established associations between ERK and Akt activities and proliferative cell fate, the heterogeneity of single-cell responses adds complexity to this understanding. This study employed a data-driven approach to address this challenge, developing machine learning models trained on a dataset of growth factor-induced ERK and Akt activity time courses in single cells, to predict cell division events. The most predictive models were developed by applying discrete wavelet transforms (DWTs) to extract low-frequency features from the time courses, followed by using Ensemble Integration, a data integration and predictive modeling framework. The results demonstrated that these models effectively predicted cell division events in MCF10A cells (F-measure=0.524, AUC=0.726). ERK dynamics were found to be more predictive than Akt, but the combination of both measurements further enhanced predictive performance. The ERK model`s performance also generalized to predicting division events in RPE cells, indicating the potential applicability of these models and our data-driven methodology for predicting cell division across different biological contexts. Interpretation of these models suggested that ERK dynamics throughout the cell cycle, rather than immediately after growth factor stimulation, were associated with the likelihood of cell division. Overall, this work contributes insights into the predictive power of intra-cellular signaling dynamics for cell fate decisions, and highlights the potential of machine learning approaches in unraveling complex cellular behaviors.

Mammalian ERK1/2 (ERK) and Akt1/2 (Akt) kinases are ubiquitous regulators of proliferation, growth and survival[1,2]. Their activity is typically controlled by upstream growth factor receptors (or other pathways affecting receptors[3,4]), and deregulated by a variety of oncogenic (e.g. EGFR, HER2, RAS, BRAF and PI3K) and tumor suppressor (e.g. PTEN and NF1) mutations[5–13]. Thus, it is not surprising that these pathways are also important drug targets for treating cancers and other diseases[14–19].

While prior work (e.g., reviewed by[1,2]) established associations between ERK and Akt activity and proliferative cell fates, single-cell signaling and response heterogeneity makes this understanding more opaque. For example, it is known that clonal cells exposed to the same conditions can exhibit markedly different ERK and Akt activity dynamics and magnitudes, as well as different proliferative (and other) fates[20–30]. It has been hypothesized that temporal differences in kinase activity profiles, such as transient vs. sustained signaling[26,31–33], oscillatory-like pulsing[22–25,34–36] and time-integrals[20,27,37–39] are associated with cell fate heterogeneity. Such kinetic control of cell fate decisions is not limited to the ERK and Akt pathways, but is also thought to be operative in systems such as p53 and NFκB[40–46]. However, a question that remains open is whether any such dynamic features of ERK and Akt activity are predictive of proliferative fates in individual cells. Answering this question can be facilitated by experiments that track kinase activities in multiple single live cells with a matched readout of the cell fate, along with rigorous data science-based modeling that can robustly assess the links between the two. While the term "cell fate" can apply to multiple phenotypes, such as differentiation and transformation, in this work, we use it to refer to cell cycle progression and cell division fate.

[1]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [2]Department of Pharmacological Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [3]Department of Chemical and Biomolecular Engineering, Clemson University, Clemson, SC, USA. ✉e-mail: mbirtwi@clemson.edu; gaurav.pandey@mssm.edu

Our previous study reported a dataset that coupled growth factor-induced ERK and Akt activity time courses to division events in the same single cells[27]. However, the question of predicting if a single, individual cell will divide or not based on such growth factor-induced time courses remains unexplored. Moreover, the previous study only considered a single cell line (MCF10A), a single set of growth factors (EGF and insulin), and acute, synchronized response to growth factors (as opposed to chronic, asynchronous scenarios). In the current work, we analyzed the previously generated dataset by utilizing time series transformations to extract features[47–49], and predicting cell divisions using effective machine learning techniques[50–52] for both MCF10A cell data with acute response to growth factors from a serum-starved state, and retinal pigment epithelial (RPE) cells asynchronously cycling in full growth media[53]. Models developed from only MCF10A data performed well on both datasets, and analyses suggested operational relationships between ERK and Akt dynamics and the probability of cell division.

Results

Data preparation

We first prepared our previously collected dataset[27] for the application of machine learning (ML) algorithms (Fig. 1a). Briefly, the dataset was generated from MCF10A cells expressing both ERK and Akt activity reporters[54,55]. Serum and growth factor-starved cells were treated with EGF and insulin (two key components of MCF10A growth media), and imaged periodically for two days (Fig. 1a). For each cell, image analysis-extracted kinase activity time course data were collected along with cell fate, the latter enabling individual cells to be assigned the label of "divided" or "undivided" that we predicted via supervised ML methods[56] (Fig. 1a). We produced two sets of data from a "high-dose" and a "low-dose" experiment, where a higher/lower dosage of EGF and insulin were used, respectively. To enable rigorous model training and evaluation, the high-dose data were randomly divided in an 80:20 ratio into train and test sets, and the "low-dose" data were used as another test set. Table 1 provides a summary of these training and test sets.

An important issue in these datasets was that divided cells had shorter time courses than undivided cells, since the original measurements terminated at the time of cell division[27]. This difference in time course length could have caused the downstream cell fate classifiers to trivially predict the outcome based on this artifact. Therefore, we processed the undivided cell time courses such that they were not trivially distinguishable from divided cell time courses (Fig. 1b; details in Methods and Supplementary Material).

Discrete wavelet transforms combined with a heterogeneous ensemble yielded the most effective classifier of MCF10A cell division fate

We employed a multi-modal data fusion approach to leverage the possible complementary information contained within ERK and Akt time courses in order to predict cell fates. We compared a diverse range of integrative methods for predicting cell fate, and also compared them to predictions from the individual ERK and Akt time courses. Throughout this paper, we refer to multi-modal methods as [ERK, Akt], and single modality time courses by the name of the corresponding kinase.

Extracting structured features that capture important information in time courses is often an effective method to enable classification[57]. We evaluated several established transformation-based feature extraction techniques: discrete wavelet transforms (DWTs)[58], MiniRocket[59], tsfresh[60] and the amplitude of the Fourier transform (FT)[61] (Fig. 2a; details in Methods). We used the features obtained from each of these transformations as input to Ensemble Integration (EI), an effective framework to develop, evaluate and interpret a suite of heterogeneous ensemble-based classifiers from multi-modal data[50]. All combinations of transformations and EI classifiers were compared using a ten-fold cross-validation strategy on the High-dose (train) dataset in terms of the F_{max} score associated with the minority (divided) class, an appropriate metric for datasets with

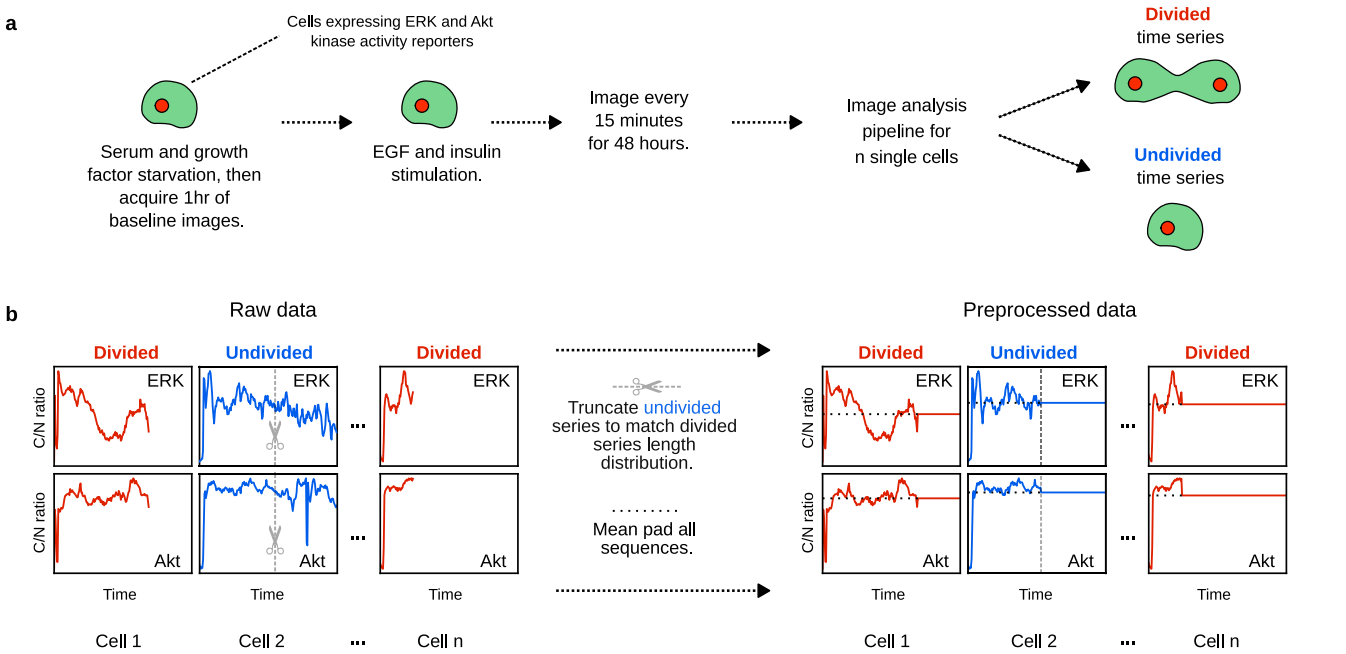


Fig. 1 | Overview of input data and their processing. a Experimental workflow of MCF10A data generation: Initially, cells expressing ERK and Akt reporters were serum- and growth factor-starved. An hour of baseline images were taken immediately before growth factor treatment. After treatment with epidermal growth factor (EGF) and insulin, cells were imaged every 15 minutes for 2 days, and the resultant image time courses were analyzed to measure kinase activity (cytoplasmic to nuclear fluorescence ratio of the kinase translocation reporter (KTR) (C/N ratio)). Time courses were labeled as divided or undivided according to the fate of the corresponding cell. **b** Pre-processing kinase time course data for machine learning analyses: Divided cells had a truncated time course at the time of cell division, whereas undivided cells did not. To address this incompatibility, all undivided cell time courses were truncated so that the distribution of time series lengths were the same between the two classes. Each time course was then padded to lengths equal to those of the undivided cell with the corresponding mean.

Table 1 | Summary of the datasets used in our study

Name	Treatment	Serum-starved?	Reporter	Number of cells			Cell type	Kinase(s) (modalities)
				Divided	Undivided	Total		
High-dose (train)	20 ng/mL EGF + 10 μg/mL Insulin	Yes	KTR	246	756	1002	MCF10A	ERK, Akt
High-dose (test)	20 ng/mL EGF + 10 μg/mL Insulin	Yes	KTR	62	189	251	MCF10A	ERK, Akt
Low-dose (test)	2 ng/mL EGF + 1 μg/mL Insulin	Yes	KTR	246	1148	1394	MCF10A	ERK, Akt
RPE (test)	62.5 nM ERKi	No	FRET	123	102	225	RPE	ERK

Treatment was either epidermal growth factor (EGF) + Insulin in the case of serum-starved cells, or ERK inhibitor (ERKi) in the case of cells asynchronously cycling in full growth medium. Reporters were either kinase translocation reporters (KTR)-, or fluorescence resonance energy transfer (FRET)-based.

imbalanced classes[62,63] and the area under the receiver operating characteristic curve (AUC) score[56,64].

The DWT can be applied successively as a cascade of transformations. We found that the approximation coefficients at level 3 (Supplementary Fig. 3) of the cascade were more predictive than other levels for our cell division prediction task (Supplementary Fig. 4). These level 3 approximation coefficients corresponded to a low frequency representation of the time courses, where time points were spaced approximately 2 hours apart. The median F_{max} score of EI classifiers with the DWT was higher than other transformations for [ERK, Akt] (0.518 vs. 0.481-0.500) and ERK (0.516 vs. 0.478-0.503), and was a close runner-up for the generally less predictive Akt modality (Fig. 2b). From this, we concluded that DWT boosted the information content of the time courses the most for the prediction of cell division events using EI. We therefore selected the best DWT+EI methods, namely stacked generalization (stacking)[56] with logistic regression for [ERK, Akt] (F_{max} = 0.542, AUC = 0.757), and stacking with random forest for ERK (F_{max} = 0.528, AUC = 0.727) and Akt (F_{max} = 0.462, AUC = 0.661). We used these DWT+EI combinations as the representative methods for further evaluation and model building from the respective time courses.

Next, we compared the above DWT+EI methods with other established prediction methods, namely (1) a deep learning-based long short-term memory (LSTM) network[65] and (2) XGBoost[51] on all processed time courses in the same cross-validation setup. The DWT+EI methods predicted cell fate more accurately than both the established methods in terms of both evaluation measures, especially for the more predictive [ERK, Akt] and ERK time courses (Fig. 2c). Thus, for each time course (ERK, Akt and [ERK, Akt]), we trained one final model using the corresponding representative DWT+EI method for predicting cell division on the entire High-dose (train) set, and proceeded with their evaluation on the test sets (Table 1).

Performance on the High-dose (test) set was close to that observed during training for each of the time courses in terms of both F-measure and AUC (Fig. 2d), indicating that the models were able to generalize to unseen data drawn from the same sample as the training data. For instance, the [ERK, Akt] model performed almost as well on the High-dose (test) dataset (F-measure = 0.524, AUC = 0.726) as it did on the training data (F_{max} = 0.542, AUC = 0.757). The models also performed comparably on the Low-dose (test) data from a separate experiment using lower growth factor concentrations (Fig. 2d; e.g., for the [ERK, Akt] model, F-measure=0.400 and AUC = 0.695). These results indicated that DWT+EI models were suitable for predicting cell division fates from ERK and Akt activity time course data in MCF10A cells.

Combined time courses of ERK and Akt activity were the most predictive of MCF10A cell division fate, but ERK was individually substantially more predictive than Akt

Throughout our results, ERK was substantially more predictive of cell fate than Akt (Fig. 2). Furthermore, the still higher performance of the [ERK, Akt] DWT+EI model consistently showed that combined information from both time courses could improve predictive performance as compared to the individual time course classifiers (Fig. 2). This suggested complementary information among the ERK and Akt time courses, which we examined further.

ERK and Akt activities across the time course were important for predicting MCF10A cell fate, albeit to different extents

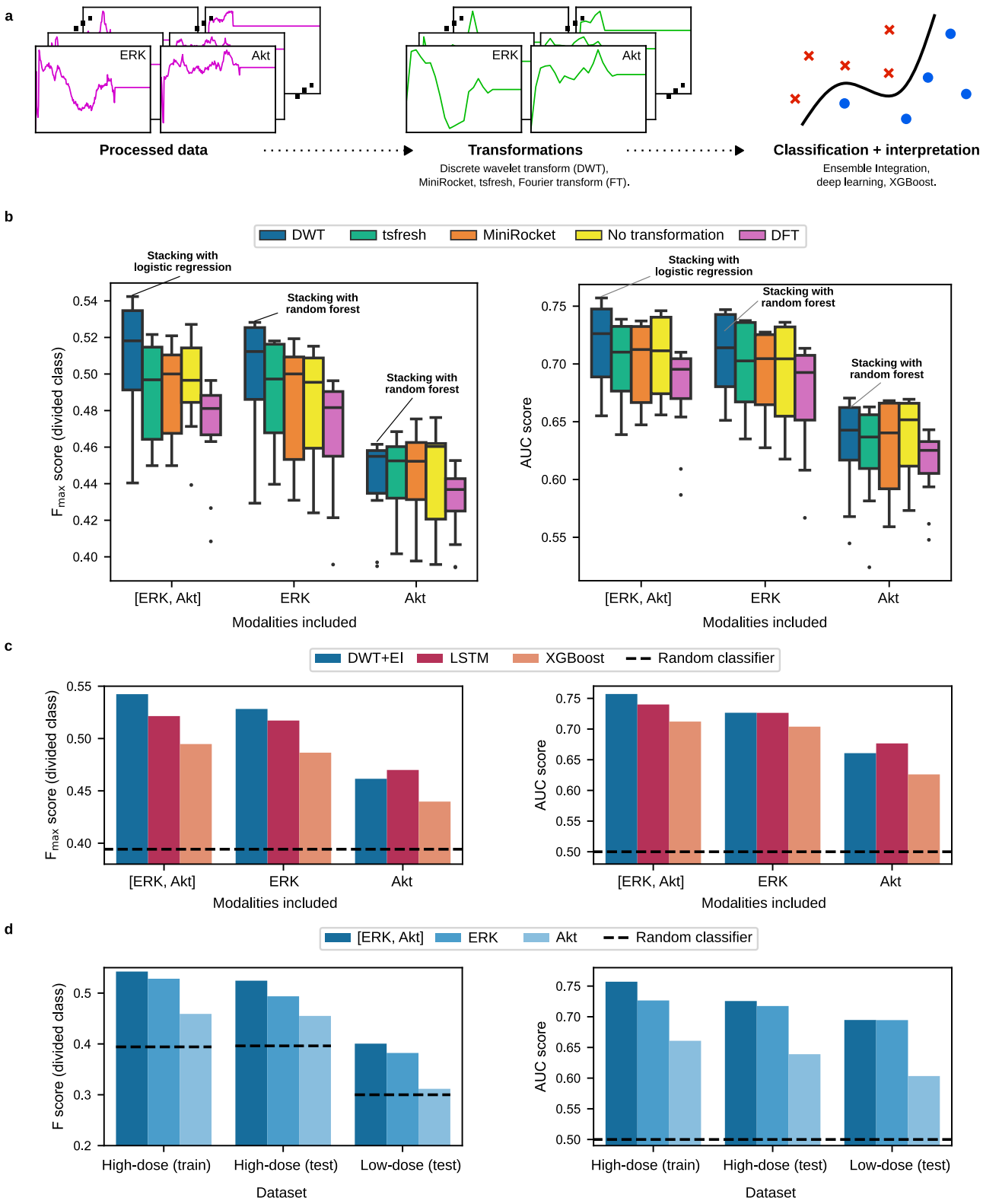
As described above, a long-standing question in signal transduction, particularly with ERK signaling, is how signaling dynamics may relate to cell fate determination[20–35,37,39]. To assess if our [ERK, Akt] model could help shed light on this phenomenon, we applied an EI-associated interpretation algorithm (Supplementary Material) to identify the time points that were the most important for predicting whether a cell would divide or not. The algorithm measured importance by quantifying the change in F_{max} of the resultant multi-modal model when a given time point was removed from the DWT`ed sequence.

This algorithm generated an importance score for each point in the low frequency ERK and Akt time course representation, which showed that the most predictive points (high values of the importance score in Fig. 3a) were distributed throughout the time courses. These scores were generally higher for ERK (median = 0.773) than Akt (median = 0.675), consistent with the relative predictive ability of the two reporters. Furthermore, the importance scores of time points in the ERK modality were correlated with the corresponding differences in median C/N ratio between the cell fate classes, though this was not really the case for the less predictive Akt (Fig. 3b; Spearman's rank correlation coefficient=0.449 and 0.093, respectively). These results show that ERK, and to a lesser extent, Akt activities throughout the cell cycle are associated with the likelihood of a cell division event, as opposed to directly after growth factor stimulation.

A potential confounder of this interpretation was if multiple subpopulations of cells were progressing through the cell cycle at different rates. To evaluate this, we analyzed the distribution of division times (Fig. 3c), which showed that they were localized towards the end of the 48 hour period in a relatively narrow window. Two important observations that follow are that the distribution of division times was much narrower than those predicted to have importance for predicting cell division, and that most cells that divided were progressing through the cell cycle at similar rates. These observations suggest that the interpretation of kinase activity being important throughout the cell cycle was likely not affected substantially by variability in cell cycle progression in dividing cells.

The ERK model generalized to predicting cell divisions in retinal pigment epithelium (RPE) cells

Our analyses thus far focused only on MCF10A cells stimulated to divide from a serum-starved condition, and ERK activity measured using a specific KTR. This raises the question of whether the models developed from this context apply to other cell types and experimental setups. To begin to answer this question, we analyzed data from a study on retinal pigment epithelium (RPE) cells containing the EKAREN5 FRET-based ERK reporter[53]. In this study, RPE cells were not serum-starved and then treated with growth factors, but rather were asynchronously cycling in full growth



medium. Applying the MCF10A-derived ERK model to this dataset (processing details in Methods) yielded F-measure and AUC scores of 0.677 and 0.694, respectively, which were comparable to those observed for the ERK time course in the High-dose (test) dataset (F-measure=0.494, AUC=0.717). Note that the higher F-measure value for the RPE (test) dataset was a result of the more balanced class distribution in this dataset, which was not the case in the other datasets we considered (Table 1). These results indicated the applicability of this model in a different biological context, as well as the utility of the DWT+EI method for model development for similar problems.

Discussion

While the biological associations between the ERK and Akt pathways and proliferative phenotypes are established, the precise differences

Fig. 2 | Cross-validated performance and testing of cell fate classification methods. a High-level overview of the machine learning analyses used. Processed data were subjected to time course transformations to enhance the signal in the ERK and Akt time courses individually, before being used as input to several machine learning classification algorithms. **b** Box plots showing the performance distribution of a family of ensembles developed by Ensemble Integration (EI) in combination with various transformations, as well as the multi-modal [ERK, Akt] time course. Centre lines and bounds of boxes correspond to the median and upper/lower quartiles, respectively. Whiskers denote maximum/minimum values within 1.5 times the interquartile range above/below the upper/lower quartile, with values outside this range (points) denoting outliers. **c** Performance of the most predictive EI algorithm combined with the Discrete Wavelet Transformation (DWT) named in **b** with arrows, as well as other established classification algorithms, namely Extreme Gradient Boosting (XGBoost) and deep learning-based Long Short Term Memory (LSTM). Both sets of results are presented in terms of the F_{max} and AUC evaluation measures. To calculate the F_{max} score, we maximized the F-measure on the training set, and then applied this threshold to discretize the test set predictions and calculate the F-measure value there. Note that the classifier with the best F_{max} score does not necessarily also have the best AUC score. The performance of a random classifier is shown for reference. **d** F-measure and AUC scores of the final DWT+EI classification models on the MCF10A train and test sets. The performance of a random classifier is shown for reference. These results show that ERK is substantially more predictive than Akt across all datasets, but utilizing both ERK and Akt time courses is even more predictive.

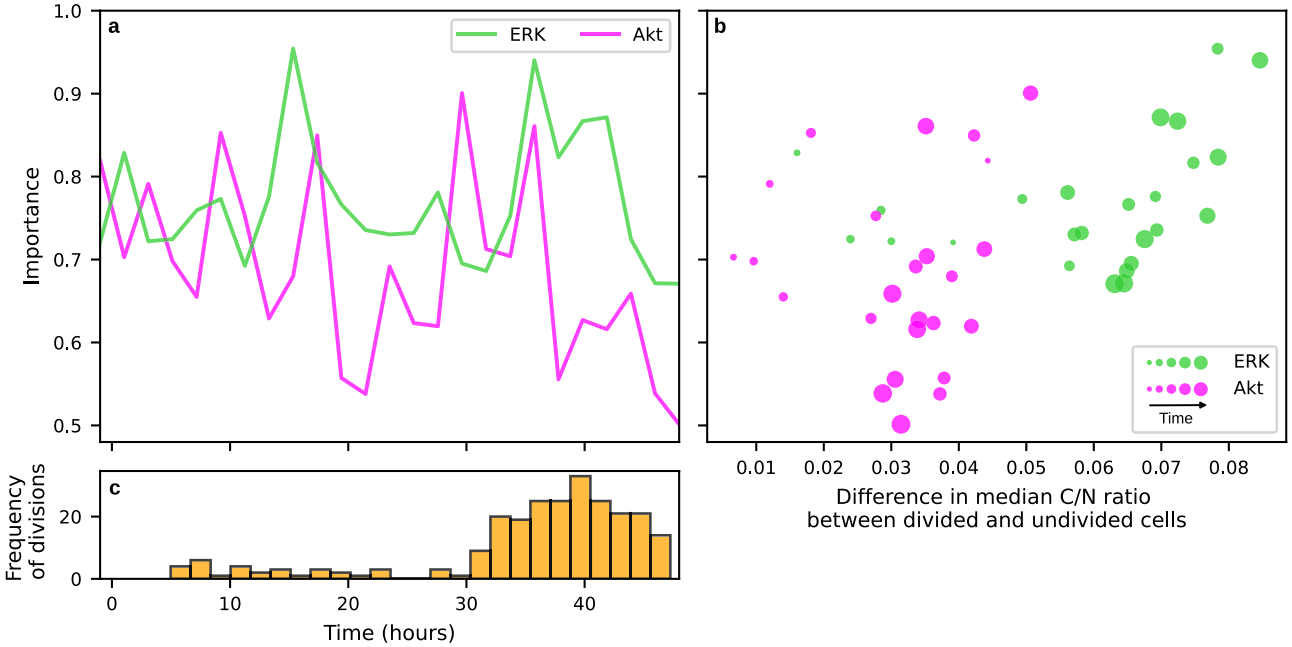


Fig. 3 | Interpretation of the multi-modal [ERK, Akt] cell fate prediction model. a The respective predictive importances of ERK and Akt activities, demonstrating that time points of high importance (higher values on the Y-axis) exist throughout the time course, in particular for ERK. **b** A scatter plot showing the importance scores of ERK and Akt activities against the differences between the median C/N ratios of the corresponding reporter for the cell fate classes (dividing and non-dividing) during the same time course, with a larger point size indicating a later time point. The most predictive/important time points in **a** were generally consistent with large differences in the median C/N ratios between cell fate classes. **c** Distribution of cell division times in the High-dose (train) dataset, showing that the division times of dividing cells were localized towards the end of the 48 hour period.

between the activities of these pathways in cells that do or do not proliferate are inadequately understood. In this study, we employed a data-driven approach to better understand the contributions of ERK and Akt kinase activities to growth factor-induced, single cell division fates by leveraging the natural cell-to-cell heterogeneity in these processes. To achieve this, we constructed supervised machine learning models to classify division events based on individual and combined ERK and Akt activity time courses measured by live-cell imaging of non-transformed breast epithelial MCF10A cells, a model system that is commonly used to study epithelial signaling biology and cell division control[3,24,66–70]. The results indicated that ERK and Akt activity time courses were predictive of cell division fate, especially when both the activities were considered together in the multi-modal Ensemble Integration (EI) framework[50]. ERK and Akt activities were somewhat correlated (r = 0.3-0.5), and this could partially explain the less than additive effects on predictive cell division fate[27]. Application of this model developed from MCF10A data to a different cell line (RPE), with a different experimental and biological context (asynchronously cycling cells in full growth medium), also showed that low-frequency ERK dynamics were predictive of cell division events.

We observed that processing the time courses with discrete wavelet transforms (DWTs)[58] prior to classification via heterogeneous ensembles yielded the best performance among the evaluated methods. Specifically, we applied a low-frequency DWT approximation to the time courses (Supplementary Fig. 3). Our finding suggested that effective classification may not require the complete detail present in the non-transformed 15-minute interval time courses, but rather the overall trend of kinase activity described by a low-frequency representation. Moreover, an inevitable characteristic of the experimental procedure used to generate the kinase activity data used in our study was the presence of noise. Due to the ability of DWTs to effectively remove noise of different frequencies, their application proved to be particularly advantageous in mitigating its impact on classification (Supplementary Fig. 4). Our results and this analysis indicate that DWTs can be a flexible, multi-resolution method to extract features from kinase activity time courses, allowing combinations of different frequency components representing variables such as cell and kinase type, as well as differences in experimental setup.

Our results also indicated that ERK dynamics were substantially more predictive of cell fate than Akt dynamics. This is consistent with the general view of the PI3K/Akt pathway being involved with growth, metabolic

regulation and survival, as opposed to direct cell cycle regulation like ERK[2,71–74]. Furthermore, the multi-modal model combining ERK and Akt dynamics to predict division events was more predictive than ERK alone. This indicates that Akt dynamics had information about cell division complementary to that in ERK, which may reflect Akt`s hypothesized regulation of some aspects of initial cell cycle entry and progression[2,75,76], but may also suggest growth and metabolic roles.

We also found that the ERK-based model developed from the starved and growth-factor induced experimental setup was predictive of division fates from chronically growth factor treated, asynchronously cycling contexts in an independent dataset[53]. This independent dataset was from a different cell line, RPE cells, which although still of epithelial origin like MCF10A, provides further evidence of the potential generality of the models developed. Recent live-cell imaging studies of ERK signaling in epithelial cells have instead looked at confluent cells in chronic growth factor treatment conditions, arguing that such a setup is closer to the epithelial biology that would be observed in tissues[23,24]. Large annotated datasets linking ERK dynamics to cell division events in confluent, chronic growth factor contexts would be needed to understand how different such scenarios truly are. While obtaining the imaging data themselves is relatively straightforward, high-quality annotated datasets like ours still often require substantial manual effort. Advances in machine learning for image analysis may be able to help generate such datasets more efficiently[77–79].

Another question we were able to address using the developed models was which time points of signaling dynamics were the most predictive of cell division events. As discussed above, there are multiple hypotheses for such relationships, including transient versus sustained signaling, the frequency and/or magnitudes of pulsing and time-integrated activities[20,22,25–27,34,37,38]. In the data analyzed here, there was not substantial evidence of pulsing, but this does not mean it might not be predictive in other contexts. Based on a systematic model interpretation algorithm, our results suggested that no single time point in the ERK and Akt time courses was solely more useful than others for predicting cell division events. This is more consistent with a time-integrator model, whereby the amount of time the pathway is on dictates the probability of cell division[20,27,37–39]. Recent studies in MCF10A cells have suggested such a relationship between ERK signaling and cell cycle progression[38]. However, proving such a relationship to be causal is challenging, since cell cycle progression depends on the activity of the ERK pathway. Optogenetic approaches are a suitable tool for such examinations, and have been used to establish relationships between ERK signaling and cell fates in other contexts[3,20,25,35].

There are multiple considerations for experimental design that could have a bearing on interpretation. Foremost, while the trained models were shown to be predictive, they were not 100% accurate, implying that other biological sources of noise, such as p53 / DNA damage[40], play operational roles in predicting cell division fate. We did not see any evidence of apoptosis in our experiments, but others have observed high vs low ERK activity leading to survival vs apoptosis, respectively[39]. In our MCF10A cell studies of Akt dynamics, large, saturating doses of insulin were used, because that is the standard for MCF10A growth media. Insulin is a strong activator of the Akt pathway[80–82]. If Akt activity was consequently very high, its fluctuations over time in a single cell may not dip below thresholds that interfere with its ability to drive cell cycle progression. Thus, growth factor conditions that do not as strongly activate the Akt pathway may lead to a different conclusion about the relative importance of ERK activity versus that of Akt for driving cell division events. Yet, we did also test the predictive models on a dataset with 10-fold lower growth factor doses, but the conclusions were similar. This could mean that insulin concentrations remained high, but it could suggest that the original interpretation of Akt activity dynamics being less predictive are more likely to be true.

There are several growth factors that induce both cell division and ERK and Akt activities. Whether the predictive relationships between ERK and/or Akt dynamics and cell division change depending on what activates the pathways remains an open question. Besides growth factor dose and type, there are other important experimental elements to consider in the context of our results. In the MCF10A experiments, somewhat sparsely seeded cells were serum- and growth factor-starved prior to the experiment, treated with growth factors, and then observed for 48 hours. Observing cells for longer could have revealed more division events, but also confounded results, as cells dividing early in the time course could have gone through the same process again. Thus, our results may be more strictly interpreted as most effective for predicting the fates of cells most likely to divide (relatively) quickly. While serum and growth factor starvation is a long-established mode of studying cell signaling[83], chronic treatment with growth factors may create a different relationship between signaling dynamics and cell division fates[4,12,24]. Yet, the models developed from starvation experiments in MCF10A cells were found to be predictive for RPE cells asynchronously cycling in full growth media, providing at least one example of generality.

In summary, our results suggest that discrete wavelet transforms of cell signaling dynamics data, combined with ensemble-based classification models, show promise as a tool for cell fate prediction. Further validation on other biological systems would be needed to establish the broader utility of this approach. Our particular application to cell division associated with ERK and Akt dynamics suggested some generality of the relationship between these dynamics and cell division for epithelial cell lines in both acute and chronic growth factor conditions. Model interpretation suggested that the time-integrator model of how these pathways influence cell division fate is more likely in the studied contexts. The availability of more annotated datasets would enable a more expansive study to understand how general such relationships are in different cell types, growth factors, doses and other experimental setups.

Methods

Dataset processing

MCF10A datasets. The MCF10A datasets were generated as described in detail previously[27]. Briefly, cells expressing reporters were seeded, allowed to attach overnight, and serum- and growth factor-starved for 24 hours. An hour of baseline images were taken before applying growth factor treatment, after which, images were taken every 15 minutes over the next 48 hours, giving a total of 49 hours of measurements. The resultant image time courses were analyzed to measure kinase activity (cytoplasmic to nuclear fluorescence ratio of the kinase translocation reporter (KTR) (C/N ratio)). Experiments in this study included biological replicates for reproducibility.

ERK and Akt KTR plasmids were obtained and used as previously described[29,55]. At least two biological replicates in each condition were performed. The imaging and image processing was performed as previously detailed[27], and is briefly described as follows. Imaging was performed with a 15 minute interval using an InCell 2200 (GE). Images were subjected to (i) flatfield correction and background subtraction using CellProfiler; (ii) nuclei identification using Ilastik; (iii) cell tracking through time using CellProfiler; (iv) cytoplasmic to nuclear fluorescence extraction using custom scripts; and (v) the identification of cell division events through KTR activity dips and manual validation.

It is appreciated that M-phase CDK activity (CDK1/CyclinB) can induce ERK-KTR responses[84–86]. We do not expect this non-specific behavior to impact our findings because M-phase is a relatively short time period in the cell cycle. Additionally, our analyses removed most time points flanking mitosis already as previously described. Specifically, cell morphology substantially changes during mitosis, which resulted in inaccurate estimates of cytoplasmic-to-nuclear fluorescence intensity ratio-a large dip not related to kinase activity.

To prepare sufficiently sized sets of examples of labeled cells to train and test machine learning methods/models, we merged the data across replicate experiments from the original study after confirming that there were no batch effects (principle component analysis; Supplementary Fig. 1). In the resultant data, experiments with a high-dose of growth factors induced more divisions, resulting in a less severe cell fate class imbalance than that in low-dose settings (24.6% divided cells for high-dose data compared to 16.6% for low-dose). To have sufficient representation of the

divided class in the data, we used the high-dose data only for both the assessment of candidate methods for cell fate classification and the subsequent final model training. We then performed an 80:20 train/test split on the high-dose data, yielding one training and two test MCF10A datasets (Table 1).

Processing steps (see Supplementary Material) were performed on the training set, and any processing parameters derived were then applied during processing of the test sets. The resulting data is summarized in Supplementary Fig. 5.

RPE dataset. The RPE experiments (part of a larger study[53]), in addition to using a different cell line, only measured ERK activity with the EKAREN5 reporter (and not Akt, or with a KTR reporter), had cells asynchronously cycling in full growth media (not initially serum-starved), and included measurements over a 4 day period with a 10 minute frequency and multiple division events along the time course (versus 2 days, 15 minute frequency, and the first division event in the MCF10A datasets). Key differences are summarized in Table 1. This study contained multiple experimental setups with different combinations of treatments. The RPE (test) dataset used in our study was generated from the experiment with no DOX and 62.5nM ERK inhibitor (ERKi), because this was the only condition where cell division fate heterogeneity was observed in the last 49 hours of the time course.

To be able to test the ERK model developed from the MCF10A dataset in this different setup, we labeled the RPE cells as divided/undivided according to whether a division occurred in the last 49 hours of their respective time courses. This also allowed us to remove cells whose division events early in their time courses reflected those committed to division prior to treatment with ERKi. We linearly interpolated each time course in RPE (test) to match the 15 minute measurements of the MCF10A datasets. We then applied the same processing steps as those described above for the MCF10A datasets. To account for different reporters, we scaled ERK activity in the resultant RPE (test) dataset to make it as consistent as possible with the measurements in the High-dose (train) dataset (details in Supplementary Material; Supplementary Fig. 5).

Classification methods and their evaluation

We evaluated a number of classification methods using a stratified, ten-fold cross-validation procedure[87] applied to the High-dose (train) dataset. We addressed class imbalance in this process by undersampling the majority (undivided) class during training, and evaluated the performance of the tested methods using the F_{max} score associated with the minority (divided) class[62,63]. This measure is defined as the maximum value of the F-measure across all classification score thresholds, allowing each classification method to achieve their most effective performance. We also report the area under the receiver operating characteristic curve (AUC) score[56,64].

We compared several approaches for classifying cell division fate in the above evaluation setup (details of the approaches in Supplementary Materials). The first approach was to extract features using a time course transformation (next subsection) prior to building predictive ensembles using the Ensemble Integration framework[50], which is designed to flexibly and effectively integrate multi-modal data like the ERK and Akt time courses. Secondly, we considered a deep learning-based Long Short Term Memory (LSTM)[88] classification method whose architecture was specifically developed to handle sequential dependencies in the time courses. Finally, we considered eXtreme Gradient Boosting (XGBoost;[51]), which has been found to be the most effective performer in many classification tasks[89–91].

A particular interest of our study was to compare multi-modal methods/models that utilized information from both ERK and Akt time courses to those built only using the information in the individual time courses. Thus, we evaluated all the classification methods considered on all three modalities, i.e, ERK, Akt and [ERK, Akt], with the aim of building three different final models. For each of the modalities, we calculated the median F_{max} scores of the EI methods tested for each transformation. In addition to analyzing the performance trends across the transformation+EI method combinations, we also selected the best-performing combination for each modality for final model building. The LSTM-based neural network and XGBoost performances were similarly evaluated and compared.

Time course transformations

To more effectively extract information from the ERK and Akt time courses prior to cell fate classification, we applied a selection of transformations[57] to them before evaluating their predictive capabilities using EI. We give a brief summary of the selected transformations below.

Discrete Fourier transform. The Discrete Fourier Transform (DFT; see, for example,[61]) transforms a signal from the time domain into the frequency domain. We computed the (real-valued) amplitudes of the DFT to use as features for classification.

Discrete wavelet transform. Discrete wavelet transforms (DWTs) decompose time series into both time and frequency components simultaneously[58]. The DWT decomposes the signal into "approximation" and "detail" coefficients. The former represent the coarse, low-frequency components of the signal, while the latter capture high-frequency information. DWTs can be applied successively as a cascade (Supplementary Fig. 3), where each level of the transformation removes progressively more high frequency detail. We used the PyWavelets package[92] to compute the approximation coefficients of a DWT, resulting in a denoised, low-dimensional representation of the signal. We used the Haar wavelet because it is the simplest wavelet, and is effective at detecting sudden transitions. This transformation yielded a set of 25 time points spaced roughly 2 hours apart for our time courses.

Minirocket. Minirocket is a time course transformation that generates features through convolution with a predefined set of kernels[59]. The result is a set of binary features indicating the presence or absence of certain patterns identified by the kernels over specific time intervals. Temporal and frequency domain characteristics are therefore not explicitly retained. We used the Sktime package[93] to implement Minirocket with default parameters to generate features.

Tsfresh. We also used the tsfresh package[60] to generate hundreds of statistical features characterizing the time courses. These features range from simple statistics, such as mean and maximum value, to more complex ones, such as descriptive statistics of the autocorrelation. We used all available features for classification.

Ensemble Integration

We utilized Ensemble Integration (EI)[50,94] to compare several interpretable ensemble methods and train our final models based on the best performers. Within the EI process for each modality, we trained a selection of base classifiers, and combined their predictions via an ensemble classifier. The specific ensemble classifier was selected from a variety of ensemble algorithms via a nested cross validation. Although EI is able to build ensembles from single data modalities (i.e. ERK or Akt alone), it is especially effective for data from multiple sources (multi-modal data), the [ERK, Akt] modality in this study. The exact base and ensemble predictor algorithms and parameters that were used in this study can be found in Supplementary Material.

Interpretation algorithm for EI models

We calculated an importance score for the time points of the discrete wavelet-transformed time courses constituting the best performing [ERK, Akt] model (Fig. 3), i.e., stacked generalization with logistic regression (Results). The algorithm used for this calculation utilized a randomized permutation procedure[95] to calculate feature ranks of each base classifier algorithm, before applying a weighted average over all the time points based on the learned weights of the logistic regression stacker (Supplementary Alg. 1). The output of this algorithm was an importance measure between 0 and

1 for each time point, where values closer to 1 indicate higher importance (more details in Supplementary Material).

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

MCF10A datasets used in this study were the dual-reporter datasets presented in a previous paper[27], and are freely available at the repository mentioned therein. The RPE datasets were produced independently[53], and were obtained by contacting the authors. With permission, we have shared all the relevant data related to this study within the following GitHub repository: https://github.com/GauravPandeyLab/predicting-cell-division.

Code availability

Our analysis is reproducible by following the instructions at https://github.com/GauravPandeyLab/predicting-cell-division.git. The code for the image analysis pipeline that produced the MCF10A datasets is available at https://github.com/birtwistlelab/Predicting-Individual-Cell-Division-Events-from-Single-Cell-ERK-and-Akt-Dynamics.

Received: 29 January 2024; Accepted: 20 May 2024;
Published online: 04 June 2024

References

1. Lavoie, H., Gagnon, J. & Therrien, M. Erk signalling: a master regulator of cell behaviour, life and fate. *Nat. Rev. Mol. Cell Biol.* **21**, 607–632 (2020).
2. Manning, B. D. & Toker, A. Akt/pkb signaling: navigating the network. *Cell* **169**, 381–405 (2017).
3. Farahani, P. E. et al. Substratum stiffness regulates erk signaling dynamics through receptor-level control. *Cell Rep.* **37** (2021).
4. Sparta, B. et al. Receptor level mechanisms are required for epidermal growth factor (egf)-stimulated extracellular signal-regulated kinase (erk) activity pulses. *J. Biol. Chem.* **290**, 24784–24792 (2015).
5. Rosell, R. et al. Coregulation of pathways in lung cancer patients with egfr mutation: therapeutic opportunities. *Br. J. Cancer* **125**, 1602–1611 (2021).
6. Ercan, D. et al. Reactivation of erk signaling causes resistance to egfr kinase inhibitors. *Cancer Discov.* **2**, 934–947 (2012).
7. Jacobsen, K. et al. Convergent akt activation drives acquired egfr inhibitor resistance in lung cancer. *Nat. Commun.* **8**, 410 (2017).
8. Ruiz-Saenz, A. et al. Her2 amplification in tumors activates pi3k/akt signaling independent of her3. *Cancer Res.* **78**, 3645–3658 (2018).
9. Bourguignon, L. Y., Gilad, E. & Peyrollier, K. Heregulin-mediated erbb2-erk signaling activates hyaluronan synthases leading to cd44-dependent ovarian tumor cell growth and migration. *J. Biol. Chem.* **282**, 19426–19441 (2007).
10. Gerosa, L. et al. Receptor-driven erk pulses reconfigure mapk signaling and enable persistence of drug-adapted braf-mutant melanoma cells. *Cell Syst.* **11**, 478–494 (2020).
11. Ryan, M. B., Der, C. J., Wang-Gillam, A. & Cox, A. D. Targeting ras-mutant cancers: is erk the key? *Trends Cancer* **1**, 183–198 (2015).
12. Gillies, T. E. et al. Oncogenic mutant ras signaling activity is rescaled by the erk/mapk pathway. *Mol. Syst. Biol.* **16**, e9518 (2020).
13. Davies, A. E. et al. Systems-level properties of egfr-ras-erk signaling amplify local signals to generate dynamic gene expression heterogeneity. *Cell Syst.* **11**, 161–175 (2020).
14. Fruman, D. A. et al. The pi3k pathway in human disease. *Cell* **170**, 605–635 (2017).
15. Mayer, I. A. & Arteaga, C. L. The pi3k/akt pathway as a target for cancer treatment. *Annu. Rev. Med.* **67**, 11–28 (2016).
16. Georgescu, M.-M. Pten tumor suppressor network in pi3k-akt pathway control. *Genes Cancer* **1**, 1170–1177 (2010).
17. Kiuru, M. & Busam, K. J. The nf1 gene in tumor syndromes and melanoma. *Lab. Investig.* **97**, 146–157 (2017).
18. Lee, S., Rauch, J. & Kolch, W. Targeting mapk signaling in cancer: mechanisms of drug resistance and sensitivity. *Int. J. Mol. Sci.* **21**, 1102 (2020).
19. Kohno, M. & Pouyssegur, J. Targeting the erk signaling pathway in cancer therapy. *Ann. Med.* **38**, 200–211 (2006).
20. Johnson, H. E. & Toettcher, J. E. Signaling dynamics control cell fate in the early drosophila embryo. *Dev. Cell* **48**, 361–370 (2019).
21. Von Kriegsheim, A. et al. Cell fate decisions are specified by the dynamic erk interactome. *Nat. Cell Biol.* **11**, 1458–1464 (2009).
22. Pokrass, M. J. et al. Cell-cycle-dependent erk signaling dynamics direct fate specification in the mammalian preimplantation embryo. *Dev. Cell* **55**, 328–340 (2020).
23. Ryu, H. et al. Frequency modulation of erk activation dynamics rewires cell fate. *Mol. Syst. Biol.* **11**, 838 (2015).
24. Albeck, J. G., Mills, G. B. & Brugge, J. S. Frequency-modulated pulses of erk activity transmit quantitative proliferation signals. *Mol. Cell* **49**, 249–261 (2013).
25. Goglia, A. G. et al. A live-cell screen for altered erk dynamics reveals principles of proliferative control. *Cell Syst.* **10**, 240–253 (2020).
26. Chen, J.-Y., Lin, J.-R., Cimprich, K. A. & Meyer, T. A two-dimensional erk-akt signaling code for an ngf-triggered cell-fate decision. *Mol. Cell* **45**, 196–209 (2012).
27. Stern, A. D. et al. Relating individual cell division events to single-cell ERK and akt activity time courses. *Sci. Rep.* **12**, 18077 (2022).
28. Gross, S. M. & Rotwein, P. Akt signaling dynamics in individual cells. *J. Cell Sci.* **128**, 2509–2519 (2015).
29. Gross, S. M., Dane, M. A., Bucher, E. & Heiser, L. M. Individual cells can resolve variations in stimulus intensity along the igf-pi3k-akt signaling axis. *Cell Syst.* **9**, 580–588 (2019).
30. Fey, D., Croucher, D. R., Kolch, W. & Kholodenko, B. N. Research topic: From structural to molecular systems biology: experimental and computational approaches to unravel mechanisms of kinase activity regulation in cancer and neurodegeneration: Crosstalk and signaling switches in mitogen-activated protein kinase cascades. *Front. Physiol.* **3** (2012).
31. Cowley, S., Paterson, H., Kemp, P. & Marshall, C. J. Activation of map kinase kinase is necessary and sufficient for pc12 differentiation and for transformation of nih 3t3 cells. *Cell* **77**, 841–852 (1994).
32. Traverse, S. et al. Egf triggers neuronal differentiation of pc12 cells that overexpress the egf receptor. *Curr. Biol.* **4**, 694–701 (1994).
33. Gross, S. M. & Rotwein, P. Mapping growth-factor-modulated akt signaling dynamics. *J. Cell Sci.* **129**, 2052–2063 (2016).
34. Nakakuki, T. et al. Ligand-specific c-fos expression emerges from the spatiotemporal control of erbb network dynamics. *Cell* **141**, 884–896 (2010).
35. Bugaj, L. et al. Cancer mutations and targeted drugs can disrupt dynamic signal encoding by the ras-erk pathway. *Science* **361**, eaao3048 (2018).
36. Aikin, T. J., Peterson, A. F., Pokrass, M. J., Clark, H. R. & Regot, S. Mapk activity dynamics regulate non-cell autonomous effects of oncogene expression. *eLife* **9**, e60541 (2020).
37. Arekatla, G. et al. Optogenetic manipulation identifies the roles of erk and akt dynamics in controlling mouse embryonic stem cell exit from pluripotency. *Dev. Cell* **58**, 1022–1036 (2023).
38. Min, M., Rong, Y., Tian, C. & Spencer, S. L. Temporal integration of mitogen history in mother cells controls proliferation of daughter cells. *Science* **368**, 1261–1265 (2020).
39. Gagliardi, P. A. et al. Collective erk/akt activity waves orchestrate epithelial homeostasis by driving apoptosis-induced survival. *Dev. Cell* **56**, 1712–1726 (2021).
40. Purvis, J. E. et al. p53 dynamics control cell fate. *Science* **336**, 1440–1444 (2012).

41. Stewart-Ornstein, J. & Lahav, G. p53 dynamics in response to dna damage vary across cell lines and are shaped by efficiency of dna repair and activity of the kinase atm. *Sci. Signal.* **10**, eaah6671 (2017).
42. Lahav, G. et al. Dynamics of the p53-mdm2 feedback loop in individual cells. *Nat. Genet.* **36**, 147–150 (2004).
43. Son, M. et al. Spatiotemporal nf-κb dynamics encodes the position, amplitude, and duration of local immune inputs. *Sci. Adv.* **8**, eabn6240 (2022).
44. Tay, S. et al. Single-cell nf-κb dynamics reveal digital activation and analogue information processing. *Nature* **466**, 267–271 (2010).
45. Cheng, Q. J. et al. Nf-κb dynamics determine the stimulus specificity of epigenomic reprogramming in macrophages. *Science* **372**, 1349–1353 (2021).
46. Adelaja, A. et al. Six distinct nfκb signaling codons convey discrete information to distinguish stimuli and enable appropriate macrophage responses. *Immunity* **54**, 916–930 (2021).
47. AlSharabi, K., Salamah, Y. B., Abdurraqeeb, A. M., Aljalal, M. & Alturki, F. A. Eeg signal processing for alzheimer's disorders using discrete wavelet transform and machine learning approaches. *IEEE Access* **10**, 89781–89797 (2022).
48. Hadrian, A., Vainshtein, R., Shapira, B. & Rokach, L. Deepcan: Hybrid method for road type classification using vehicle sensor data for smart autonomous mobility. IEEE Transactions on Intelligent Transportation Systems (2023).
49. Fatimah, B., Singh, P., Singhal, A. & Pachori, R. B. Detection of apnea events from ecg segments using fourier decomposition method. *Biomed. Signal Process. Control* **61**, 102005 (2020).
50. Li, Y. C., Wang, L., Law, J. N., Murali, T. M. & Pandey, G. Integrating multimodal data through interpretable heterogeneous ensembles. *Bioinforma. Adv.* **2**, vbac065 (2022).
51. Chen, T. & Guestrin, C. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. http://arxiv.org/abs/1603.02754. 1603.02754 [cs].
52. Han, Z., Zhao, J., Leung, H., Ma, K. F. & Wang, W. A review of deep learning models for time series prediction. *IEEE Sens. J.* **21**, 7833–7848 (2019).
53. Chen, J.-Y. et al. Multi-range erk responses shape the proliferative trajectory of single cells following oncogene induction. *Cell reports***42** (2023).
54. Gross, S. M., Dane, M. A., Bucher, E. & Heiser, L. M. Individual cells can resolve variations in stimulus intensity along the IGF-PI3k-AKT signaling axis. *Cell Syst.* **9**, 580–588 (2019).
55. Regot, S., Hughey, J. J., Bajar, B. T., Carrasco, S. & Covert, M. W. High-sensitivity measurements of multiple kinase activities in live single cells. *Cell* **157**, 1724–1734 (2014).
56. Alpaydin, E.*Introduction to machine learning* (MIT press, 2020).
57. Ruiz, A. P., Flynn, M., Large, J., Middlehurst, M. & Bagnall, A. The great multivariate time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data Min. Knowl. Discov.* **35**, 401–449 (2021).
58. Van Fleet, P. J. *Discrete wavelet transformations: An elementary approach with applications* (John Wiley & Sons, 2019).
59. Dempster, A., Schmidt, D. F. & Webb, G. I. MINIROCKET: A very fast (almost) deterministic transform for time series classification. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 248–257. http://arxiv.org/abs/2012.08791. 2012.08791 [cs, stat].
60. Christ, M., Braun, N., Neuffer, J. & Kempa-Liehr, A. W. Time series FeatuRe extraction on basis of scalable hypothesis tests (tsfresh - a python package). *Neurocomputing* **307**, 72–77 (2018).
61. Sundararajan, D. *The discrete Fourier transform: theory, algorithms and applications* (World Scientific, 2001).
62. Radivojac, P. et al. A large-scale evaluation of computational protein function prediction. *Nat. methods* **10**, 221–227 (2013).
63. Zhou, N. et al. The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome Biol.* **20**, 1–23 (2019).
64. Lever, J., Krzywinski, M. & Altman, N. Points of significance: model selection and overfitting. *Nat. methods* **13**, 703–705 (2016).
65. Hochreiter, S. & Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **9**, 1735–1780 (1997).
66. Soule, H. D. et al. Isolation and characterization of a spontaneously immortalized human breast epithelial cell line, mcf-10. *Cancer Res.* **50**, 6075–6086 (1990).
67. 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**, 1066 (2022).
68. Min, M. & Spencer, S. L. Spontaneously slow-cycling subpopulations of human cells originate from activation of stress-response pathways. *PLoS Biol.* **17**, e3000178 (2019).
69. Spencer, S. L. et al. The proliferation-quiescence decision is controlled by a bifurcation in cdk2 activity at mitotic exit. *Cell* **155**, 369–383 (2013).
70. Ender, P. et al. Spatiotemporal control of erk pulse frequency coordinates fate decisions during mammary acinar morphogenesis. *Developmental cell* **57**, 2153–2167 (2022).
71. Plas, D. R. & Thompson, C. B. Akt-dependent transformation: there is more to growth than just surviving. *Oncogene* **24**, 7435–7442 (2005).
72. Hoxhaj, G. & Manning, B. D. The pi3k–akt network at the interface of oncogenic signalling and cancer metabolism. *Nat. Rev. Cancer* **20**, 74–88 (2020).
73. Lien, E. C., Lyssiotis, C. A. & Cantley, L. C. Metabolic reprogramming by the pi3k-akt-mtor pathway in cancer. *Metabolism in Cancer* 39–72 (2016).
74. He, Y. et al. Targeting pi3k/akt signal transduction for cancer therapy. *Signal Transduct. Target. Ther.* **6**, 425 (2021).
75. Xu, N., Lao, Y., Zhang, Y., Gillespie, D. A. et al. Akt: a double-edged sword in cell proliferation and genome stability. *Journal of oncology***2012** (2012).
76. Chang, F. et al. Involvement of pi3k/akt pathway in cell cycle progression, apoptosis, and neoplastic transformation: a target for cancer chemotherapy. *Leukemia* **17**, 590–603 (2003).
77. Robitaille, M. C., Byers, J. M., Christodoulides, J. A. & Raphael, M. P. Self-supervised machine learning for live cell imagery segmentation. *Commun. Biol.* **5**, 1162 (2022).
78. Pylvänäinen, J. W., Gómez-de Mariscal, E., Henriques, R. & Jacquemet, G. Live-cell imaging in the deep learning era. *Curr. Opin. Cell Biol.* **85**, 102271 (2023).
79. Van Valen, D. A. et al. Deep learning automates the quantitative analysis of individual cells in live-cell imaging experiments. *PLoS computational Biol.* **12**, e1005177 (2016).
80. Hermann, C., Assmus, B., Urbich, C., Zeiher, A. M. & Dimmeler, S. Insulin-mediated stimulation of protein kinase akt: A potent survival signaling cascade for endothelial cells. *Arteriosclerosis, thrombosis, Vasc. Biol.* **20**, 402–409 (2000).
81. Hemmings, B. A. & Restuccia, D. F. Pi3k-pkb/akt pathway. *Cold Spring Harb. Perspect. Biol.* **4**, a011189 (2012).
82. Boucher, J., Kleinridders, A. & Kahn, C. R. Insulin receptor signaling in normal and insulin-resistant states. *Cold Spring Harb. Perspect. Biol.* **6**, a009191 (2014).
83. Pirkmajer, S. & Chibalin, A. V. Serum starvation: caveat emptor. *Am. J. Physiol.-Cell Physiol.* **301**, C272–C279 (2011).
84. Ponsioen, B. et al. Quantifying single-cell erk dynamics in colorectal cancer organoids reveals egfr as an amplifier of oncogenic mapk pathway signalling. *Nat. cell Biol.* **23**, 377–390 (2021).
85. Aoki, K. et al. Stochastic erk activation induced by noise and cell-to-cell propagation regulates cell density-dependent proliferation. *Mol. cell* **52**, 529–540 (2013).

86. Wilcockson, S. G., Guglielmi, L., Rodriguez, P. A., Amoyel, M. & Hill, C. S. An improved erk biosensor detects oscillatory erk dynamics driven by mitotic erasure during early development. *Dev. Cell* **58**, 2802–2818 (2023).
87. Arlot, S. & Celisse, A. A survey of cross-validation procedures for model selection (2010).
88. Hochreiter, S. & Schmidhuber, J. Lstm can solve hard long time lag problems. In Mozer, M., Jordan, M. & Petsche, T. (eds.) *Advances in Neural Information Processing Systems*, vol. 9 (MIT Press, 1996). https://proceedings.neurips.cc/paper_files/paper/1996/file/a4d2f0d23dcc84ce983ff9157f8b7f88-Paper.pdf.
89. Liew, X. Y., Hameed, N. & Clos, J. An investigation of xgboost-based algorithm for breast cancer classification. *Mach. Learn. Appl.* **6**, 100154 (2021).
90. Torlay, L., Perrone-Bertolotti, M., Thomas, E. & Baciu, M. Machine learning–xgboost analysis of language networks to classify patients with epilepsy. *Brain Inform.* **4**, 159–169 (2017).
91. Le, N. Q. K. et al. Xgboost improves classification of mgmt promoter methylation status in idh1 wildtype glioblastoma. *J. Personalized Med.* **10**, 128 (2020).
92. Lee, G., Gommers, R., Waselewski, F., Wohlfahrt, K. & O'Leary, A. Pywavelets: A python package for wavelet analysis. *J. Open Source Softw.* **4**, 1237 (2019).
93. Löning, M. et al. sktime: A unified interface for machine learning with time series. *arXiv preprint arXiv:1909.07872* (2019).
94. Bennett, J. J. R., Li, Y. C. & Pandey, G. ensemble-integration. https://pypi.org/project/ensemble-integration/0.1.2/ Accessed: November 5, 2023 (2023).
95. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).

Acknowledgements

We thank Jia-Yun Chen, Peter Sorger and Galit Lahav for sharing the RPE dataset generated in their study, as well as associated information. We also thank Alexander Davies and Rawan Makkawi for helpful discussions. This work was supported by NIH grants R35GM141891 to MRB, and R01HG011407 and U01CA271318 to GP. It was also supported in part by Oracle Cloud credits and related resources provided by the Oracle for Research program.

Author contributions

J.J.R.B performed the analysis. ADS and XZ performed the initial imaging data acquisition and analysis and provided guidance on use of the data for this study. J.J.R.B, M.R.B and G.P wrote the manuscript. M.R.B and G.P supervised the study. All authors discussed the results and contributed to the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41540-024-00389-7.

Correspondence and requests for materials should be addressed to Marc R. Birtwistle or Gaurav Pandey.