Control of cell state transitions

Citation for published version:
Rukhlenko, OS, Halasz, M, Rauch, N, Zhernovkov, V, Prince, T, Wynne, K, Maher, S, Kashdan, E, Macleod, K, Carragher, NO, Kolch, W & Kholodenko, BN 2022, 'Control of cell state transitions', *Nature*. https://doi.org/10.1038/s41586-022-05194-y

Digital Object Identifier (DOI):
10.1038/s41586-022-05194-y

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Nature

Control of cell state transitions

Oleksii S. Rukhlenko[1], Melinda Halasz[1&], Nora Rauch[1&], Vadim Zhernovkov[1], Thomas Prince[1], Kieran Wynne[1], Stephanie Maher[1], Eugene Kashdan[1], Kenneth MacLeod[3], Neil O. Carragher[3], Walter Kolch[1,2] and Boris N. Kholodenko[1,2,4,*]

[1]Systems Biology Ireland, School of Medicine, University College Dublin, Belfield, Dublin 4, Ireland

[2]Conway Institute of Biomolecular and Biomedical Research, University College Dublin, Belfield, Dublin 4, Ireland

[3]Cancer Research UK Edinburgh Centre, Institute of Genetics and Molecular Medicine, The University of Edinburgh, Edinburgh, UK.

[4]Department of Pharmacology, Yale University School of Medicine, New Haven, USA

[&] Equal contribution.

*Corresponding author. Email: boris.kholodenko@ucd.ie

Supplementary Information

Testing cSTAR predictions using unbiased mass spectrometry (MS) acquired datasets.

The reverse phase protein array (RPPA) dataset contains a limited number of protein features selected based on prior knowledge and detected by validated antibodies. MS based phosphoproteomics allows for the unbiased acquisition of much greater amounts of data characterizing cell states. Importantly, most of the phosphosites detected by MS have no known functions and thus also add potentially uninformative data to the analysis. Therefore, we tested cSTAR with MS data to assess how well it scales and copes with a background of uninformative data. Using the same conditions as in the RPPA dataset, we acquired quantitative MS datasets for TrkA and TrkB cells. Ca. 5000 phosphosites were reliably detected (localization score > 0.75, Table S6), and fold changes relative to control were calculated (Methods). Similar as for the RPPA dataset, the SVM separated data points corresponding to a NGF-stimulated TrkA differentiation state and a BDNF-stimulated TrkB proliferation state in the MS dataspace (Extended Data Fig. 15 and Table S7, the sheet "confusion_matrix"). To capture core network components controlling cell fate outcomes, we built the STV in a ca. 5000-dimensional space (Table S7). After selecting the top STV contributors, we applied Kinase Enrichment Analysis (KEA) in order to find kinase-phosphosite relationships. We used the latest edition of KEA (KEA-2018), which is publicly available as a part of the X2K project (https://maayanlab.cloud/X2K/). The Fisher exact test was used to compute kinase enrichment p-values[1]. By selecting over-represented kinases, we identified components of a core signaling network (and Table S8). Most of these components match the kinases found by the STV ranking of RPPA data, but additionally, p38 and cell cycle kinases, CDK1, CDK4 and CK2α were found. These kinases were not included in the RPPA, and it is not surprising that the much larger MS dataset identified additional core network components. However, our experiments showed that the inhibition of p38 does not change the percentage of differentiated TrkA and TrkB cells (Extended Data Figs. 14A and 14B). This conflicting result is likely due to the similarity between the p38 and ERK phosphorylation site motifs, which cannot be confidently distinguished, likely leading to a spurious representation of p38 in the kinase enrichment analysis. The cell cycle kinases, CDK1, CDK4 and CK2α, are regulated by upstream signaling kinases, which are already present in our core network. Because cSTAR analyses the responses of both core and global network components, it is reasonable to include CDK1, CDK4 and CK2α in the DPD module that describes cell cycle and other cell machinery modules. After removing the core network components (12 phosphosites) from the STV, we calculated the DPD that represents the overall signaling network downstream of the core components. The calculated DPD values were -49.7 and +49.7 (in arbitrary distance units in the space of ~5000 phosphosites) for TrkA and TrkB cells stimulated with NGF and BDNF, respectively.

A key prediction of our dynamic model was the synergy between inhibitors of the ERBB family and the ERK pathway in inducing differentiation of TrkB cells. To test this prediction on MS data, we calculated the changes in the DPD for Gefitinib and Trametinib given separately or in combination with a half dose of each inhibitor. The inhibitor combination decreases the TrkB DPD values more effectively than each inhibitor on its own, thereby driving TrkB cells into the differentiation phenotype, whereas this combination does not change the DPD of TrkA cells (Fig. 6D). Applying the Bliss synergy score to the DPD changes showed statistically significant synergy between Gefitinib and Trametinib for both datasets of TrkB samples that were analyzed on different MS machines (Fig. 6D and Supplementary Experimental Procedures). Thus, the DPD changes, which were calculated for the whole cell signaling pattern of ca. 5000 phosphosites, suggest that Trametinib and Gefitinib selectively and synergistically induce differentiation in TrkB cells without affecting the state of TrkA cells. These results confirm the conclusions of the RPPA analysis. We conclude that (i) cSTAR is a robust and scalable method, which can handle much bigger, global omics type datasets of different nature; and that (ii) cSTAR can extract salient information out of large unbiased data sets where much of the information is not associated with the specific biological changes. Thus, cSTAR gives robust and reproducible results even when the input data differ vastly in scale and bias.

Applying cSTAR to Epithelial-Mesenchymal Transition (EMT).

cSTAR quantifies cell phenotypic changes by the DPD. This opens the possibility to integrate different omics datasets by comparing the normalized DPD changes in response to perturbations. To test this, we applied cSTAR to two datasets that analyzed the suppression of EMT by kinase inhibitors. One used single-cell RNA sequencing (scRNA-seq) of four cancer cell lines (A549, DU145, MCF7 and OVCA420), each stimulated with three different ligands, TGFβ, EGF and TNFα[2]. The other study used single-cell resolution mass cytometry of phosphoproteomic responses in Py2T breast cancer cells stimulated with TGFβ[3]. TGFβ led to robust EMT in all cancer cell lines fully converting epithelial into mesenchymal states within 7 days[2,3]. For each cell line, we distinguished epithelial and TGFβ-induced mesenchymal states using the SVM, built the STV, and calculated the DPD for control and drug-treated conditions (Extended Data Figs. 18 and 19). To compare phenotype changes across five different cell lines and two different omics dataspaces, we normalized the DPD to 1 for TGFβ-induced mesenchymal states and to 0 for control epithelial states. Using the normalized DPD values, we were able to quantitate and compare the extent of EMT for each ligand as compared to TGFβ-induced EMT, and the potency of each drug to suppress EMT in different cell lines and in individual cells (Extended Data Fig. 19 and Table S11). Below, we illustrate the conclusions drawn from this analysis.

Compared to TGFβ, treatment with EGF or TNFα results in a weaker induction or even the absence of EMT, as was also observed by Cook and Vanderhyden based on the expression of EMT hallmark genes[2]. Both EGF and TNFα induce broad single-cell DPD distributions with the medians closer to the epithelial state in all cells except MCF7, in contrast to a narrow DPD peak induced by TGFβ (Extended Data Fig. 20). Only Py2T cells stimulated with TGFβ show broad single-cell distributions detected by mass cytometry (Extended Data Fig. 21). TGFβ receptor (TGFβR) inhibitors efficiently suppress EMT across all cell lines and ligands, suggesting a key role of the TGFβ/SMAD pathway in EMT induction, which is in line with the findings of the original publications[2,3] (Extended Data Fig. 19). After TGFβR inhibitors, the next most potent EMT inhibitor is Necrostatin-5, a drug that inhibits Receptor-Interacting Serine/Threonine-Protein Kinase 1 (RIPK1), which is a regulator of TNFα-mediated apoptosis, necroptosis and inflammatory pathways. Necrostatin-5 fully suppresses the partial EMT induced by EGF or TNFα and partially reduces TGFβ-induced EMT (Extended Data Fig. 19). Single-cell DPD data in cell lines treated with TGFβ plus Necrostatin-5 show broad distributions with the median centered between epithelial and mesenchymal states (Extended Data Figs. 20B, 20D, 20F, 20H). This spread of the DPD values suggests that cells can adopt a continuum of states between the epithelial and mesenchymal states, which can be described as hybrid epithelial/mesenchymal phenotypes. A recent study distinguished six distinct populations of tumor cells in a mouse model of skin squamous cell carcinoma, which were interpreted as successive hybrid epithelial/mesenchymal states[4]. The broad single-cell DPD distributions that we found would emerge if discrete local minima of Waddington's landscape between fully epithelial and mesenchymal states are shallow, resulting in a stochastic blurring of discrete states and allowing cells to cross over between valleys as they maneuver through the landscape (see below "The impact of noise on cell fate decisions" for details). This bridges two concepts of continuum and discrete hybrid epithelial/mesenchymal states.

As RIPK1 also can activate the NF-kB pathway[5], one may expect that IKK inhibitors have similar effects as Necrostatin-5. Indeed, IKK inhibition substantially suppresses EMT in TGFβ stimulated Py2T and A549 cells but not in other cell lines (Extended Data Fig. 19). Among all cell lines, TGFβ-induced EMT in DU145 cells is most resistant to suppression. In these cells, only TGFβR and RIPK1 inhibition leads to strong suppression of TGFβ-induced EMT, while its slight suppression is observed following the inhibition of RTKs, IKK, Aurora-A kinases (Extended Data Fig. 19). These observations are in line with the findings of both original publications[2,3].

In general, kinase inhibitor effects on EMT are cell line specific, as we observed for responses to EGF and TNFα. Simultaneous inhibition of multiple RTKs (VGFR/PDGFR/FGFR) strongly suppresses EMT in Py2T and A549 cells induced by TGFβ, but weakly inhibits EMT in DU145, MCF7 and OVCA420 cell lines (Extended Data Fig. 19). This suggests that positive autocrine or paracrine signaling loops acting via different RTKs are important drivers of EMT in some cell lines, but not in others. Inhibition of individual RTKs, such as EGFR, PDGFR, FGFR or VEGFR, does not efficiently suppress EMT in any cell line except Py2T. MEK inhibition suppresses EMT in TGFβ-treated A549, MCF7 and Py2T cells, but not in DU145 and OVCA420 cells, also supporting the conclusions of the original publications[2,3]. Inhibition of the p38, PI3K/mTOR, JAK, JNK and ROCK kinases do not markedly suppress EMT in any cell line except Py2T[2,3].

The DPD values identify differential, activating or inhibiting, drug effects on EMT depending on the cell and stimulation context, which remained undetected in the original analyses[2,3]. Whereas A549 cells do not undergo EMT following EGF stimulation, the concomitant inhibition of the PI3K/mTOR pathway induces EMT in these cells, as well as in MCF7, DU145 and OVCA420 cells, and in TNFα-treated DU145 and OVCA420 cells. By contrast, PI3K/mTOR inhibition moderately suppresses TGFβ-induced EMT in these cells (Extended Data Fig. 19). MEK inhibitors partially suppress EMT in TGFβ-stimulated A549 cells, while promoting EMT in response to TNFα. Likewise, PKC inhibition suppresses TGFβ-induced EMT to different extents, strongly in Py2T, weakly in A549, MCF7 and OVCA420, and very weakly in DU145. However, in EGF-treated A549, DU145 and MCF7, and in TNFα-treated DU145 and OVCA420, PKC inhibition promotes EMT. Inhibition of Aurora-A kinase moderately suppresses EMT in TGFβ-treated A549 and DU145, but promotes EMT in MCF7 cells treated with either EGF or TNFα. Inhibition of GSK3β weakly suppresses EMT in TGFβ-treated A549 and OVCA420, and promotes EMT in MCF7 cells treated with either EGF or TNFα.

In summary, our results correspond well to the original phenomenological observations and conclusions drawn from the experiments in these papers[2,3], showing that cSTAR correctly captures the relationships between phenotypical and underlying molecular states. Moreover, cSTAR adds new insights to the original conclusions. For instance, the DPD analysis demonstrates that the observed partial EMT states comprise a continuum of intermediate states between fully epithelial and fully mesenchymal states.

Inference of signaling networks driving epithelial-mesenchymal transition (EMT)

To underpin different cell states with mechanistic interpretations, which was previously not possible, we applied BMRA to reconstruct the kinase signaling networks underlying these phenotypes under each treatment condition. To enable kinase network reconstructions, the changes in the kinase activities must be measured at a time where the phosphorylation dynamics can be causally connected to the subsequent changes in the cell phenotype. In the CYTOF experiments[3], phosphorylation responses were measured at 7 days after inhibitor perturbations despite that they may occur within minutes following these inhibitor treatments. This is a too long gap following causative changes in the signaling dynamics. For instance, phospho-SMAD2/3, which are *bona fide* downstream effectors of TGFβR and well-known promoters of EMT, were downregulated in TGFβ-treated cells relative to control, whereas the inhibition of TGFβR resulted in upregulation of phospho-SMAD2/3 up to their control levels. These surprising observations can be explained by a faster degradation of active forms of SMAD transcription factors[6,7], yet we cannot determine the activities of TGFβR and other kinases, which might affect EMT, at the earlier time points.

Although the single cell RNAseq datasets were also measured at 7 days after perturbations, the ensuing transcriptional responses develop at later times than the more immediate phosphorylation responses. However, the causal kinase network inference by BMRA requires the data on kinase activity responses, rather than transcriptomics responses to kinase inhibitors. Therefore, we first extrapolated the changes in the kinase activities based on the expression of downstream transcriptional targets of these kinases and the signaling pathways responsible for regulating the transcriptional targets. These extrapolated values allowed us to calculate the global response coefficients for each pathway inhibited by a specific kinase inhibitor in the dataset. These coefficients, together with the DPD changes, served as an input for BMRA network inference (Supplementary Experimental Procedures).

Kinase networks reconstructed by BMRA (Table S12) demonstrate that network wiring changes not only between different cell line but also in response to ligands that activate different RTKs[8]. However, due to the overlap between transcriptional downstream targets of different kinases, the calculated global response coefficients are not completely independent. As a result, the inferred matrices of the connection coefficients are denser than the matrices inferred for directly measured protein responses (cf., Table S12 with Tables S5 and S10). Yet, these networks help us understand the mechanistic regulation of the EMT induction for each kinase, showing whether it directly changes the DPD, driving EMT through its own transcriptional response, or engages other network modules by activating or inhibiting their kinase components. The immediate effects of each signaling node on DPD can be seen by the last row of the connection (r) matrix presenting the signaling pathway connections to the DPD module. The network-mediated effects of signaling nodes on the DPD are given (as the first, linear approximation) by the last row of the inverse connection matrix (r^{-1}). Both matrices are presented in Table S12.

Because TGFβR and RIPK1 inhibitors were the most potent drugs to suppress EMT, we investigated the molecular mechanisms how the TGFβR and RIPK1 pathways drive EMT. The reconstructed networks demonstrate that the TGFβR pathway directly changes the DPD in TGFβ-stimulated A549 and OVCA420 cells. In DU145 cells and MCF7 cells, the TGFβR pathway drives EMT indirectly by activating the PI3K/mTOR and JNK modules and the MEK/ERK and RIPK1 modules, respectively (Table S12). Of note, even in A549 and OVCA420 cells, mutual cooperation of the TGFβR module with other kinase modules is important. For instance, in these cell lines GSK3β, in addition to its direct positive effect on the DPD and hence pro-EMT effects, activates the TGFβR pathway, explaining why the GSK3β inhibition substantially but not fully suppresses both TGFβR/SMAD signaling and EMT[2]. The inferred network connections show that like the TGFβR pathway, the RIPK1 module directly induces EMT in TGFβ-stimulated A549 and OVCA420 cells, and this induction cannot be explained by RIPK1 crosstalk with the NFkB/IKK, TNFα or ERK pathways (Table S12). Moreover, while in A549 and OVCA420 cells a direct, strong connection from RIPK1 to the DPD module is observed only upon TGFβ stimulation, in MCF7 cells the RIPK1 module is a key direct driver of EMT regardless of the stimulating ligand. This explains why the partial EMT induced by EGF and TNFα in MCF7 cells is completely abrogated by RIPK1 inhibition[2].

Signaling by EGFR, VGFR, PDGFR and FGFR affects EMT only via crosstalk with the TGFβR and the activation of downstream kinase pathways, because these RTKs lack direct connections to the DPD module. For instance, in EGF-stimulated DU145 cells, a partial EMT is mediated by EGFR-driven activation of VGFR, PDGFR and FGFR, which activate TGFβR. This cross-activation can be explained by autocrine or paracrine feedback loops, or via receptor heterodimerization in the cell membrane. A partial EMT induction in EGF-stimulated DU145 cells is counteracted by the PKC pathway, which is activated by PI3K/mTOR signaling. This explains why the inhibition of either of these two pathways can promote EMT.

The partial EMT in EGF-stimulated MCF7 cells is mainly driven by RIPK1 signaling, activated via the MEK/ERK pathway. Here TGFβR and other RTKs contribute to MEK/ERK activation via mutual positive connections. Thus, the inhibition of RTKs, MEK/ERK or RIPK1 results in a similar, almost complete suppression of EGF-driven partial EMT, as observed in the original publication[2]. At the same time, the PI3K/mTOR, PKC and GSK3β modules negatively influence the DPD, thereby directly suppressing EMT, whereas Aurora-A inhibits the main EMT driver, RIPK1. Not surprisingly, the inhibition of these signaling pathways promotes EMT in EGF-stimulated MCF7 cells, as the DPD changes caused by these inhibitors demonstrate.

In TNFα-stimulated MCF7 cells, TGFβR and RTKs contribute to a partial EMT via the MAPK modules (MEK/ERK and JNK) that activate RIPK1 (Table S12). Here, the EMT induction is limited by GSK3β, which directly negatively affects the DPD, and Aurora-A, which inhibits RIPK1. Accordingly, the inhibition of GSK3β and Aurora-A promotes EMT in TNFα-stimulated MCF7 cells.

We conclude that a partial EMT is generated by the intricate balance between EMT promoting and EMT suppressing signals. Depending on the cell line and ligand, well-studied signaling pathways, such as the MEK/ERK, PI3K/mTOR, PKC, GSK3β, Aurora-A and other pathways, can promote or suppress EMT. Reconstructed network connections, including connections to the DPD module, underpin the roles of these pathways in the EMT induction or suppression with mechanistic explanations. While the role of the MEK/ERK, PKC, GSK3β and Aurora-A pathways is impressively cell-specific, PI3K/mTOR signaling tends to promote TGFβ-induced EMT and suppress EMT in response to EGF or TNFα stimulation in all cell lines. However, the magnitude of these effects depends on the cell line (Table S12).

In summary, twelve BMRA-reconstructed networks indicate how differential network topologies and connection strengths cause cell type and stimulation-specific responses. Yet, these networks, which reveal the mechanistic meaning of the DPD calculations, can only be considered as hypotheses (Table S12). A detailed experimental validation is part of future work. The mechanistic, predictive simulations enabled by our reconstructed networks provide valuable suggestions for designing the most informative experiments to disentangle the relationships between these multiple EMT states.

Supplementary Experimental Procedures

Testing cSTAR predictions using unbiased mass spectrometry (MS) acquired datasets

The data were analyzed using the python scripts provided at https://github.com/OleksiiR/cSTAR_Nature. 3 biological replicates were done in 2 technical replicates each. The STV ranking, kinase enrichment analysis results and calculated DPD values for perturbations are presented in Tables S7 and S8.

To determine synergy between Trametinib and Gefitinib we have used the Bliss independence criterion. The drug independence means that the relative effect of a drug is independent of the presence of the other drug[9]. The effect of a combination of two independent drugs is additive[9], and the Bliss synergy score equals zero[10].

Let DPD_d be the DPD value for the differentiation state of NGF-stimulated TrkA cells, DPD_p be the DPD value for proliferation state for BDNF-stimulated TrkB cells, and DPD_{drug} be the DPD value for BDNF-stimulated TrkB cells treated with a specific drug or a drug combination. Then, the induction of differentiation by a drug for TrkB cells is described by the effect value, Y, as follows,

$$Y = \frac{DPD_{drug} - DPD_d}{DPD_p - DPD_d} \tag{S1}$$

.

Y is 1 if drug has no effect, Y is 0 for a complete switch to a differentiation state, and Y is between 0 and 1 for a partial increase in the number of differentiated cells. According to the Bliss criterion, if the drug 1 produces effect Y_1 and drug 2 produces effect Y_2, and drugs 1 and 2 act independently, i.e. there is neither synergy no antagonism between them, the effect of their combination will be the following,

$$Y_{comb}^{ind} = Y_1 \cdot Y_2 \qquad \text{(S2)}$$

.

In our experiments, we combined Trametinib and Gefitinib at twice lower doses than the doses of these drugs given separately. Then, if Trametinib and Gefitinib would be merely additive, the effect Y_{Tr+Gef} of a combination would be the following,

$$Y_{Tr+Gef}^{ind} = \sqrt{Y_{Tr}} \cdot \sqrt{Y_{Gef}} \qquad \text{(S3)}$$

.

Here Y_{Tr} and Y_{Gef} are the phenotypic effects of Trametinib and Gefitinib given in twice larger doses on their own. The Bliss synergy score[10] was defined as follows,

$$Score = Y_{Tr+Gef}^{ind} - Y_{Tr+Gef}^{obs} = \sqrt{Y_{Tr}} \cdot \sqrt{Y_{Gef}} - Y_{Tr+Gef}^{obs} \qquad \text{(S4)}$$

.

Here Y_{Tr+Gef}^{obs} is the observed effect of a combination. The synergy scores calculated for each replicate are presented in Table S7 and demonstrated statistically significant synergy between Gefitinib and Trametinib for both datasets of TrkB samples that were analyzed on separate MS machines. The average Bliss synergy score was equal to 20% ± 2%).

Calculating global response coefficients from single cell RNAseq data

The BMRA algorithm takes as an input the global response coefficients, R_{ij}, calculated from experimental data as follows[11]

$$R_{ij} = \frac{x_j^i - x_0^i}{x_{av}^i} \qquad \text{(S5).}$$

Here x_0^i is the activity of i-th module before perturbation, x_j^i is the activity of i-th module after j-th perturbation, and x_{av}^i is a value of the i-th module activity within an interval $[x_0^i, x_j^i]$ (cf. Eq. 22 in Materials and Methods). Because R_{ij} are defined as dimensionless ratios, the activity of a kinase module can be estimated using any value that is proportional to the module activity. Because only transcriptional responses to kinase inhibitors have been measured in the original publication[2], we have to extrapolate the changes in the kinase activities using measured scRNAseq data. For this estimation, we need to determine immediate downstream transcriptional targets of these kinases. Below we present an algorithm of finding kinase activity descriptors and calculating the global response coefficients based on the available transcriptomics data. It consists of the following 6 steps.

First, single cell RNAseq data were normalized using the standard LogNormalize method from the Seurat R package[12]. Briefly, the LogNormalize command (*i*) normalizes the expression of a particular mRNA by the total mRNA expression level in this cell, (*ii*) multiplies this by a scale factor 10,000, (*iii*) adds 1, and (*iv*) log-transforms the result. Next, we obtain the population average values for each mRNA and each condition and using log-transforms obtain the log-fold changes with respect to no-treatment control.

Second, for each cell line and each ligand we determined differentially expressed genes (DEGs) for pairs of the control state and a state resulting from a specific kinase inhibitor perturbation using the non-parametric Wilcoxon rank sum test from the Seurat R package and cut-off criteria of a FDR corrected p value < 0.05. Then, we determined overrepresented transcription factors (TFs) using the TF enrichment analysis (TFEA) method and ChEA & ENCODE consensus database[13].

Third, using the interaction proteomics databases BioGRID, IntAct,MINT, ppid[14,15] we found the proteins known to physically interact with these TFs and termed these proteins interactors. Then, we enriched each list of overrepresented transcription factors by adding corresponding interactors.

Fourth, using the kinome database KEA[16], for each perturbed kinase we determined its substrates that are present in the corresponding list of overrepresented TFs and their interactors[1]. For each kinase, we called this list of kinase targets a preliminary activity descriptor list for the given kinase.

Fifth, for each cell line (A549, DU145, MCF7 and OVCA420), a ligand (TGFβ, EGF and TNFα) and a drug perturbation, we calculated the global response coefficients for each gene in the preliminary activity descriptor of each perturbed kinase using the following expression,

$$R_{ij}^{k} = \frac{y_j^k - y_0^k}{\left(\left|y_j^k\right|, \left|y_0^k\right|\right)} \qquad \text{(S6).}$$

Here R_{ij}^{k} is a global response coefficient for k-th gene in the preliminary activity descriptor of i-th kinase upon j-th perturbation, y_0^k is the expression level of k-th gene before perturbation, y_j^k is the expression level of this gene after j-th perturbation. This expression is only a first-order approximation, but it gives robust estimates for R_{ij} when y_j^i or y_0^i are small values.

Next, for every cell line and each perturbed kinase (i), we analyzed the signs of R_{il}^{k} for each k-th gene across different ligands for a perturbation (l) that inhibited this i-th kinase (one or two inhibitors were used for the suppression of the same kinase, therefore the index l has one or two values here). A negative sign shows the reduced k-th gene expression by the i-th kinase inhibition with its inhibitor l, and a positive sign shows that the k-th gene expression was facilitated by the inhibitor. If R_{il}^{k} does not change its sign for cell stimulation with different ligands, this gene remains in the activity descriptor list of i-th kinase, but if R_{il}^{k} changes the sign, this gene is excluded from the list. The resulting gene list is our final activity descriptor list of i-th kinase. If the global response coefficient R_{il}^{k} is positive for a given kinase i, the inhibitor of this kinase (l), and a specific k-th gene in this list, this means that the expression of the k-th gene is activated upon the inhibition of this kinase, i.e. its expression level negatively correlates with this kinase activity. Therefore, we change the sign of R_{il}^{k} for this gene to ensure the positive correlation with the kinase activity.

Sixth and finally, we calculated the average value of transcriptomic responses over all genes from the kinase activity descriptor list for each kinase i and each inhibitor j, as follows,

$$R_{ij} = \frac{1}{N} \cdot \sum_{k=1}^{N} \delta_i^k R_{ij}^{k} \qquad \text{(S7).}$$

Here, R_{ij} are global response coefficient of the activity of i-th kinase to j-th perturbation estimated using the transcriptional response; $\delta_i^k = 1$ if k-th gene is suppressed when kinase i is inhibited, and $\delta_i^k = -1$ if k-th gene is activated when kinase i is inhibited. Although upon an arbitrary perturbation j the terms, $\delta_i^k R_{ij}^k$, can be negative or positive, upon a direct inhibition of i-th kinase by its inhibitor (l), the estimated global response coefficient R_{il} of the kinase i activity is negative (we assume that the inhibitors used in the original publication[2] do not paradoxically activate their primary kinase targets[1]).

Using this algorithm, we calculated the global response coefficients for all kinases perturbed by inhibitors in the original publication[2]. The global responses of the DPD module to each inhibitor perturbation were calculated as previously described. The python code generating global response matrices for the scRNA data is available at https://github.com/OleksiiR/cSTAR_Nature. Finally, we applied BMRA to the approximated global responses to find cell- and ligand-specific connections in the regulatory network, which includes both kinases and the DPD module.

Predicting systemic responses to perturbations using linear approximation

To derive mechanistic systems of nonlinear ODEs that can simulate the dynamics of these 12 kinase networks and the movement of the cell states through a Waddington landscape would require measurements of the kinase activities at earlier timepoints after perturbation than 7 days. This gap is too long to infer a full nonlinear dynamic system that causally connects the molecular perturbation to the phenotypical outcomes. Also, the estimates of the global kinase responses to inhibitors from cRNAseq responses[2] involve several approximative steps, and therefore the inference of the connections strengths and the normalized Jacobian elements, which are needed to build mechanistic models, is qualitative rather than quantitative. However, the MRA framework allows making qualitative predictions of systems responses to perturbations using a linear approximation[17,18]. Given the vector r_I describes the local perturbations to signaling nodes (such as a receptor or a cytoplasmic kinase to an inhibitor), and r is the matrix of connection coefficients (aka the local response matrix), then the vector of the global responses R_I can be obtained using the following expression[18]

$$R_I = -r^{-1} \cdot r_I \qquad \text{(S8).}$$

Thus, for the first order estimate of how specific perturbation propagates through the network and changes cellular phenotype, we must calculate the inverse connection matrix r. Table S12 contains $-r^{-1}$ matrices for all 12 conditions, in which single cell RNAseq data were acquired. Each j-th column $(-(r^{-1}))_j$ of this matrix contains the vector of system responses of all signaling nodes to a perturbation of node j. For example, the first column in matrices inferred for TGFβ-stimulated cells quantifies the response of each node to the activation of TGFβR (approximated as the percentage change in the node brought about by 1% activation of TGFβR). To obtain system responses to the inhibition of a signaling node, the elements of this matrix must be multiplied by -1 (accordingly, the first column of $(r^{-1})_j$ quantifies the systems response to the inhibition of TGFβR). Likewise, to estimate the responses of all nodes to MEK inhibition, the MEK/ERK column of $-r^{-1}$ must be multiplied by -1. The last row of each of twelve $-r^{-1}$ matrices contains the DPD responses, thereby giving the phenotypic responses to a stimulation of the corresponding signaling module. For example, we can see that regardless of a specific cell line and a ligand, the DPD response to small changes in the TGFβR activity is positive under all conditions, meaning that TGFβR signaling node positively contributes to EMT, even if the direct connection from TGFβR to DPD module is negligible.

In summary, even when a precise nonlinear model that would quantitatively predict signaling and phenotypic effects of every perturbation and every combination of perturbations cannot be built, MRA allows qualitative estimations of these effects (Eq. S8) as a linear, first-order approximation.

The impact of noise on cell fate decisions

In the main text we have not considered the impact of noise on our nonlinear dynamical models of SH-SY5Y-TrkA/B and SKMEL-133 cells, because these models are developed based on the RPPA data, which represent the population averages, rather than single cell values. Yet, cSTAR allows to consider stochastic features intrinsic to cell fate decision processes. Here, we show how cSTAR can be applied to single cell signaling data and account for the impact of noise on cell fate outcomes.

cSTAR conceptually divides cell signaling patterns into two unequal parts: a mechanistically described core network and the cell-wide signaling network, described using a distinct DPD module. A cell maneuvering through Waddington's landscape is described by the changes in the DPD, which is driven by core network signaling (signaling driving force) and a phenomenological quadratic potential that creates the gradient force field (restoring force). To analyze the impact of noise on cell-wide signaling and cell fate decision, we introduce probabilistic terms in an equation describing the evolution of DPD values (Eqs. 28 and 29), making it the Langevin equation for describing single-cell behavior or the Fokker-Planck equation for describing the population dynamics. Under general assumptions about noise in cell signaling, the stochastic differential equation governing DPD (S) dynamics reads (cf. Eqns. 29 and 30),

$$dS = -\frac{dW(S,\vec{x},t)}{dS}dt + \sigma \cdot dB \qquad \text{(S9).}$$

Where B denotes a Wiener process, $\vec{x} = (x_i)$ is a vector of the activities of core network components, W is a Waddington's landscape potential defined in Eqn. 30, and σ is the noise quantity[19].

Since for every fixed time point, the $dW(S,\vec{x},t)/dS$ term consists of linear with respect to S terms (see Eqns. 25-30), the Freidlin-Wentzell theorem[20] readily gives an analytical estimate for the probability of a stochastic escape from a local W minimum for each time point.

In addition, Eq. S9 allows us to estimate the width of the DPD distribution in the vicinity of a local minimum of W. Because $W \sim \frac{1}{2}\alpha(S - S_0)^2 + o((S - S_0)^2)$, where S_0 is the local minimum of W, from Eq. S9 it follows that the width (ΔS_0) of the DPD distribution in the vicinity of S_0 is

$$\Delta S_0 \approx \sigma/\sqrt{2\alpha} \qquad \text{(S10).}$$

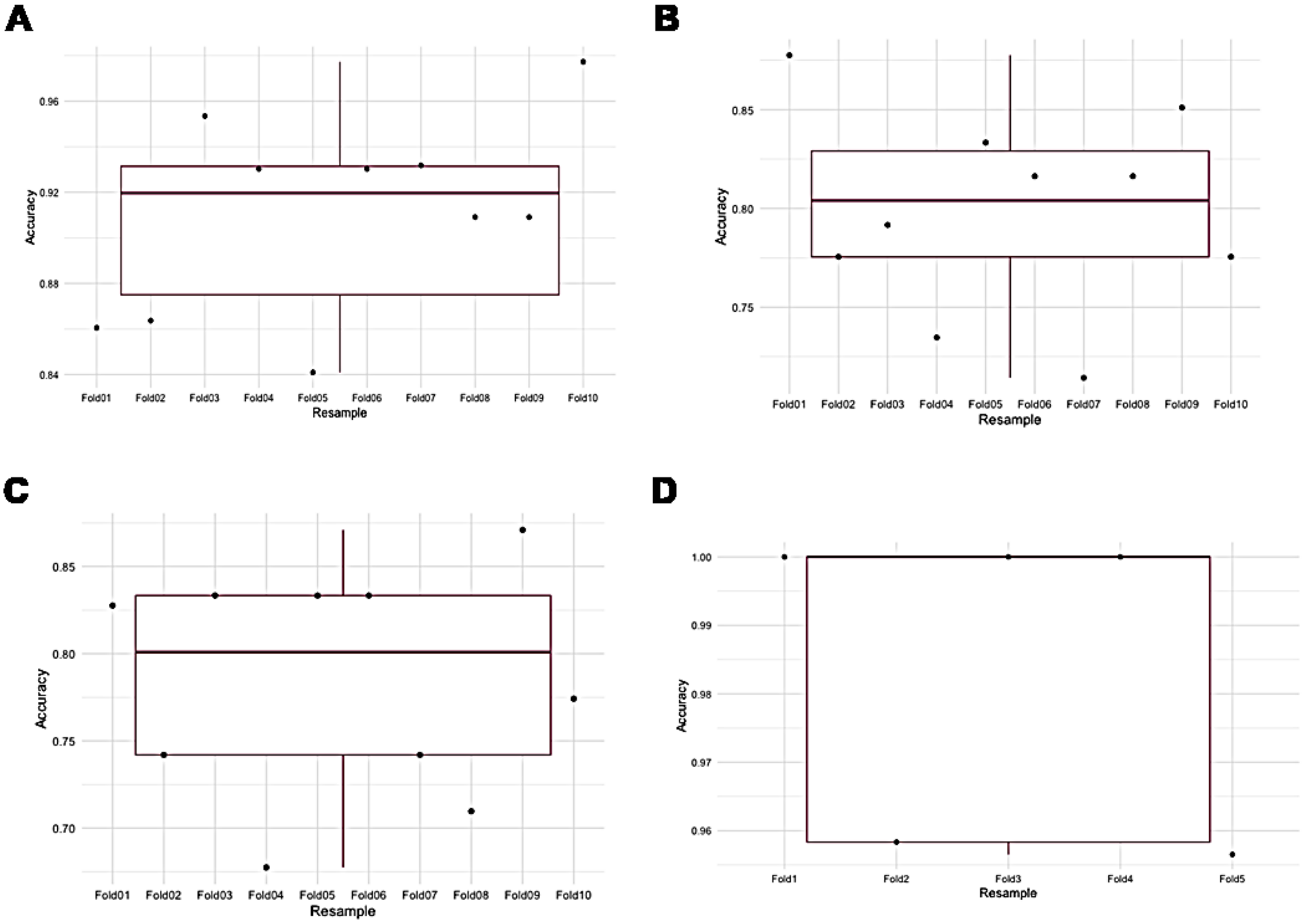
Consequently, if the local minimum of W is shallow, i.e. α is small, the distribution of DPD is wide, explaining how a discrete number of stable steady cell states (that are local W minima) results in a wide spectrum of single cell states. If a core network consists of kinases, as in our work, the noise in the chemical reactions comprising the network can be neglected because of large protein numbers, although there is cell-to-cell variability in the protein abundances considered as the initial conditions. Consequently, only the DPD dynamics is described by a stochastic differential equation (or a partial differential equation, if the population cell behavior is studied), while the rest of the model consists of deterministic ODEs.

Robustness of the SVM classification and the STV ranking of analytes to noise

In this section we explored robustness of (i) the classification of cellular states by SVM with linear kernel and (ii) the STV ranking of analytes to noise. We used our own data (RPPA proteomics for SH-SY5Y-TrkA/B cells) and publicly available datasets that include both RPPA data (SKMEL-133 cells[21]) and single cell RNAseq transcriptomics[2]. The accuracy of the SVM classification and the selection of top-ranking analytes was studied using a k-fold cross validation (with k=5 or k=10 depending on the number of samples)[22].

We first added random noise to datasets labelled SKMEL-133 (16 samples, 235 analytes) and TrkA/B (24 samples and 117 analytes) to investigate how various noise levels affect the accuracy of the method. The magnitude of noise was a certain percentage (n_r) of the measured value for each analyte, so that signal to noise ratio was equal to $1/n_r$. We considered both a full dataset and a reduced dataset obtained by removing core network components (16 samples and 179 analytes and 24 samples and 70 analytes, respectively). In these four cases, the value of each measured analyte x was modified by the adding a uniformly distributed random value within a range between $-n_r x$ and $n_r x$ where n_r is a positive number between 0 and 1. To test SVM robustness on the RPPA datasets we used 5-times repeated 5-fold cross validation. The results show that a SVM with a linear kernel is highly robust to noise, although results become slightly less accurate on reduced datasets (Table S14). A main effect of the added noise is increasing variance in the datapoint distances to the separation plane, but since the classification is binary, its accuracy is robust with respect to random movement of data points in the clouds corresponding to different states. Because the number of measured samples was typically low (<10 for each condition) whereas the number of analytes was high (> 100) for population averaged RPPA data (a condition aka "High Dimension Low Sample Size (HDLSS)"[23]), there is a risk of model over-fitting that can explain very high SVM robustness to noise. Single cell RNAseq data are free from the over-fitting problem due to the greater number of datapoints. It allowed us to designate a subset of the data for additional validation to confirm the reliability of our modelling approach as presented next.

The second set of tests was carried out on the single cell RNAseq data[2]. The large number of samples for each cell line allowed us to subdivide the entire dataset into 75% and 25% randomly selected parts. We applied cross validation procedure for a 75% part, while a 25% part was used for independent validation using a single cross validation run. Thus, we obtained two different estimates of classification accuracy. The noise was added to every analyte x as a uniformly distributed random vector with values within $[(x) - (x), (x) - (x)]$ interval. The datasets are abbreviated as A549 (12,924 genes, 494 datapoints for no treatment, epithelial state and 88 datapoints for TGF-stimulation, mesenchymal state), DU145 (13,016 genes, 344 datapoints for epithelial state and 304 datapoints for mesenchymal state), MCF7 (13,370 genes, 274 datapoints for epithelial state and 132 datapoints for mesenchymal state) and OVCA420 (13,057 genes, 145 datapoints for epithelial state and 10 points for mesenchymal state). The results are presented in Table S15 and Supplementary Fig. 2. Although the OVCA420 dataset was ill-balanced, (93.5% of datapoints for epithelial state and 6.5% of datapoints for mesenchymal state), due to the large number of analytes and a good separation of classes, the SVM with linear kernel showed its robustness to noise (Table S15).



Supplementary Figure 2. Accuracy of classification of scRNAseq data with added artificial noise tested using 10-fold cross validation. Data are taken from Ref.[62]. Dots correspond to the classification accuracy calculated for each fold cross validation. The box plot summarizes the individual experiments by showing the median (bold horizontal line) and the distance between the third (upper box bound, Q3) and the first (lower box bound, Q1) quartiles. The inter-quartile range is defined as IQR=Q3-Q1. The ends of the whiskers correspond to Q1 - 1.5*IQR and Q3 + 1.5*IQR. (A) – A549 cells, (B) – DU145 cells, (C) – MCF7 cells, (D) – OVCA420 cells.

Summarizing, when applied to the original data, SVMs completely separated population averaged RPPA data achieving a higher than 90% accuracy of cell state separation for each cell line. Added noise moderately decreased the SVM accuracy of classification of single cell RNAseq data but, surprisingly, only very high level of noise (noise to signal ratio greater than 0.5) resulted in separation errors for population averaged RPPA data.

After showing that the cell state separation is robust, we asked whether the STV ranking that identifies core molecule species is robust to noise. To answer this question, we analysed an extreme case of very noisy data with the added 50% noise (SNR = 2). Within the same k-cross validation procedure, we built the STVs for the original RPPA and single cell RNAseq transcriptomics data and the computationally obtained very noisy data (SNR = 2, Table S15). We observed that the highly ranked STV components that defined core network modules were largely conserved between the original and noisy RPPA data. As expected, the ranking of analytes which are outside of the core network changed more substantially, but these changes did not alter the cell state classification determined by the DPD. For single cell RNAseq data we found that many analytes had very comparable contributions to STV, while the exact rank of each gene is determined less accurately than for RPPA data. However, the top 100 components essentially overlap between the original and noisy datasets, and the top 1000 components were determined even more robustly.

In summary, the results suggest that cSTAR is robust to noise originating from different types of omics data.

References

1 Lachmann, A. & Ma'ayan, A. KEA: kinase enrichment analysis. *Bioinformatics* **25**, 684-686, doi:10.1093/bioinformatics/btp026 (2009).

2 Cook, D. P. & Vanderhyden, B. C. Context specificity of the EMT transcriptional response. *Nature Communications* **11**, 2142, doi:10.1038/s41467-020-16066-2 (2020).

3 Chen, W. S. *et al.* Uncovering axes of variation among single-cell cancer specimens. *Nature Methods* **17**, 302-310, doi:10.1038/s41592-019-0689-z (2020).

4 Pastushenko, I. *et al.* Identification of the tumour transition states occurring during EMT. *Nature* **556**, 463-468, doi:10.1038/s41586-018-0040-3 (2018).

5 Ang, R. L., Chan, M. & Ting, A. T. Ripoptocide – A Spark for Inflammation. *Frontiers in Cell and Developmental Biology* **7**, 163 (2019).

6 Xu, L. Regulation of Smad activities. *Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression* **1759**, 503-513, doi:https://doi.org/10.1016/j.bbaexp.2006.11.001 (2006).

7 Izzi, L. & Attisano, L. Regulation of the TGFβ signalling pathway by ubiquitin-mediated degradation. *Oncogene* **23**, 2071-2078, doi:10.1038/sj.onc.1207412 (2004).

8 Santos, S. D., Verveer, P. J. & Bastiaens, P. I. Growth factor-induced MAPK network topology shapes Erk response determining PC-12 cell fate. *Nat Cell Biol* **9**, 324-330 (2007).

9 Greco, W. R., Bravo, G. & Parsons, J. C. The search for synergy: a critical review from a response surface perspective. *Pharmacological Reviews* **47**, 331 (1995).

10 Holbeck, S. L. *et al.* The National Cancer Institute ALMANAC: A Comprehensive Screening Resource for the Detection of Anticancer Drug Pairs with Enhanced Therapeutic Activity. *Cancer Research* **77**, 3564, doi:10.1158/0008-5472.CAN-17-0489 (2017).

11 Kholodenko, B. N. *et al.* Untangling the wires: A strategy to trace functional interactions in signaling and gene networks. *Proceedings of the National Academy of Sciences* **99**, 12841, doi:10.1073/pnas.192442699 (2002).

12 Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e3529, doi:https://doi.org/10.1016/j.cell.2021.04.048 (2021).

13 Lachmann, A. *et al.* ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics* **26**, 2438-2444, doi:10.1093/bioinformatics/btq466 (2010).

14 Stelzl, U. *et al.* A Human Protein-Protein Interaction Network: A Resource for Annotating the Proteome. *Cell* **122**, 957-968, doi:10.1016/j.cell.2005.08.029 (2005).

15 Berger, S. I., Posner, J. M. & Ma'ayan, A. Genes2Networks: connecting lists of gene symbols using mammalian protein interactions databases. *BMC Bioinformatics* **8**, 372, doi:10.1186/1471-2105-8-372 (2007).

16 Kuleshov, M. V. *et al.* KEA3: improved kinase enrichment analysis via data integration. *Nucleic Acids Research* **49**, W304-W316, doi:10.1093/nar/gkab359 (2021).

17 Bruggeman, F. J., Westerhoff, H. V., Hoek, J. B. & Kholodenko, B. N. Modular Response Analysis of Cellular Regulatory Networks. *Journal of Theoretical Biology* **218**, 507-520, doi:https://doi.org/10.1006/jtbi.2002.3096 (2002).

18 Kholodenko, B. N., Rauch, N., Kolch, W. & Rukhlenko, O. S. A systematic analysis of signaling reactivation and drug resistance. *Cell Reports* **35**, 109157, doi:https://doi.org/10.1016/j.celrep.2021.109157 (2021).

19 Brackston, R. D., Lakatos, E. & Stumpf, M. P. H. Transition state characteristics during cell differentiation. *PLoS Comput Biol* **14**, e1006405, doi:10.1371/journal.pcbi.1006405 (2018).

20 Freidlin, M. I., Szucs, J. & Wentzell, A. D. *Random Perturbations of Dynamical Systems*. (Springer New York, 2012).

21 Korkut, A. *et al.* Perturbation biology nominates upstream-downstream drug combinations in RAF inhibitor resistant melanoma cells. *Elife* **4**, doi:10.7554/eLife.04640 (2015).

22 Berrar, D. in *Encyclopedia of Bioinformatics and Computational Biology* (eds Shoba Ranganathan, Michael Gribskov, Kenta Nakai, & Christian Schönbach) 542-545 (Academic Press, 2019).

23 Hall, P., Marron, J. S. & Neeman, A. Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**, 427-444, doi:https://doi.org/10.1111/j.1467-9868.2005.00510.x (2005).