Tensor clustering with algebraic constraints gives interpretable groups of crosstalk mechanisms in breast cancer

Anna Seigal,[1, *] Mariano Beguerisse-Díaz,[2, †] Birgit Schoeberl,[3] Mario Niepel,[4] and Heather A. Harrington[2, ‡]

[1] *Department of Mathematics, University of California, Berkeley, CA 94702, USA*
[2] *Mathematical Institute, University of Oxford, Oxford OX2 6GG, UK*
[3] *Merrimack Pharmaceuticals Cambridge, MA 02139, USA*
[4] *Harvard Medical School, 200 Longwood Avenue, Boston, MA 02115*
(Dated: May 1, 2017)

We introduce a tensor-based clustering method to extract sparse, low-dimensional structure from high-dimensional, multi-indexed datasets. Specifically, this framework is designed to enable detection of clusters of data in the presence of structural requirements which we encode as algebraic constraints in a linear program. We illustrate our method on a collection of experiments measuring the response of genetically diverse breast cancer cell lines to an array of ligands. Each experiment consists of a cell line-ligand combination, and contains time-course measurements of the early-signalling kinases MAPK and AKT at two different ligand dose levels. By imposing appropriate structural constraints and respecting the multi-indexed structure of the data, our clustering analysis can be optimized for biological interpretation and therapeutic understanding. We then perform a systematic, large-scale exploration of mechanistic models of MAPK-AKT crosstalk for each cluster. This analysis allows us to quantify the heterogeneity of breast cancer cell subtypes, and leads to hypotheses about the mechanisms by which cell lines respond to ligands. Our clustering method is general and can be tailored to a variety of applications in science and industry.

Keywords: Algebra; Tensors; Data clustering; Signaling networks; Systems biology; Model selection and parameter inference

I. INTRODUCTION

Muti-dimensional datasets are now prevalent across the sciences; their ubiquity and importance will only continue to grow [1–4]. The analysis of data demands methods that preserve multi-dimensional structures, and that exploit them. We introduce a versatile data clustering framework based on tensors (high dimensional arrays) and algebra to analyze multi-dimensional datasets. One key feature of this method is that it can incorporate general, application-specific constraints on the composition of a cluster, and is guaranteed to find optimal partitions. The flexibility of the method allows it to be used directly on a dataset (i.e., as a standalone clustering tool), or in combination with other clustering methods.

We apply our method on an extensive set of time-course measurements of the activation levels of the mitogen-activated protein kinase (MAPK) and phospho-inositide 3-kinase (PI3K) pathways that are involved in cellular decisions and fates [10–13], and are known to dysfunction in cancer [10–13, 16]. The key signaling proteins and subtype responses in breast cancer cells are known; however, among genetically diverse cell lines the dysfunction varies and is not well understood [1, 15, 16]. Our objective is to find groups of cell lines whose signal transduction networks have similar dynamics. A high similarity suggests that the cell lines share pathway features that can be relevant for the responses to the ligands.

We find these groups by clustering experiments by their temporal profile similarity, subject to 'interpretability' constraints, which are that each cluster must match a subset of cell lines (with similar response) with a subset of growth factors (whose effect on the signaling pathways is perhaps altered by that feature). In other words, the behavior of the cell line/ligand combinations in each must be attributable to the same signaling mechanism. We examine a set of experimental data [1] containing the response of 36 breast cancer cell lines after exposure to 14 ligands (growth factors/signaling molecules). Each experiment measures the temporal phosphorylation response of a cell line to one ligand. Because the dataset is *complete* (i.e., there is a measurement for every combination of times, proteins, cell lines, ligands, and doses), we can represent it as a tensor in five dimensions (Fig. 1A).

Our analysis consists of two stages. In stage 1, we first cluster experimental data encoded in tensors subject to *structural or interpretability constraints* in the form of algebraic inequalities (Fig. 1B,C). Then, in stage 2, we perform a systematic search for nonlinear ordinary differential equation (ODE) models that reproduce the key dynamical features of the time series in each cluster (Fig. 1D). The partition of the experiments into clusters from stage 1 can be visualized by color-coding the grid of experiments (see Figs. 1C and 2) according to their cluster assignment. The constraints force the clusters to pair a subset of the cell lines with a subset of the ligands in such a way that each cluster must be rectangle-shaped on the grid (although possibly disconnected). The assumption behind this constraint is that the observations of the experiments in each cluster must be generated by the same biological mechanism: If two experiments, belong

* seigal@berkeley.edu, joint first author
† beguerisse@maths.ox.ac.uk, joint first author
‡ harrington@maths.ox.ac.uk

to the same cluster, then so should the experiments in the diagonally opposite entries of the grid. We find clusters using a novel notion of tensor similarity, and global optimality is ensured by leveraging results from integer programming. One of the strengths of our approach, is that it can incorporate a pre-existing non-rectangular partition obtained with other methods (e.g., k-means, spectral methods, community detection on graphs). Starting from an initial partition can be computationally advantageous, and our method is able to find the 'closest' optimal rectangular clustering (where the distance between partitions is given by number of reassignments).

Once an optimal partition of the data is obtained, we search for mechanisms that can explain the behavior of the experiments in each cluster. We explore differences between clusters to understand the range of signaling mechanisms across cancer cell lines. To this end, we construct, parametrize, and rank models for each cluster from a pool of 729 candidate models.

II. TENSORS AND ALGEBRA

Data tensor. We represent a multi-indexed dataset (e.g., the complete dataset in Fig. 1A) as a tensor $\mathbf{Z}$ of order h in the real numbers with size $n_1 \times \ldots \times n_h$ (i.e., $\mathbf{Z} \in \mathbb{R}^{n_1 \times \ldots \times n_h}$, where $n_i \in \mathbb{N}$, and $i = 1, \ldots, h$). When the data is complete, every entry of the tensor is filled with a number. A full treatment of tensors is available in [1] and references therein. We introduce here the tensor theory required for our analysis.

Similarity tensors. In a similarity matrix the entry (i, j) records the pairwise similarity of the two items labelled by unidimensional indices i and j. We now introduce the high-dimensional generalization of a similarity matrix, which extends this to multi-indexed data. Suppose we want to compute the similarity of the data indexed by $\mathbf{i} = (i_1, i_2)$ and that indexed by $\mathbf{j} = (j_1, j_2)$:

$$s_{\mathbf{i,j}} = \mathrm{sim}\left(\mathbf{Z}(i_1, i_2, :, \ldots, :), \mathbf{Z}(j_1, j_2, :, \ldots, :)\right), \quad (1)$$

where $i_1, j_1 \in \{1, \ldots, n_1\}$ and $i_2, j_2 \in \{1, \ldots, n_2\}$. The similarity function $\mathrm{sim} : \mathbb{R}^{n_3 \times \ldots \times n_h} \times \mathbb{R}^{n_3 \times \ldots \times n_h} \to \mathbb{R}$ computes the similarity between the data indexed by $\mathbf{i}$ and $\mathbf{j}$ (e.g., correlation or cosine similarity). In general, for data indexed by the first d dimensions, we have the multi-indices $\mathbf{i} = (i_1, \ldots, i_d)$ and $\mathbf{j} = (j_1, \ldots, j_d)$. The dimensions of $\mathbf{Z}$ can be re-ordered as needed. We can construct a *similarity tensor* $\mathbf{S}$ of order $2d$. The shape of $\mathbf{S}$ is determined by the chosen dimensions of the data: $\mathbf{S} \in \mathbb{R}^{n_1 \times \ldots \times n_d \times n_1 \ldots \times n_d}$. The similarity tensor and the similarity matrix are related by *flattening* the tensor as follows. The original data tensor $\mathbf{Z}$ can be flattened (reshaped) into a data matrix $\widetilde{\mathbf{Z}} \in \mathbb{R}^{N_1 \times N_2}$, where $N_1 = \prod_{r=1}^{d} n_r$, and $N_2 = \prod_{r=d+1}^{h} n_r$. Each row of $\widetilde{\mathbf{Z}}$ is a N_2-dimensional vector that corresponds to multi-index $\mathbf{i}$, and the length N_2 is the product of the dimensions of $\mathbf{Z}$ that are *not* included in $\mathbf{i}$.
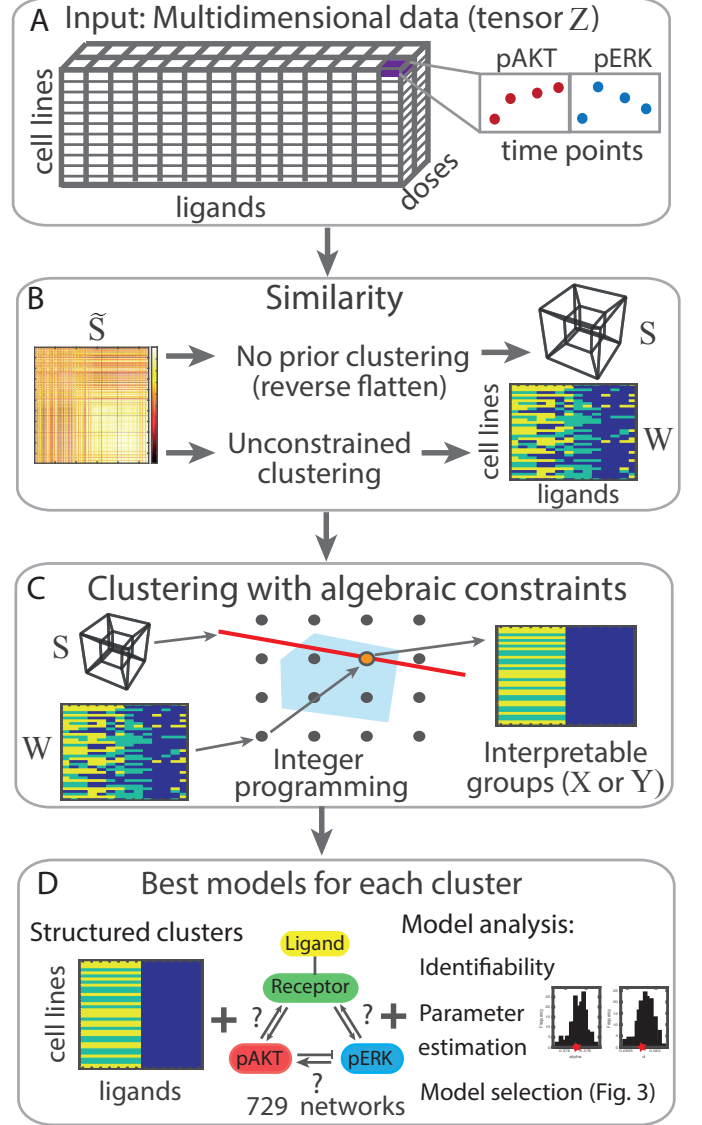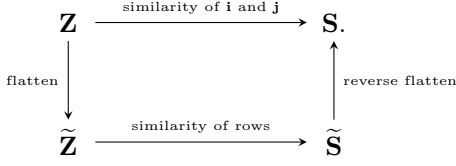

FIG. 1. Schematic of constrained tensor clustering method and model identification. (A) The complete set of experiments can be represented by the multi-indexed tensor $\mathbf{Z}$; see Data section. (B) The similarity scores between experiments (each cell line/ligand combination) can be stored in a similarity matrix $\widetilde{\mathbf{S}}$, that can be used to construct a similarity tensor $\mathbf{S}$, or to find a preliminary clustering of the data $\mathbf{W}$ that may not comply with the constraints. (C) Structured clustering via integer programming. The starting point can either be the similarity tensor $\mathbf{S}$ or the pre-existing clustering $\mathbf{W}$. The possible clusterings are represented by points on the grid. The red line is the value of the objective function (equations 6 and 7). The best integer value (orange point) is found inside the convex feasible region (blue). (D) A large scale search for mechanistic models for each cluster involves parametrising, and ranking the best ODE models for each cluster.

The similarity matrix between the rows of $\widetilde{\mathbf{Z}}$ is $\widetilde{\mathbf{S}} \in \mathbb{R}^{N_1 \times N_1}$, which is obtained by flattening the similarity tensor, $\mathbf{S}$. We summarize this relationship in the following diagram:

$$\mathbf{Z} \xrightarrow{\text{similarity of } \mathbf{i} \text{ and } \mathbf{j}} \mathbf{S}.$$

with vertical maps: flatten $\mathbf{Z} \to \widetilde{\mathbf{Z}}$, reverse flatten $\widetilde{\mathbf{S}} \to \mathbf{S}$, and $\widetilde{\mathbf{Z}} \xrightarrow{\text{similarity of rows}} \widetilde{\mathbf{S}}$.

To compute the similarity tensor $\mathbf{S}$, we can simply flatten the data tensor $\mathbf{Z}$ into $\widetilde{\mathbf{Z}}$, construct a similarity matrix $\widetilde{\mathbf{S}}$, and then reverse flatten it into the desired $\mathbf{S}$. Note that $\mathbf{Z}$ and $\mathbf{S}$ have the same number of entries as $\widetilde{\mathbf{Z}}$ and $\widetilde{\mathbf{S}}$ respectively.

Example. *Let $\mathbf{Z} \in \mathbb{R}^{10 \times 5 \times 3}$ be a tensor of order 3. If $\mathbf{i} = (i_1, i_2)$ is the multi-index, then $d = 2$, $N_1 = 10 \cdot 5 = 50$, and $N_2 = 3$. The (order 4) similarity tensor $\mathbf{S}$ has size $10 \times 5 \times 10 \times 5$. The similarity matrix $\widetilde{\mathbf{S}}$ has size 50×50. The flattened data matrix $\widetilde{\mathbf{Z}}$ has size 50×3.*

Algebraic interpretability condition. When clustering a set of data points we typically seek to find a partition such that the points within a cluster are more similar (or close) to each other than to the rest of the data [17]. In the simplest cases, there are few restrictions on the clusters other than the similarity or distance be reflected in the cluster assignments. In certain cases, imposing restrictions on the clusters can be desirable or even required. Here we pursue *structured clustering*; that is, we impose restrictions on the shape of the clusters in the tensor, which in this application, allow us to interpret clusters in terms of data-generating mechanisms (i.e., grouping cell lines/ligand combinations to ensure mechanistic interpretation). See Results section for the biological justification of these constraints.

A hard partition of a dataset represented as a tensor $\mathbf{Z}$ of size $n_1 \times \cdots \times n_h$ into m clusters can be encoded in two ways:

1. An $(n_1 \times \ldots \times n_d) \times (n_1 \times \ldots \times n_d)$ tensor $\mathbf{X}$ in which the data has multi-indices $\mathbf{i} = (i_1, \ldots, i_d)$ and $\mathbf{j} = (j_1, \ldots, j_d)$, and:

$$x_{\mathbf{ij}} = \begin{cases} 0 & \text{if } \mathbf{i} \text{ and } \mathbf{j} \text{ belong to the same cluster,} \\ 1 & \text{otherwise.} \end{cases} \quad (2)$$

The tensor $\mathbf{X}$ can be seen as a Boolean approximation of the distances between pairs of data points: $x_{\mathbf{ij}} = 0$ if $\mathbf{i}$ and $\mathbf{j}$ are 'close' (in the same cluster), and $x_{\mathbf{ij}} = 1$ if they are 'far' (in different clusters). To ensure that $\mathbf{X}$ encodes a valid clustering of the data, the three conditions of an equivalence relation must be met. These conditions are given by the following algebraic equations and inequality:

$$\begin{aligned} \text{Reflexivity:} \quad & x_{\mathbf{ii}} = 0, \\ \text{Symmetry:} \quad & x_{\mathbf{ij}} = x_{\mathbf{ji}}, \\ \text{Transitivity:} \quad & 0 \leq -x_{\mathbf{ik}} + x_{\mathbf{ij}} + x_{\mathbf{jk}} \leq 2. \end{aligned} \quad (3)$$

2. In a $n_1 \times \ldots \times n_d \times m$ tensor $\mathbf{Y}$ where

$$y_{\mathbf{i}k} = \begin{cases} 1 & \text{if the data indexed by } \mathbf{i} \text{ belongs to cluster } k, \\ 0 & \text{otherwise.} \end{cases} \quad (4)$$

We require that $\sum_{k=1}^{m} y_{\mathbf{i}k} = 1$, to ensure that each data item has been assigned to exactly one cluster.

The tensors $\mathbf{X}$ and $\mathbf{Y}$ are related by the following equation:

$$1 - x_{\mathbf{i},\mathbf{j}} = \sum_{k=1}^{m} y_{\mathbf{i},k} y_{\mathbf{j},k}.$$

Integer Optimization. The structural or interpretability conditions we have imposed on the clusters take the form of linear constraints. These constraints, along with the fact that the tensors are Boolean, allow us to find optimal tensors $\mathbf{X}$ and $\mathbf{Y}$ by solving an integer linear program [18, 19]. Specifically we use the branch and cut algorithm [20] as we describe in the Structured clustering section below.

III. DATA

We examine an extensive experimental dataset detailing the temporal phosphorylation response of signaling molecules in genetically diverse breast cancer cell lines in response to different growth factors [1]. This dataset is complete and can be represented by a tensor $\mathbf{Z}$ of order 5 whose dimensions correspond to 36 cell lines, 14 ligands, 2 doses, 3 time points, and 2 proteins (pERK, pAKT) (see the SI Appendix). In this work each experiment is a set of measurements (ie doses, time points and proteins) for each cell line/ligand combination (i.e., $36 \cdot 14 = 504$ experiments). Our goal is to find sets of experiments with a similar response; consequently, the data structures we require are the following:

$$\begin{aligned} \mathbf{Z} &\in \mathbb{R}^{36 \times 14 \times 2 \times 3 \times 2}, && \text{(data tensor)} \\ \widetilde{\mathbf{Z}} &\in \mathbb{R}^{504 \times 12}, && \text{(flattened data tensor)} \\ \mathbf{S} &\in \mathbb{R}^{36 \times 14 \times 36 \times 14}, && \text{(similarity tensor)} \\ \widetilde{\mathbf{S}} &\in \mathbb{R}^{504 \times 504}. && \text{(similarity matrix)} \end{aligned}$$

Each experiment has a multi-index $\mathbf{i} = (i_1, i_2)$, where $i_1 \in \{1, \ldots, 36\}$ and $i_2 \in \{1, \ldots, 14\}$. We compute the 504×504 cosine similarity matrix $\widetilde{\mathbf{S}}$ of the normalized rows of $\widetilde{\mathbf{Z}}$ (see SI Appendix).

IV. STRUCTURED CLUSTERING

Clustering biological data is not a straightforward task; challenges include choosing the an appropriate method for the data [17], working with high-dimensional data [21, 22], and the consideration of the context of the problem, which must be done almost on a case-by-case basis [23]. Constrained clustering is an active field of research [24]. One of the most common approaches incorporates pairwise *must-link* and *cannot-link* constraints to indicate whether two items must or must not be in the same cluster [25, 26]. Other methods set constraints on what the possible clusters can be, rather than constraining the elements in a cluster; there is a large pool of candidate clusters from which those that meet selection criteria can be chosen [27].

Given a similarity tensor $\mathbf{S}$, we seek to obtain the best partition of experiments subject to constraints that ensure the interpretability of the clusters. To ensure that the clusters can be interpreted in terms of biological mechanisms, we impose that the clusters be rectangular-shaped with respect to cell lines and ligands (see eq. (5) and results section). This approach is similar to the one in Ref. [22]; however, we do not require the rectangles to be connected. This is because we do not require a fixed order for the rows and columns of the data. This is an important strength of our method: an ordering of the data is artificial, and we seek clustering results that are not biased by order. We describe the biological motivation for these constraints in the results section and mathematical details of the method here.

We present two implementations of our method. The first one does not require previous knowledge about the clustering assignment of the experiments, and provides an optimal clustering directly from the similarity data. This implementation can, however, be prohibitively expensive to compute. To sidestep high computational costs we present a second implementation that begins with a pre-existing partition of the experiments into clusters (not necessarily compliant with the constraints), which might originate from *any* clustering method (e.g., using the reshaped similarity tensor $\widetilde{\mathbf{S}}$). This implementation then reconstructs $\mathbf{S}$ and finds the nearest optimal clustering compliant with the constraints. Starting with an initial clustering has the advantage that it is less computationally expensive than starting with no prior knowledge of the clusters (see SI Appendix).

A. No prior clustering

This implementation works with the similarity tensor $\mathbf{S}$ whose entries record the similarity of experiments $\mathbf{i}$ and $\mathbf{j}$, where $\mathbf{i} = (i_1, i_2)$, $\mathbf{j} = (j_1, j_2)$, and the ranges of indices are $i_1, j_1 \in \{1, \ldots, 36\}$ and $i_2, j_2 \in \{1, \ldots, 14\}$.

The clustering assignments are recorded by the tensor $\mathbf{X}$ defined in equation (2). The rectangular-shaped interpretability condition corresponds to three simultaneous algebraic constraints on the entries of $\mathbf{X}$:

$$\begin{aligned}
&x_{i_1 i_2 j_1 j_2} = x_{i_1 j_2 j_1 i_2}, \\
&0 \le x_{i_1 i_2 j_1 j_2} - x_{i_1 i_2 j_1 i_2} \le 1, \\
&0 \le x_{i_1 i_2 j_1 j_2} - x_{i_1 i_2 i_1 j_2} \le 1.
\end{aligned} \tag{5}$$

We search over arrays $\mathbf{X}$ that satisfy these conditions, and find the one that best maximizes internal similarity, according to the following criterion. Experiments in the same cluster should have high similarity, so we maximize the similarity between experiments in the same cluster. This maximization is equivalent to solving the integer optimization problem

$$\begin{aligned}
\max_{\mathbf{X}} \quad &\langle \mathbf{S}, (\mathbf{1} - \mathbf{X}) \rangle + \lambda \langle \mathbf{1}, \mathbf{X} \rangle, \qquad (6) \\
\text{subject to} \quad &b_l \le \mathbf{V} \cdot \text{vec}(\mathbf{X}) \le b_u,
\end{aligned}$$

where the tensors $\mathbf{X}$ and $\mathbf{S}$ are as above, $\langle \cdot, \cdot \rangle$ denotes the entry-wise inner product, and $\cdot$ represents matrix multiplication of the matrix $\mathbf{V}$ by the vector $\text{vec}(\mathbf{X})$. The $504^2 \times 1$ vector $\text{vec}(\mathbf{X})$ is the vectorized form of $\mathbf{X}$, and $\mathbf{1}$ is the tensor of ones with the same size as $\mathbf{X}$. The coefficient λ is a regularization term introduced to control the number of clusters. The matrix $\mathbf{V}$ encodes the constraints on $\mathbf{X}$ given in eq. (3) and eq. (5). This matrix has over 1 million rows, 504^2 columns and is extremely sparse. The kth row of $\mathbf{V}$ represents the kth constraint to the values of $\text{vec}(\mathbf{X})$: the entry is the coefficient (which can be 0, 1, or −1, depending on the constraint) with which each entry of $\text{vec}(\mathbf{X})$ appears in the constraint. The kth entry of b_l and b_u (which can be 0, 1, or 2) give the lower and upper bounds respectively of each linear inequality. We solve this optimization program using the branch and cut algorithm [20] via the IBM ILOG CPLEX Optimization Studio [28].

The resulting rectangle-shaped clusters are a sparse, low-rank representation of the data. The tensor $\mathbf{1} - \mathbf{X}$, of size $(36 \times 14) \times (36 \times 14)$, gives a binary measure of the distance between any two experiments. This tensor has sparse block structure: it consists of m cuboids of 1s along the diagonal, where m is the number of clusters, and has zeros everywhere else. As a consequence $\mathbf{X}$ has low multilinear rank [29], bounded above by (m, m, m, m), which is less than the maximum possible value of $(36, 14, 36, 14)$.

B. Pre-existing clusters

Our method can find interpretable, structured clusters from an initial unstructured partition of the data. As before, we find structured clusters using linear integer optimization. The input is an initial partition of the 504 experiments into m clusters that do not have to be rectangular. We then modify the cluster assignments to reach the closest possible interpretable, structured clustering.

Mathematically, this implementation is set up in the

following way. The initial clustering is encoded by a partition tensor, $\mathbf{W}$, of size $36 \times 14 \times m$

$$w_{\mathbf{i}k} = \begin{cases} 1, & \mathbf{i} \text{ is in cluster } k, \\ 0, & \text{otherwise,} \end{cases}$$

where $\mathbf{i} = (i_1, i_2)$ indexes an experiment. The new clusters are encoded by a tensor $\mathbf{Y}$ of the same size (defined according to equation (4)). In order to have rectangular clusters, the entries of $\mathbf{Y}$ must satisfy the conditions

$$\sum_{r=1}^{m} y_{ijr} = 1, \quad \text{(unique cluster assignment)}$$

$$-1 \leq y_{ikr} + y_{jlr} - y_{ilr} \leq 1. \quad \text{(interpretability condition)}$$

As before, we use the branch and cut algorithm to obtain the global optimum (given $\mathbf{W}$) for the optimization problem

$$\max_{\mathbf{Y}} \quad \langle \mathbf{W}, \mathbf{Y} \rangle. \tag{7}$$

The inner product $\langle \mathbf{W}, \mathbf{Y} \rangle$ sums the number of clustering assignments unchanged by converting the initial unstructured clustering into a clustering that satisfies the interpretability constraints.

We obtain the tensor $\mathbf{Y}$, of size $36 \times 14 \times m$ by solving the optimization problem in eq. (7). As with $\mathbf{X}$, the tensor $\mathbf{Y}$ also has sparse and low-rank structure. Its m two-dimensional slices, each a matrix of size 36×14, have rank two and block structure with a rectangle shape populated by 1s and all other values equal to 0.

V. RESULTS

Each experiment in our data is indexed by (c, l), where c is the cell line, and l is the ligand. A high similarity between experiments suggests the possibility of a common underlying biological mechanism. This is the basic notion that underpins the constraints in our clustering method. The biological hypothesis is that if experiments (c_1, l_1) and (c_2, l_2) are in the same cluster, it is possible that cell lines c_1 and c_2 share a feature such as a genetic mutation. This feature causes them to respond similarly to ligands l_1 and l_2, and the genetic mutation is relevant to the part of the signaling pathways these ligands activate. Such biological causality would imply that if we swapped the ligands, experiments (c_2, l_1) and (c_1, l_2) should also respond in a similar way. The biological mechanism that gives rise to the similarity between (c_1, l_1) and (c_2, l_2) can be discounted if it is not also reflected in experiments (c_1, l_2) and (c_2, l_1); this is why we require that the clusters of experiments have a rectangular shape.

A. Interpretable groups by mutation and receptor subtype

In a clinical setting, prognosis and treatment decisions for breast cancer are guided by tumor grade, stage and clinical subtype [30], which is based on the presence of cellular receptors:

- HER2$^{\mathrm{amp}}$ cells are characterized by amplification of the HER2 gene, leading to over-expression of the ErbB2 receptor tyrosine kinase;

- HR$^+$ cells are characterized by the expression of the estrogen receptor (ER) or progesterone receptor (PR);

- Triple negative breast cancer (TNBC) cells are negative for HER2 amplification and express ER and PR at low levels.

We compare the clusters from our method with the three standard clinical subtypes above. We also compare our clusters with the mutational status of the cell lines [31, 32], and with their drug response [33].

Clustering clinical subtypes. We first investigate a fine-grained classification within each of the three clinical subtypes. A summary statistic between 0 and 1 (based on the cosine similarity, see SI Appendix) quantifies the within-class variation for each clinical subtype. A score of 0 indicates complete homogeneity, and 1 indicates complete heterogeneity. The HER2$^{\mathrm{amp}}$ cell lines show comparatively little variation, with an average difference score of 0.086. The TNBC and the HR$^+$ cell lines have an average difference score of 0.224 and 0.334. We obtained clusters without prior knowledge of an initial clustering by solving the optimization problem (6). The results (shown in Fig. 2A and 2B) identify heterogeneity within each subtype as well as cell lines of particular interest.

The TNBC cell lines are divided into 12 clusters (Fig. 2A) which mirror the heterogeneous behavior of TNBC in the clinic [34]. All but one TNBC cell lines with a PTEN mutation appear in the green cluster. The only exception is the HCC1937 cell line which has a PTEN mutation but appears in the yellow cluster. The cluster assignment of cell lines MDA-MB-231 and MDA-MB-157 is markedly different than the other cells across the ligands. These assignments might be explained by the mutational status of the cell lines; MDA-MB-231 is the only cell line with an NF2 mutation or a BRAF mutation, whereas MDA-MB-157 is the only cell line with an NF1 mutation. The bright orange cluster contains five cell lines (all but HCC1937) with the same two CDKH2A mutations.

Figure 2B shows the clustering of the HR$^+$ cell lines. Cell line MDA-MB-415 stands out for its response to so-called high-response ligands [1] (ligands to the left of HRG in Fig. 2B). Among all cell lines, MDA-MB-415 has the second highest susceptibility to the drugs Ixabepilone, Methylglyoxal and PD [33]. CAMA-1 cell line, is distinctive in its response to the low-response ligands (to
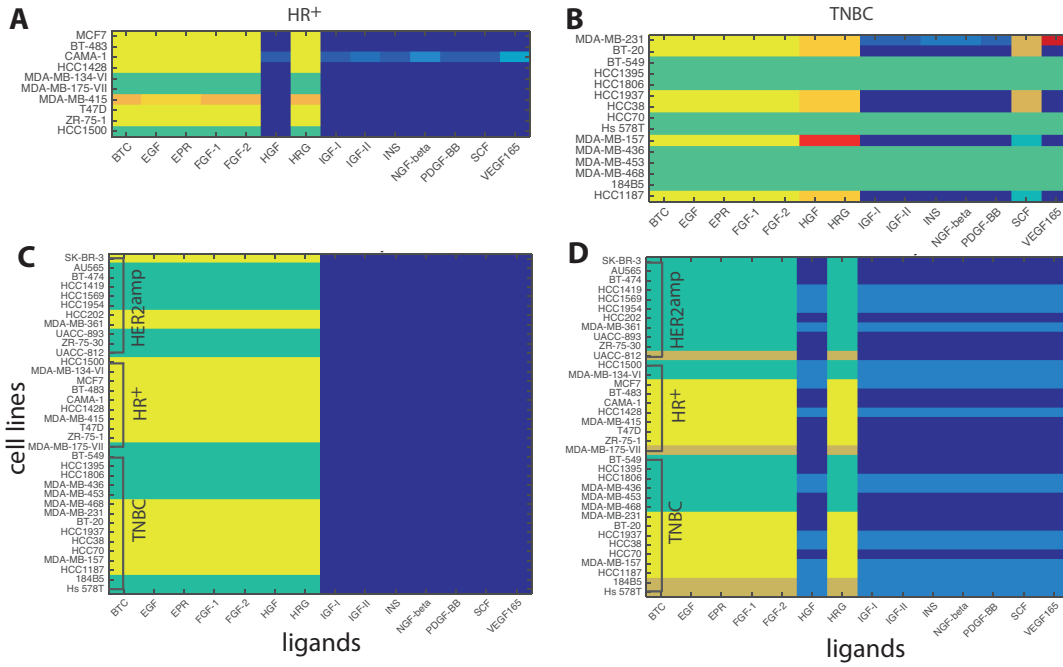
FIG. 2. Tensor-based structured clustering. (A) TNBC clustering with no prior clustering information. (B) HR$^+$ clustering with no prior information. (C) Clustering of all cell lines starting from an initial partition into three clusters. (D) Clustering from an initial partition into five clusters.

the right of HRG), which might help explain why it is particularly susceptible to both (Z)-4-Hydroxytamoxifen and TCS PIM-11 [33].

The HER2^{amp} cell lines cluster together for all ligands except for the MDA-MB-361 cell line. This is the HER2^{amp} cell line most resistant to HER2-targeted therapy such as Lapatinib [33]. In fact, its resistance to Lapatinib exceeds that of some TNBC cell lines (HCC2185 and MDA-MB-453). The grouping of the rest indicates the consistency among all other HER2^{amp} cell lines (see SI Appendix).

Clustering all cell lines. We now apply our method to the full dataset. Here we solve the optimization problem (7), which means that we rely on an initial clustering to speed up computations (see the SI Appendix). The groups we find respect the broad division of the cell lines seen in Fig. 2A,B, which is a sign of the consistency between the two implementations of our method.

We obtained our initial clustering by first constructing a graph of experiments using the Relaxed Minimum Spanning Tree algorithm [2–4]. Then we used the Markov Stability community detection method [5, 6] to obtain the clusters. Markov Stability identifies robust partitions of the experiments into three, five and seven groups (see SI Appendix).

From the initial partition into three clusters, we obtain three rectangular clusters (Fig. 2C). Of these, we find that two groups of ligands correspond to previously reported high active expression profiles (yellow and green) and one to muted profiles (blue), respectively [1]. Within the more highly active group, the HR$^+$ cell lines are pre-

dominantly in the yellow cluster, while the HER2^{amp} cells are in the green cluster. This separation of the HR$^+$ and HER2^{amp} clinical subtypes is entirely data driven and supports the notion that our method is indeed able to find interpretable groups. The cell lines that are not clustered according to their subtype reflect previous findings that neither growth factor responses nor sensitivity to drugs that target signal transduction pathways is uniform within clinical subtypes [1, 16, 19]. The TNBC cell lines are divided between the yellow and green clusters, providing further evidence of the heterogeneity in TNBC cell lines [19, 41–45].

When we use initial clustering into five groups, the rectangular clusters split the ligands into a low response group (blues) and high response (green, yellow, brown). This split is nearly the same as we obtained before (Fig. 2D). Note that the difference in the ligand HGF may be due to the fact that it is not part of the ErbB nor the FGF families of ligands. The HER2^{amp} cell lines are now all assigned to the green cluster and there are only three HR$^+$ cell lines not assigned to the yellow cluster. A new brown cluster consists of cell lines: MDA-MB-175-VII (classified as a HR$^+$), UACC-812 (HER2^{amp}), 1845B5 (TNBC) and HS578T (TNBC). While none of them have the same cell classification or genetic mutation, all cell lines in the brown cluster show high susceptibility to Gefitinib [33]. Note that MDA-MB-175-VII is the only HR$^+$ cell line that is not assigned to the yellow group in either three or five clusters; this might be due to the fact that this cell line carries a unique chromosomal translocation. The translocation leads to the fusion

and amplification of neuregulin-1 which signals through ErbB2/ErbB3 heterodimers [46, 47], and could be the underlying cause of the cell line's unique sensitivity to ErbB-targeting drugs like Lapatinib or Afatinib [16, 19].

Finally, the clustering that begins with an initial partition into seven groups shows high consistency with the five cluster case (see SI Appendix). We therefore continue our analysis on the five clusters.

B. Systematic model identification

We now analyze the response of the five structured groups found in the previous section (Fig. 2D) to obtain a mechanistic insight about how the cell line/ligand combinations in them respond to the stimuli. We perform systematic model analysis of 729 possible ODE network models, and find that with our data 44 are structurally identifiable, a prerequisite for performing parameter estimation and model selection. Then, we parametrize, rank and choose the models that best represent each cluster's response. As a result, we have a list of candidate signaling mechanisms for each cluster which provides more information than the statistical predictions of the sensitivity of MAP Kinase drug targets (e.g., ErbB drug class) [19].

Models of the MAPK and AKT pathways have been studied under a variety of biological and modeling assumptions [8, 9], including pathway crosstalk [13, 15–17]. Here we consider simple models to ensure the parameters are at least locally identifiable so there are a finite number of parameter values to fit the data. See the SI Appendix for a brief synopsis of MAPK models.

We construct nonlinear ordinary differential equation models to describe the dynamics of the AKT and ERK signaling pathways. These models include three molecular species: Receptor (R), pERK (E) and pAKT (A). Since the data contains the response of pERK and pAKT, we assume that the receptor must phosphorylate ERK and/or AKT. We consider positive, negative, or no interaction between pERK and pAKT under different types of kinetic regimes (mass-action or Michaelis–Menten) and different types of inhibition (blocking/sequestration or removal/degradation). The combination of these features result in the 44 structurally identifiable models that we study in further detail. Each model corresponds to a different mechanistic hypothesis of the dynamics in the pathways (see SI Appendix). To find the models that best describe the response of each of the five clusters, we estimate parameters using the Squeeze-and-Breathe algorithm [21], and rank them using the Akaike information criterion score (AICc) (see SI Appendix). The best models for each cluster are shown in Fig. 3.

The AICc penalizes more complex models; therefore it is not surprising that the top models are the simplest ones. The best models for each cluster have different feedback strengths (parameter values) and network topologies (see Fig. 3); this supports the hypothesis that
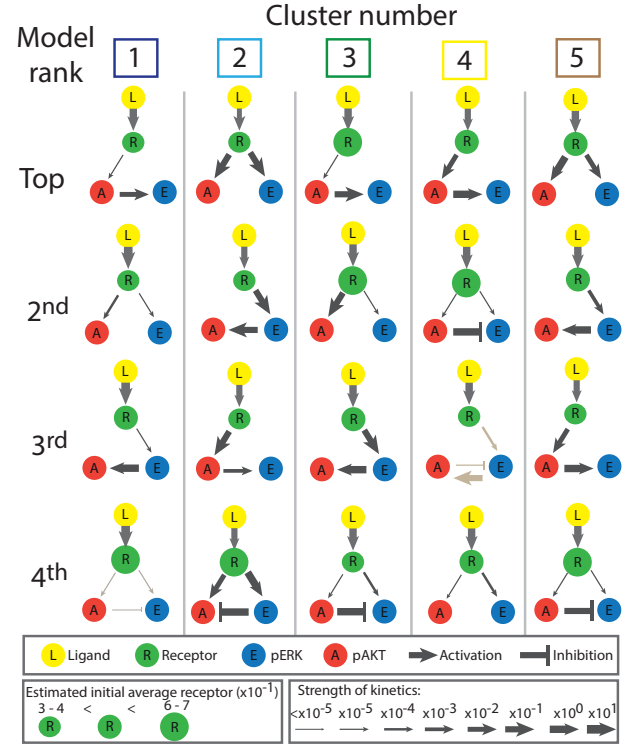

FIG. 3. The top four models for each cluster. The strength of interactions are indicated by the size of the arrow. The grey arrows indicate a blocking mechanism for inhibition. Black inhibition arrows indicate a removal mechanism for inhibition.

mutations may play a role in the dynamics. Although the values of the parameters vary, the model with arrows from the receptor to pAKT and pERK appears in all clusters, which is in line with how cells are understood to operate. We remark that cluster 4, which corresponds to HR^+ cells (yellow in Fig. 2D), includes inhibition crosstalk as the second best model, whereas in all other clusters this mechanism appears in fourth place. These results support the notion that clusters may have different parameter values/network structure and that our approach is viable.

VI. DISCUSSION

We have introduced a novel framework to cluster multi-indexed data based on tensors that allows structural constraints to be incorporated using algebraic relationships. This method can be used to extract clusters directly from the data, but if an initial clustering (which may not satisfy the constraints) is provided it can find the closest optimal partition that satisfies the constraints and reduce computation time. A key advantage of this framework is that it allows more control over the composition of clusters than in many unsupervised methods, and to tailor the clustering to the requirements of the problem at hand.

We applied this method on a dataset charting the response of genetically diverse breast cancer cell lines to ligands. We identified both similarities (e.g., HER2$^{\text{amp}}$) and heterogeneities (e.g., TNBC) within clinical subtypes. The heterogeneity of our clustering analysis (see Fig. 2B) seems to be related to both the mutational status of the cells as well as their response to inhibitors. This result means that similar analyses in patient tissues might be able to identify patients that respond differently to therapeutic methods commonly used within a clinical subtype. By analyzing clusters from all subtypes, we also showed that we cannot explain the dynamics of each data cluster with only one mathematical model, which helps explain network model differences across cell type.

The applicability of our method goes beyond the biological problem presented here. It can be used in *any* context in which the constraints on the clusters can be expressed in algebraic form (as equalities and inequalities), such as when there are size restrictions on the clusters, or to impose/prohibit particular combinations of data beyond *must-link* and *cannot-link* constraints. For example, this method could be used to construct optimal portfolios that comply with rules about their composition [53], to help the formation of teams that maximize members' preferences and are compliant with skill requirements [54], to find communities in networks with quotas, among others. The presented pipeline (a sophisticated and interpretable data analysis method that feeds into a nonlinear modeling framework) will be ever more necessary as increasingly more large-scale, comprehensive datasets become available.

[1] Kolda TG, Bader BW (2009) Tensor decompositions and applications. *SIAM REVIEW* 51(3):455–500.

[2] Hore V, et al. (2016) Tensor decomposition for multiple-tissue gene expression experiments. *Nature Genetics* 48:1094–1100.

[3] Austin W, Ballard G, Kolda TG (2016) Parallel tensor compression for large-scale scientific data. *Proc. IPDPS.*

[4] Sankaranarayanan P, Schomay TE, Aiello KA, Alter O (2015) Tensor gsvd of patient- and platform-matched tumor and normal dna copy-number profiles uncovers chromosome arm-wide patterns of tumor-exclusive platform-consistent alterations encoding for cell transformation and predicting ovarian cancer survival. *PLoS One.*

[5] von Kriegsheim A, et al. (2009) Cell fate decisions are specified by the dynamic ERK interactome. *Nature* 11(12):1458–1464.

[6] Marshall CJ (1995) Specificity of receptor tyrosine kinase signaling: transient versus sustained extracellular signal-regulated kinase activation. *Cell* pp. 179–185.

[7] Purvis JE, Lahav G (2013) Encoding and Decoding Cellular Information through Signaling Dynamics. *Cell* 152(5):945–956.

[8] Chen WW, et al. (2009) Input–output behavior of erbb signaling pathways as revealed by a mass action model trained against dynamic data. *Molecular Systems Biology* 5(1).

[9] Won JK, et al. (2012) The crossregulation between erk and pi3k signaling pathways determines the tumoricidal efficacy of mek inhibitor. *Journal of Molecular Cell Biology* 4(3):153–163.

[10] McCubrey JA, et al. (2011) Therapeutic resistance resulting from mutations in raf/mek/erk and pi3k/pten/akt/mtor signaling pathways. *Journal of Cellular Physiology* 226(11):2762–2781.

[11] Serra V, et al. (2011) Pi3k inhibition results in enhanced her signaling and acquired erk dependency in her2-overexpressing breast cancer. *Oncogene* 30(22):2547–2557.

[12] Hanahan D, Weinberg R (2011) Hallmarks of cancer: The next generation. *Cell* 144(5):646 – 674.

[13] Baselga J (2006) Targeting tyrosine kinases in cancer: The second wave. *Science* 312(5777):1175–1178.

[14] Niepel M, et al. (2014) Analysis of growth factor signaling in genetically diverse breast cancer lines. *BMC Biol* 12:20.

[15] Kolch W, Halasz M, Granovskaya M, Kholodenko BN (2015) The dynamic control of signal transduction networks in cancer cells. *Nat Rev Cancer* 15(9):515–527.

[16] Heiser LM, et al. (2012) Subtype and pathway specific responses to anticancer compounds in breast cancer. *Proc Natl Acad Sci* 109(8):2724–2729.

[17] Lebart L, Morineau A, Warwick K (1984) *Multivariate Descriptive Statistical Analysis: Correspondence Analysis and Related Techniques for Large Matrices*, Probability and Statistics Series. (Wiley).

[18] Nemhauser GL, Wolsey LA (1999) *Integer and Combinatorial Optimization.* (Wiley).

[19] Bertsimas D, Weismantel R (2005) *Optimization over the Integers.* (Dynamic Ideas).

[20] Mitchell JE (2002) Branch-and-cut algorithms for combinatorial optimization problems. *Handbook of Applied Optimization* pp. 65–77.

[21] Ver Steeg G, Galstyan A (2014) Discovering structure in high-dimensional data through correlation explanation in *Advances in Neural Information Processing Systems.* pp. 577–585.

[22] Madeira SC, Oliveira AL (2004) Biclustering algorithms for biological data analysis: A survey. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 1(1):24–45.

[23] Luxburg UV, Williamson RC, Isabelle (2009) Clustering: Science or art? in *NIPS 2009 workshop on clustering theory.* pp. 1–11.

[24] Basu S, Davidson I, Wagstaff K (2008) *Constrained clustering: Advances in algorithms, theory, and applications.* (CRC Press).

[25] Wagstaff K, Cardie C, Rogers S, Schrödl S (2001) Constrained k-means clustering with background knowledge in *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01. (Morgan Kaufmann Publishers Inc., San Francisco, CA, USA), pp. 577–584.

[26] Davidson I, Basu S (2007) A survey of clustering with instance level constraints. *ACM Transactions on Knowledge Discovery from Data* pp. 1–41.

[27] Mueller M, Kramer S (2010) *Integer Linear Programming Models for Constrained Clustering*, eds. Pfahringer B, Holmes G, Hoffmann A. (Springer Berlin Heidelberg, Berlin, Heidelberg), pp. 159–173.

[28] IBM (2011) *IBM ILOG CPLEX Optimization Studio CPLEX User's Manual.*

[29] Lathauwer LD, Moor BD, Vandewalle J (2000) A multilinear singular value decomposition. *SIAM J. Matrix Anal. Appl.* 21(4):1253–1278.

[30] (2016) http://www.cancer.gov/.

[31] Hollestelle A, Elstrodt F, Nagel JHA, Kallemeijn WW, Schutte M (2007) Phosphatidylinositol-3-OH kinase or RAS pathway mutations in human breast cancer cell lines. *Molecular cancer research : MCR* 5(2):195–201.

[32] Bamford S, et al. (2004) The cosmic (catalogue of somatic mutations in cancer) database and website. *Br J Cancer* 91(2):355–358.

[33] Hafner M, Niepel M, Chung M, Sorger PK (2016) Growth rate inhibition metrics correct for confounders in measuring sensitivity to cancer drugs. *Nature Methods* 13(6):521–532.

[34] Podo F, et al. (2010) Triple-negative breast cancer: Present challenges and new perspectives. *Molecular Oncology* 4(3):209 – 229. Thematic Issue: The Molecular Biology of Breast Cancer.

[35] Vangelov B (2014) Ph.D. thesis (Imperial College London).

[36] Beguerisse-Díaz M, Vangelov B, Barahona M (2013) Finding role communities in directed networks using Role-Based Similarity, Markov Stability and the Relaxed Minimum Spanning Tree. *2013 IEEE Global Conference on Signal and Information Processing (GlobalSIP)* pp. 937–940.

[37] Beguerisse-Díaz M, Garduño Hernández G, Vangelov B, Yaliraki SN, Barahona M (2014) Interest communities and flow roles in directed networks: the Twitter network of the UK riots. *J R Soc Interface* 11(101).

[38] Delvenne JC, Yaliraki S, Barahona M (2010) Stability of graph communities across time scales. *Proc Nat Acad Sci USA* 107(29):12755–12760.

[39] Delvenne JC, Schaub MT, Yaliraki SN, Barahona M (2013) The stability of a graph partition: A dynamics-based framework for community detection in *Dynamics On and Of Complex Networks, Volume 2*, Modeling and Simulation in Science, Engineering and Technology, eds. Mukherjee A, Choudhury M, Peruani F, Ganguly N, Mitra B. (Springer New York), pp. 221–242.

[40] Niepel M, et al. (2013) Profiles of basal and stimulated receptor signaling networks predict drug response in breast cancer lines. *Sci Signal* 6(294):ra84.

[41] Kirouac DC, et al. (2016) Her2+ cancer cell dependence on pi3k vs. mapk signaling axes is determined by expression of egfr, erbb3 and cdkn1b. *PLoS Comput Biol* 12(4):e1004827.

[42] Kirouac DC, et al. (2013) Computational modeling of erbb2-amplified breast cancer identifies combined erbb2/3 blockade as superior to the combination of mek and akt inhibitors. *Sci Signal* 6(288):ra68.

[43] Kirouac DC, et al. (2015) Model-based design of a decision tree for treating her2+ cancers based on genetic and protein biomarkers. *CPT Pharmacometrics Syst Pharmacol* 4(3):e00019.

[44] Shah SP, et al. (2012) The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature* 486(7403):395–399.

[45] Shastry M, Yardley DA (2013) Updates in the treatment of basal/triple-negative breast cancer. *Curr Opin Obstet Gynecol* 25(1):40–48.

[46] Schaefer G, Fitzpatrick VD, Sliwkowski MX (1997) Gamma-heregulin: a novel heregulin isoform that is an autocrine growth factor for the human breast cancer cell line, MDA-MB-175. *Oncogene* 15(12):1385–1394.

[47] Liu X, Baker E, Eyre HJ, Sutherland GR, Zhou M (1999) Gamma-heregulin: a fusion gene of DOC-4 and neuregulin-1 derived from a chromosome translocation. *Oncogene* 18(50):7110–7114.

[48] Heinrich R, Neel BG, Rapoport TA (2002) Mathematical Models of Protein Kinase Signal Transduction. *Molecular Cell* 9:957–970.

[49] Huang CY, Ferrell JE (1996) Ultrasensitivity in the mitogen-activated protein kinase cascade. *Proc Natl Acad Sci USA* 93(19):10078–10083.

[50] Fujita KA, et al. (2010) Decoupling of receptor and downstream signals in the akt pathway by its low-pass filter characteristics. *Sci Signal* 3(132):ra56.

[51] Fey D, Croucher DR, Kolch W, Kholodenko BN (2012) Crosstalk and signaling switches in mitogen-activated protein kinase cascades. *Frontiers in physiology* 3:355.

[52] Beguerisse-Díaz M, Wang B, Desikan R, Barahona M (2012) Squeeze-and-breathe evolutionary Monte Carlo optimization with local search acceleration and its application to parameter fitting. *J R Soc Interface* 9(73):1925–1933.

[53] McNeil A, Frey R, Embrechts P (2015) *Quantitative Risk Management: Concepts, Techniques and Tools*, Princeton Series in Finance. (Princeton University Press).

[54] Davis EW, Heidorn GE (1971) An algorithm for optimal project scheduling under multiple resource constraints. *Management Science* 17(12):B–803–B–816.

SUPPLEMENTARY INFORMATION FOR: TENSOR CLUSTERING WITH ALGEBRAIC CONSTRAINTS GIVES INTERPRETABLE GROUPS OF CROSSTALK MECHANISMS IN BREAST CANCER

Appendix A: Data

We analyse a data set first presented in Ref. [1]. The data consists of time-course measurements (at 0, 10, 30, and 90 minutes) of the fold change in the phosphorylation levels of the mitogen activated protein kinase (MAPK) pERK and the phosphoinositide 3-kinase (PI3K) pAKT in 36 breast cancer cell lines (Table I), each exposed to two doses (low: 1 ng/ml and high: 100 ng/ml) of 14 different ligands (Table II). Figure S1 shows an example of the temporal response of pAKT and pERK in the cell line MCF7 to doses of betacellulin.

The data set is complete in the sense that it contains measurements of pAKT and pERK in every cell line, exposed to 2 doses of each of the 14 ligands at 4 time points. As a result, we are able to represent the data as a $\times 36 \times 14 \times 2 \times 3 \times 2$ tensor $\mathbf{Z}$, with entries z_{ijptd}. The index $p \in \{\text{pAKT, pERK}\}$ denotes which protein was measured, i corresponds to the cell lines in Table I, j corresponds to the ligands in Table II, $d \in \{1, 100\}$ denotes the doses, and $t \in \{10, 30, 90\}$ is time after the stimulus.

Appendix B: Clustering

We first describe how we normalize the experimental data for clustering. Then we provide details for how we clustered directly from the data, and then how we obtained and incorporated an initial clustering.

1. Normalization

We normalize the data as follows. We scale all the pAKT and pERK foldchange responses such that their average takes each the value 1. This normalization balances the effects of AKT and ERK, so that the behavioural features and not the scale are the dominating features, and to ensure we treat them with equal significance in our study. The mean value (pre-normalization) across the AKT responses is 1.7754 and that across the ERK responses is 11.4190.

2. No prior clustering

As a summary statistic of the three clinical subtypes, we compute the average distance score within each subtype. The score for the 11 HER2$^{\text{amp}}$ cell lines is obtained as follows. There are $154 = 11 \times 14$ experiments in the dataset that involve HER2$^{\text{amp}}$ cell lines, each consisting of 12 measurements. For each pair of experiments, we find the dissimilarity between the 12 measurements, using cosine dissimilarity `1 - dot(v1,v2)/(norm(v1,2)*norm(v2,2))`. The average

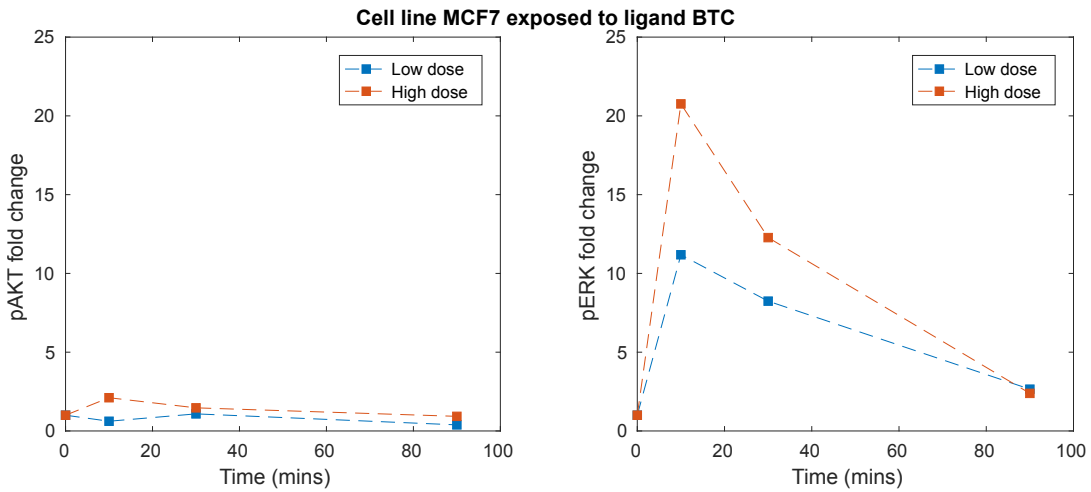
FIG. S1. Time course measurements of cell line MCF7 exposed to two doses of betacellulin (BTC).

Cell line	Subtype
MCF7	HR+
SK-BR-3	HER2amp
MDA-MB-231	TNBC
AU-565	HER2amp
BT-20	TNBC
BT-474	HER2amp
BT-483	HR+
BT-549	TNBC
CAMA-1	HR+
HCC-1395	TNBC
HCC-1419	HER2amp
HCC-1428	HR+
HCC-1569	HER2amp
HCC-1806	TNBC
HCC-1937	TNBC
HCC-1954	HER2amp
HCC-202	HER2amp
HCC-38	TNBC
HCC-70	TNBC
Hs 578T	TNBC
MDA-MB-134VI	HR+
MDA-MB-157	TNBC
MDA-MB-175VII	HR+
MDA-MB-361	HER2amp
MDA-MB-415	HR+
MDA-MB-436	TNBC
MDA-MB-453	TNBC
MDA-MB-468	TNBC
T47D	HR+
UACC-812	HER2amp
UACC-893	HER2amp
ZR-75-1	HR+
ZR-75-30	HER2amp
184-B5	TNBC
HCC-1187	TNBC
HCC-1500	HR+

TABLE I. Breast cancer cell lines used in the data set [1].

Ligand name	Abbreviation
Betacellulin	BTC
Epidermal Growth Factor	EGF
Epiregulin	EPR
Fibroblast Growth Factor (acidic)	FGF-1
Fibroblast Growth Factor (basic)	FGF-2
Hepatocyte Growth Factor	HGF
Heregulin $\beta 1$	HRG
Insulin-like Growth Factor 1	IGF-1
Insulin-like Growth Factor 2	IGF-2
Insulin	INS
Nerve Growth Factor	NGF-beta
Platelet Derived Growth Factor BB	PDGF-BB
Stem Cell Factor	SCF
Vascular endothelial growth factor A	VEGF165

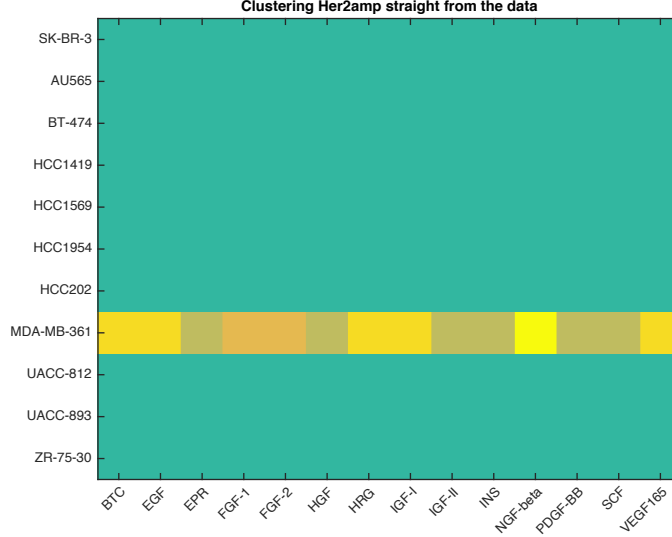TABLE II. Ligands used in the data set [1].

FIG. S2. Clusters from the data of HER2$^{\mathrm{amp}}$ cell line.

dissimilarity is obtained by averaging these pairwise distances across the $\binom{154}{2} = 11781$ pairs. Similarly for the HR$^+$ cell lines and for the TNBC cell lines. The averages obtained are 0.086 for HER2$^{\mathrm{amp}}$, 0.334 for HR$^+$, and 0.224 for TNBC. The partitioning of the HER2$^{\mathrm{amp}}$ cell lines is given in Fig. S2.

3. Pre-existing clusters

a. *Computing pre-existing clusters*

We find an initial clustering of the experiments. For this initial clustering, we label each experiment by a single index. The data for the ith experiment is:

$$\widetilde{\mathbf{Z}}(i,:) = \left[\mathbf{AKT}_i^1 \middle| \mathbf{ERK}_i^1 \middle| \mathbf{AKT}_i^{100} \middle| \mathbf{ERK}_i^{100} \right], \tag{B1}$$

where $\mathbf{AKT}_i^1$ is the normalised time series of fold-change response of pAKT under dose 1ng/ml, and so on. We compute the 504×504 similarity matrix $\widetilde{\mathbf{S}}$, in which s_{ij} indicates the cosine similarity of experiments i and j:

$$s_{ij} = \frac{\langle \tilde{\mathbf{z}}_i, \tilde{\mathbf{z}}_j \rangle}{||\tilde{\mathbf{z}}_i||_2 \, ||\tilde{\mathbf{z}}_j||_2} = \cos\left(\tilde{\mathbf{z}}_i, \tilde{\mathbf{z}}_j \right). \tag{B2}$$

where $\tilde{\mathbf{z}}_i = \widetilde{\mathbf{Z}}(i,:)$ and $\tilde{\mathbf{z}}_j = \widetilde{\mathbf{Z}}(j,:)$. The entries of $\tilde{\mathbf{z}}_i$ and $\tilde{\mathbf{z}}_j$ are nonnegative, which means that $s_{ij} \in [0,1]$. If $s_{ij} = 1$, experiments i and j have an *identical* response to the treatments in both AKT and ERK (up to a scaling constant). When $s_{ij} = 0$, the data for the experiments are orthogonal. The task of clustering the experiments presents two challenges: the number of clusters is not known a priori, and the matrix $\widetilde{\mathbf{S}}$ is full matrix and is noisy due to experimental error. To tackle these challenges we use a combination of tools from manifold learning and network science. We create a network (graph) in which each of the 504 experiments is represented by a node, and where connections exist between similar experiments. We first define the dissimilarity matrix $\mathbf{D}$ ($d_{ij} = 1 - s_{ij}$). We then use the *Relaxed Minimum Spanning Tree* (RMST) algorithm [2–4], which extracts a network representation from high-dimensional point clouds (in this case the $\tilde{\mathbf{z}}_i$) that are embedded in a lower dimensional manifold.

Specifically, the algorithm creates an undirected, unweighted network with an edge between i and j if they are neighbors in a minimum spanning tree (MST) from $\mathbf{D}$. The algorithm adds extra edges to the network if they are consistent with the continuity of the data, i.e. if the distances between the points in $\mathbf{D}$ is comparable to their separation in the MST and is consistent with the continuity of the data, according to the equation

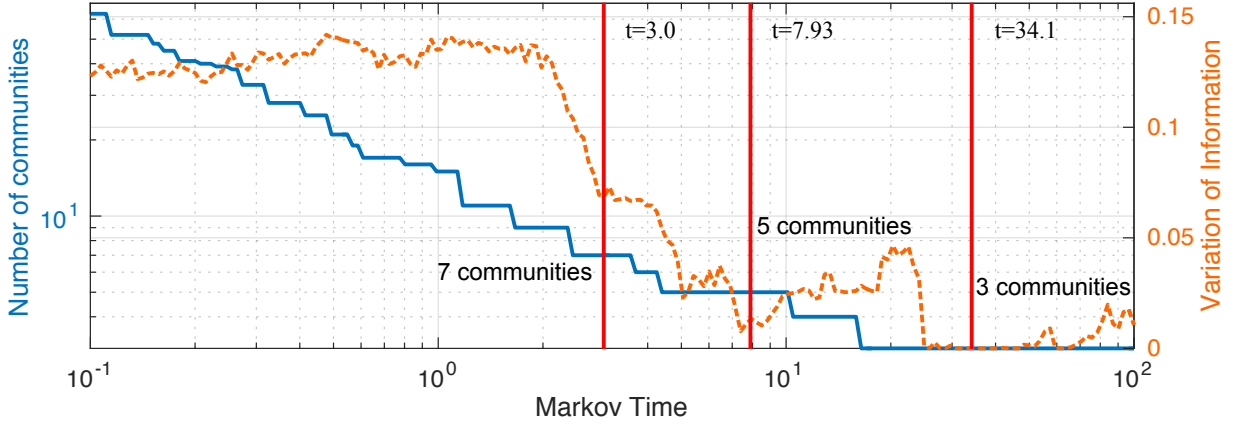$$d_{ij} < \mathrm{mlink}_{ij} + \frac{1}{2}(\mathrm{k_i} + \mathrm{k_j}).$$

FIG. S3. Number of communities and variation of information for the network obtained from RMST similarity graph.

Here mlink_{ij} is the maximal edge weight in the MST path connecting i to j, and k_i is the the distance to the nearest neighbour of i (i.e., the minimum value on the ith row of $\mathbf{D}$, excluding d_{ii}). Basically, what the RMST algorithm does is allows edges to be added to the MST (it 'relaxes' the MST), so that we obtain an network description of high dimensional data that is embedded on a lower dimensional manifold.

Once we have obtained the network from the similarity matrix, we extract communities using the Markov Stability (MS) community detection algorithm [5, 6]. This method employs continuous time random walks of varying duration (Markov time) to extract communities of the network at different levels of resolution. Shorter Markov times produce small communities, whereas longer Markov times lead to coarse partitions of the network. Obtaining the optimal partition of a network into communities is an NP complete problem, so MS uses heuristics to find communities. Because there is no guarantee of finding a global optimum, MS repeats the heuristic search 100 times for each Markov time. The variability in each set of 100 solutions is measured with the Variation of Information (VI) [7]. A low value of the VI for a Markov time indicates that the solutions obtained are similar to each other, we take this similarity as a sign that there is a robust partition of the network for this Markov time. In Fig. S3 we show that the network A has a robust partition into 3, 5 and 7 communities.

b. Structured clusters from pre-existing clusters

We present the clustering assignments after using MS to obtain our initial partition into clusters. The pre-existing and structured cluster results for three clusters (Fig. S4), five clusters (Figs. 2D, S4) and seven clusters (Fig. S6).

A finer clustering into seven groups (see Fig. S6) divides the ligands into three groups: high response (yellow, green, brown) {BTC, EGF, EPR, FGF-1, FGF-2, HGF, HRG} and lower response (blues) {IGF1, IGF2, INS} and { NGF-β, PDGF-BB, SCF, VEGF175}. The assignment of cell types is remarkably similar to the five cluster results. The exception cell lines are: HCC1419 (green to brown), and ZR-75-30 (green to the new seventh cluster). Given this consistency, we restrict to mechanistic interpretation of the five cluster case.

4. Comparison of implementations

When we employ our method assuming no initial clusterings, the integer optimization is learning the values, 0 or 1, for an array of size $c \times c \times l \times l$ (where $c = $ #cell lines and $l = $ # ligands). When we use an initial clustering, CPLEX uses the same branch and cut algorithm on an array of size $c \times l \times m$, where $c = $ #cell lines, $l = $ #ligands and $m = $ #clusters, so the array is much smaller for the pre-existing clustering implementation. We show a comparison of the performance of the two implementations in Fig. S7.
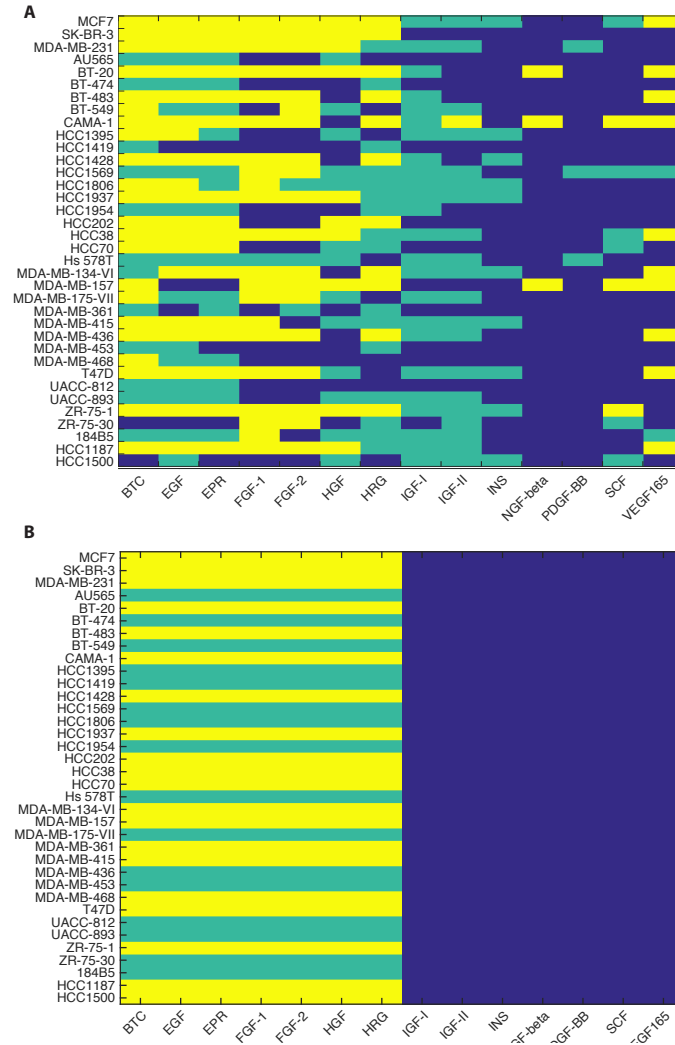
FIG. S4. A. The clustering assignments are represented by yellow, green and blue squares. B. The tensor clustering are a close approximation to the clusters obtained with MS.

Appendix C: Models

1. Models of the MAPK pathway

The MAP Kinase pathway has been widely studied. Models have focused on various features of the cascade, such as its three-tier phosphorylation feedback structure [8, 9]. The activation profile of these kinases is directly related to cellular decisions and fates [10–13]. The AKT pathway has been modeled by EGF-dependent activation, including phosphorylation of AKT (pAKT) and its downstream intracellular proteins [14]. Few models of crosstalk between ERK and AKT exist. One model was created for studying PC12 cells, and they found that AKT acts as a low-pass filter which decouples the EGF signal [15]. Another model was created to study HEK293 cells in the presence of a MEK inhibitor; they found crosstalk is reinforced between Ras and PI3K [16]. Another model found that JNK is regulated by AKT and MAPK feedbacks in these pathway [17]. Chen and co-authors constructed and analyzed an ErbB model focusing on the receptor dynamics and early activation response of the MAPK and AKT pathways in response to ligands in A431 or H1666 cells [13]; however their model was unidentifiable, meaning there were an infinite number of parameter values to fit the data [13]. Weakly activated models of MAPK activation cascades with optimal amplification under a variety of stimuli were analyzed in [18].
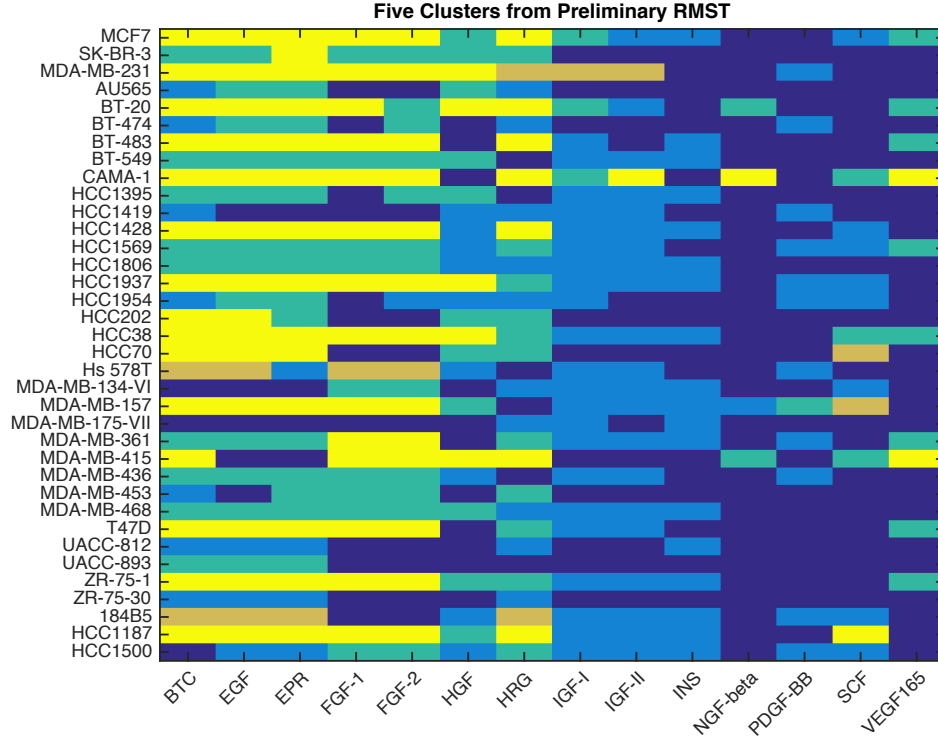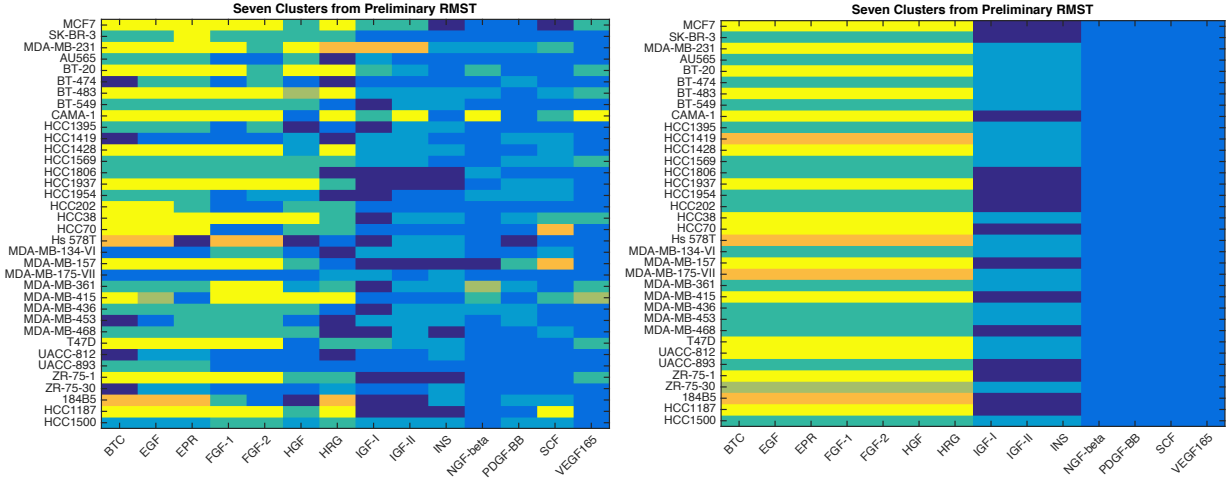
FIG. S5. Five clusters from Markov Stability

FIG. S6. Seven clusters from Markov Stability.

2. All wiring diagrams

We consider all possible wiring diagrams to describe the interactions between the receptor, the Erk pathway and the Akt pathway. These can be written as a wiring-diagram where we assume an arrow exists between the ligand L and receptor R:
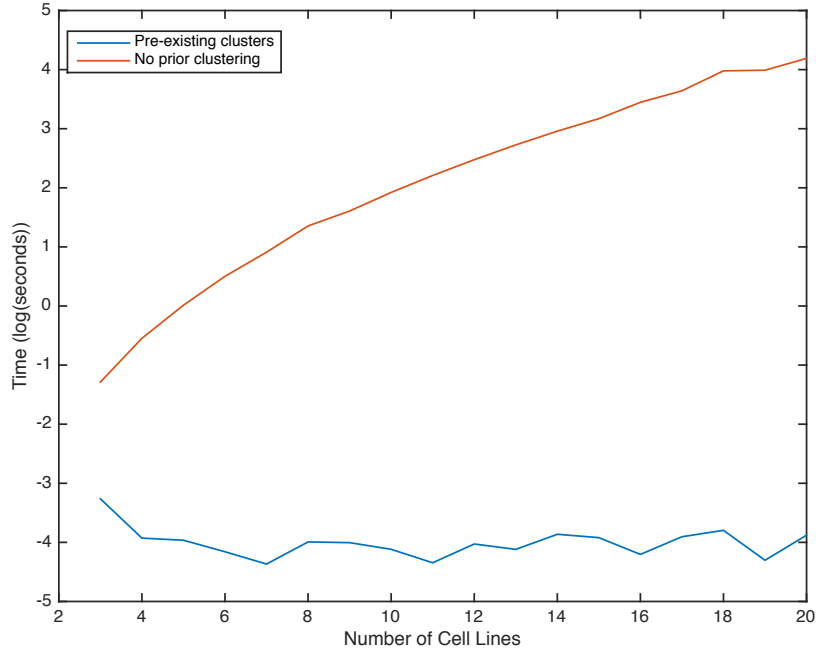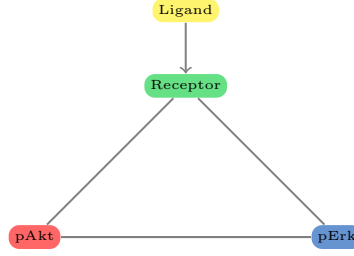
FIG. S7. Computational complexity of two implementations with eight ligands and a varying number of cell lines.

We first consider all possible network topologies with interactions between the variables R, E and A. There are three possibilities for the directed interaction from one variable to another: positive ($\rightarrow$), negative ($\dashv$), or no significant interaction (no arrow). There are six potential directed interactions in the network which give a total of $3^6 = 729$ networks.

Many of these networks can be ruled out. The data shows some response in pERK and pAKT for each stimuli, therefore we require that both pERK and pAKT have at least one arrow coming into each of these (to ensure a response). Given this restriction from the data, we can eliminate many networks, for example a network where the receptor inhibits pAKT and pERK is biologically infeasible. We also do not have dynamic data of the phosphorylated receptor, therefore we cannot distinguish a network topology that has an arrow feeding back from the phopho-form to the receptor; thus we fix the interaction from pERK to R and pAKT to R to none (no arrow). All of these restrictions produce Table III.

Based on these wiring diagrams, we now consider different possible kinetics for each arrow, summarized in the Table, and described in the next subsection.

3. Construction of mechanistic models

We construct systems of ordinary differential equation models to describe interaction dynamics between the receptor (R), pAKT (A) and pERK (E). The equation describing the evolution of the phosphorylated receptor is $dR/dt = \alpha L(R_{tot} - R) - \delta R$. The total amount of receptor, R_{tot}, is estimated from the receptor abundance data. The unphosphorylated receptor is given by $(R_{tot} - R)$. The parameter α determines the rate at which it is phosphorylated by the ligand dose ($L = 1$ or 100 ng/ml). The time evolution of pERK and pAKT can be activated by R. There are two other ordinary differential equations that describe the change in the phosphorylation (in fold change) of A and E with respect to time, whose left hand side is given by dA/dt and dE/dt. Crosstalk between the ERK and

Number of arrows	Total number of networks	Number of networks we consider
All	$3^4 = 81$	15
0	$\binom{4}{0} \cdot 2^0 = 1$	None
1	$\binom{4}{1} \cdot 2^1 = 8$	None
2	$\binom{4}{2} \cdot 2^2 = 24$	3
3	$\binom{4}{3} \cdot 2^3 = 32$	8
4	$\binom{4}{4} \cdot 2^4 = 16$	4

TABLE III. In total, we consider 15 network topologies: those that are biologically plausible given the data. There are different possible kinetics for each model: mass-action or Michaelis-Menten. This gives $15 \times 2 = 30$ possible models. Furthermore, seven of these include inhibition (⊣) which we model via either: blocking or removal (described in the text). Accounting for all our kinetic models gives a total of 44 models.

AKT pathways is encoded by interactions between pERK and pAKT, which can either activate or inhibit the other pathway. Activation terms are modeled using either mass action or Michaelis-Menten kinetics. We consider two types of inhibition: blocking through a saturation term or through a removal term using mass action kinetics.

Without data from receptor dynamics, we write the change in phosphorylated receptor (in arbitrary units) as a function of time as:

$$\frac{\mathrm{dR}}{\mathrm{dt}} = \alpha L(R_{tot} - R) - \delta R,$$

where $\alpha L(R_{tot} - R)$ describes the fraction of non-phosphorylated receptor that becomes phosphorylated at some rate proportional to the ligand L, and δ is the rate at which R is de-phosphorylated.

The other two equations, dA/dt and dE/dt, describe the change in the phosphorylation (in fold change) of A and E with respect to time. These equations change based on the assumed interactions, each different set of equations describes a different mechanistic model.

We assume that activation is either via mass-action kinetics or Michaelis-Menten kinetics, and that inhibition is either via removal or blocking. For example, if we consider the model $R \to E \to A, \quad R \dashv A$, we can write this as:

$$\frac{\mathrm{dR}}{\mathrm{dt}} = \alpha L(R_{tot} - R) - \delta R, \tag{C1}$$

$$\frac{\mathrm{dE}}{\mathrm{dt}} = \frac{k_1 R}{K_{m1} + R} - \delta E, \tag{C2}$$

$$\frac{\mathrm{dA}}{\mathrm{dt}} = \frac{k_2 E}{K_{m2} + E} - k_3 AR - \delta A, \tag{C3}$$

where the blue denotes the Michaelis-Menten term (ignoring the blue is mass-action), and the red term $k_3 AE$ describes inhibition of A as a removal interaction. However, when A is inhibiting by blocking, now the red term is written in a the following form:

$$\frac{\mathrm{dR}}{\mathrm{dt}} = \alpha L(R_{tot} - R) - \delta R, \tag{C4}$$

$$\frac{\mathrm{dE}}{\mathrm{dt}} = \frac{k_1 R}{K_{m1} + R} - \delta E, \tag{C5}$$

$$\frac{\mathrm{dA}}{\mathrm{dt}} = \left(\frac{k_2 E}{K_{m2} + E} \right) \left(\frac{k_3}{k_3 + R} \right) - \delta A. \tag{C6}$$

We summarize the 44 models analyzed in more detail in Fig. S8.

Appendix D: Model identification

1. Estimating the total abundance of receptor, R_{tot}

Our mechanistic analysis requires us to estimate the total abundance of receptor (both phosphorylated and un-phosphorylated) prior to the addition of any ligand. We do this based on the Receptor Abundance Data from [19,
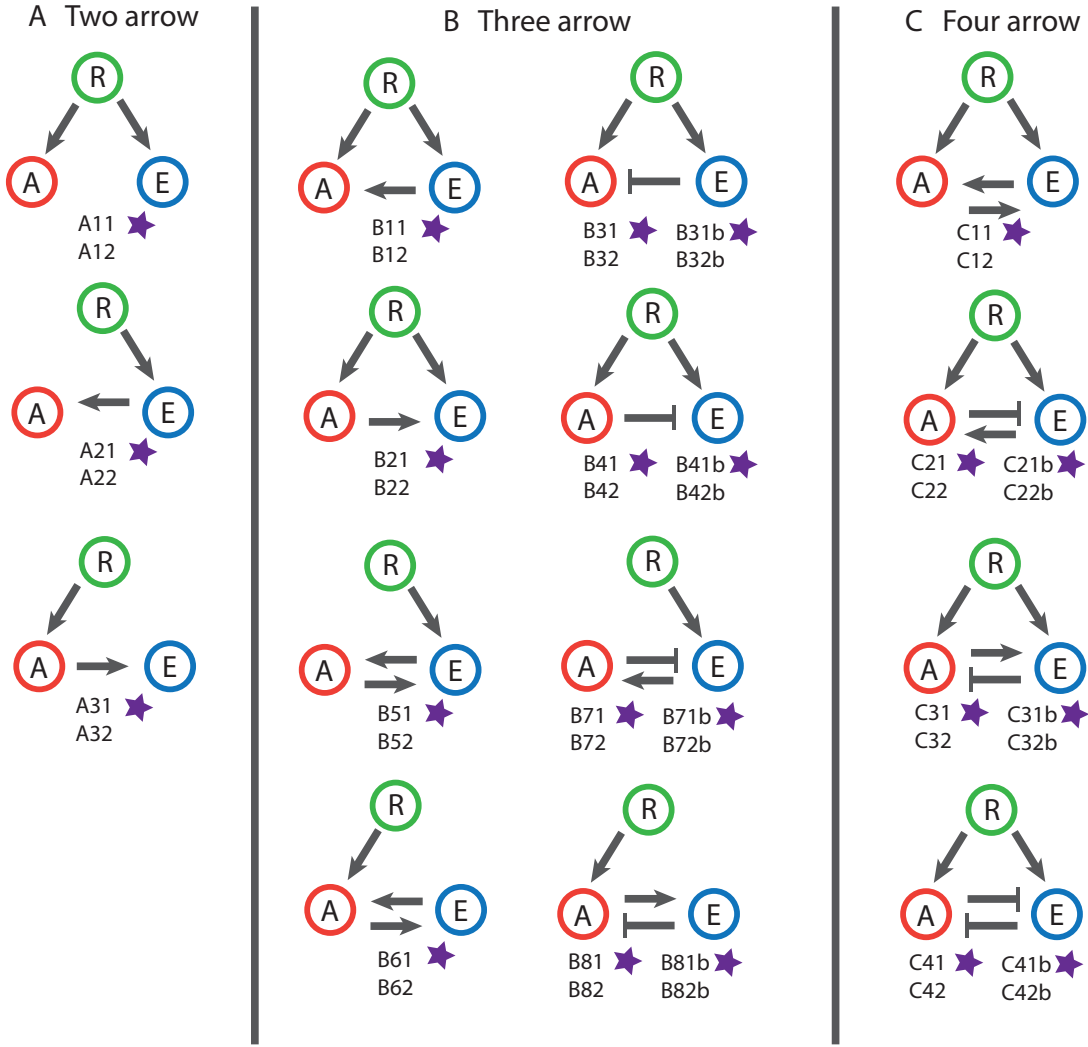
FIG. S8. Mechanistic models of breast cancer cell lines. The name of network model is indexed first by whether it is a two arrow (A), three arrow (B), or four arrow (C) model. The second index assigns a number to each network topology. Each network can be further subdivided to describe a model with mass action kinetics (1 in third index) and a model with Michaelis-Menten kinetics (2 in third index). Any network with inhibition (⊣) has 4 models considered: mass action removal (1 in third index), Michaelis-Menten removal (2 in third index), mass action blocking (1b in third index), or Michaelis-Menten removal (2b in third index). Stars next to the model name are globally structurally identifiable models, all other models are locally structurally identifiable.

Figure 1]. Each ligand has one or more receptors associated to it, as shown in Table IV.

Each experiment involves a cell line and a ligand. For each cell line/ligand pair, we estimate the receptor abundance by averaging the receptor abundances for that cell line, for each of the receptors associated to the ligand. We do this averaging over all receptor, cell line pairs for which we have data.

We then estimate the receptor abundance for a cluster by averaging the values obtained above for each cell line, ligand pair that is present in the cluster.

2. Identifiability analysis

Before estimating model parameters from the data, we determine whether a model is identifiable. Models that are globally identifiable have parameters that are uniquely identifiable under ideal data conditions. Models that are locally identifiability have a finite number of indistinguishable parameter values. Since we only have time-course measurements for A and E, we use a differential algebra approach for eliminating the species R. We test identifiability

Ligand	Associated Receptors
BTC	ErbB1, ErbB4
EGF	ErbB1
EPR	ErbB1, ErbB4
FGF-1	FGFR-1, FGFR-2, FGFR-3, FGFR-4
FGF-2	FGFR-1, FGFR-2, FGFR-3, FGFR-4
HGF	cMET
HRG	ErbB4
IGF-1	IGF1R
IGF-2	IGF1R, IGF2R
INS	InsR
NGF-beta	TrkA
PDGF-BB	PDGFRa, PDGFRb
SCF	c-Kit
VEGF165	VEGFR-1, VEGFR-2, VEGFR-3

TABLE IV. Receptors associated to each ligand

using the algorithm `DAISY` [20]. All of our models are locally, if not globally identifiable given the experimental data. Globally identifiable models are denoted by brown boxes in Fig. S8.

3. Parameter estimation

We estimate parameters using the average time-course for each cluster. The model simulated at a parameter vector $\boldsymbol{\theta}$ gives a vector of model predictions of E and A at dose L and time point t_j and data $\mathcal{D}$ is the set of normalised measurements of $\hat{E}$ and $\hat{A}$. The squared sum of errors of the model with parameter set $\boldsymbol{\theta}$ is:

$$E_{\mathcal{D}}(\boldsymbol{\theta}) = \sum_{L \in \text{ doses}} \sum_{t_j} \left(\hat{A}_{ij}(t_j;\, L) - A_{ij}(t_j;\, \boldsymbol{\theta},\, L) \right)^2$$
$$+ \left(\hat{E}_{ij}(t_j;\, L) - E_{ij}(t_j;\, \boldsymbol{\theta},\, L) \right)^2 .$$

We seek the parameter set $\boldsymbol{\theta}^*$ that minimises the discrepancy between the model and the data:

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, E_{\mathcal{D}}(\boldsymbol{\theta}), \quad \text{subject to } \boldsymbol{\theta}^* \geq \mathbf{0}. \tag{D1}$$

We find $\boldsymbol{\theta}^*$ using the Squeeze-and-Breathe (SB) evolutionary optimisation algorithm [21]. Given an initial estimate of the distribution of the parameter values (a 'prior'). SB generates a large number of parameter sets using Monte-Carlo simulation. These points are used as the starting guess for a local minimisation of $E_{\mathcal{D}}(\boldsymbol{\theta})$ using a derivative free method such as Nelder-Mead [22]. The local minima are ranked and the best points are used to recompute the distribution of the parameters (the 'posterior'). These new posteriors are used as priors in a new iteration of the algorithm, which again looks for new local minima and keeps the best. This process goes on until the error function converges. One of the key advantages of SB is that it can handle situations where little is known about the parameter values by efficiently exploring the parameter space, even venturing to regions outside the original prior (for this reason the posteriors are not true posteriors in the Bayesian sense). Figure S9 shows an example for the type of output produced by SB.
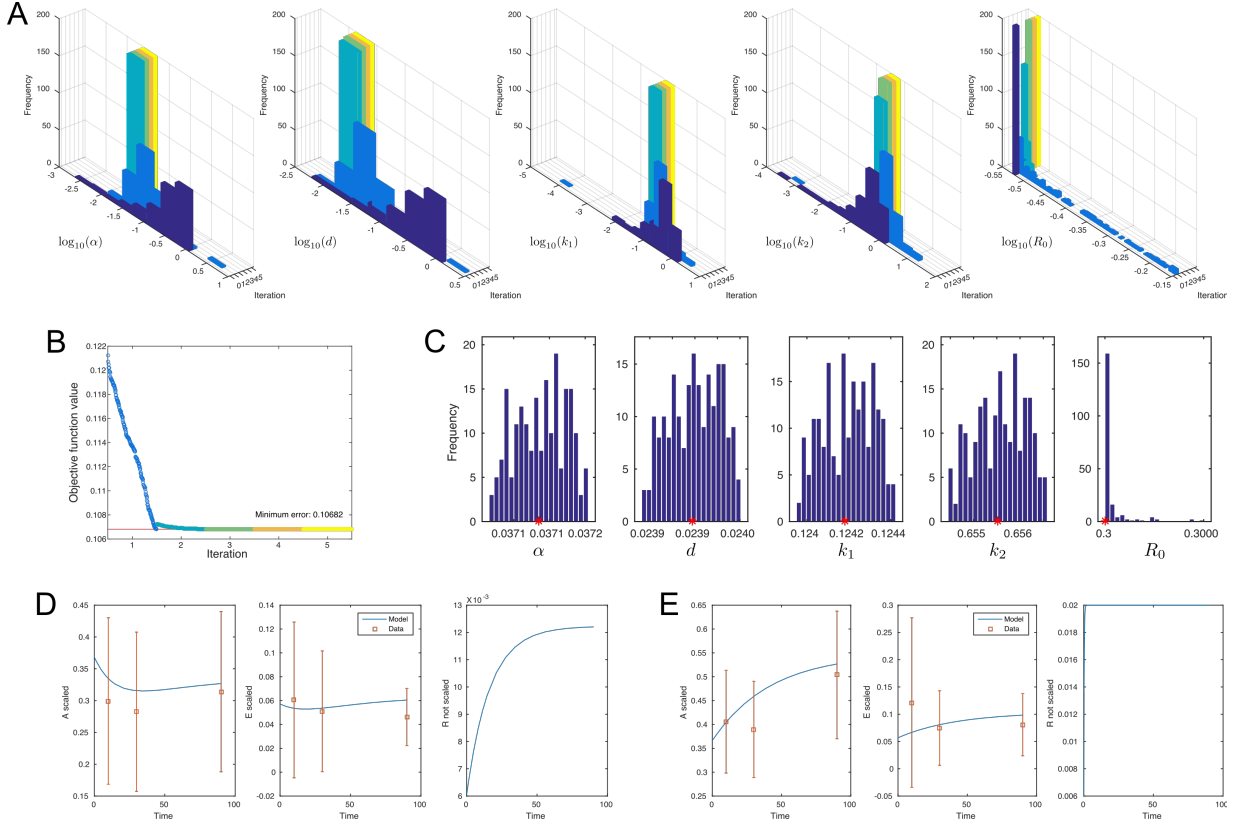
FIG. S9. Example output of the Squeeze and Breathe parameter estimation algorithm. A: Sequence of posteriors for each parameter after each iteration. B: Convergence of the objective function. The plot shows the error of best parameter sets after each iteration. C: Final distribution of the parameters after convergence. The red star marks the value of the parameter with the lowest error. D and E: Time course of the model with the best parameters for one cluster.

4. Model selection

We perform model selection on the 40 models and 5 clusters using the Akaike Information Criterion with a correction for finite sample size (AICc) [23]. For a given model i with p_i parameters,

$$AICc_i = n \ln(nRSS_i) + 2p_i + \frac{2p_i(p_i + 1)}{n - (p_i + 1)},$$

where n is the number of observations and RSS_i is the residual sum of squares of the model. The AICc balances how well the model fits the data with the complexity of the model (the number of parameter values). The lowest $AICc_i$ is the preferable model. Figure S10 shows the AIC scores of the models for each cluster, ranked from first (best) to tenth.

[1] Niepel M, et al. (2014) Analysis of growth factor signaling in genetically diverse breast cancer lines. *BMC Biol* 12:20.

[2] Vangelov B (2014) Ph.D. thesis (Imperial College London).

[3] Beguerisse-Díaz M, Vangelov B, Barahona M (2013) Finding role communities in directed networks using Role-Based Similarity, Markov Stability and the Relaxed Minimum Spanning Tree. *2013 IEEE Global Conference on Signal and Information Processing (GlobalSIP)* pp 937–940.

[4] Beguerisse-Díaz M, Garduño Hernández G, Vangelov B, Yaliraki SN, Barahona M (2014) Interest communities and flow roles in directed networks: the Twitter network of the UK riots. *J R Soc Interface* 11.

[5] Delvenne JC, Yaliraki S, Barahona M (2010) Stability of graph communities across time scales. *Proc Nat Acad Sci USA* 107:12755–12760.
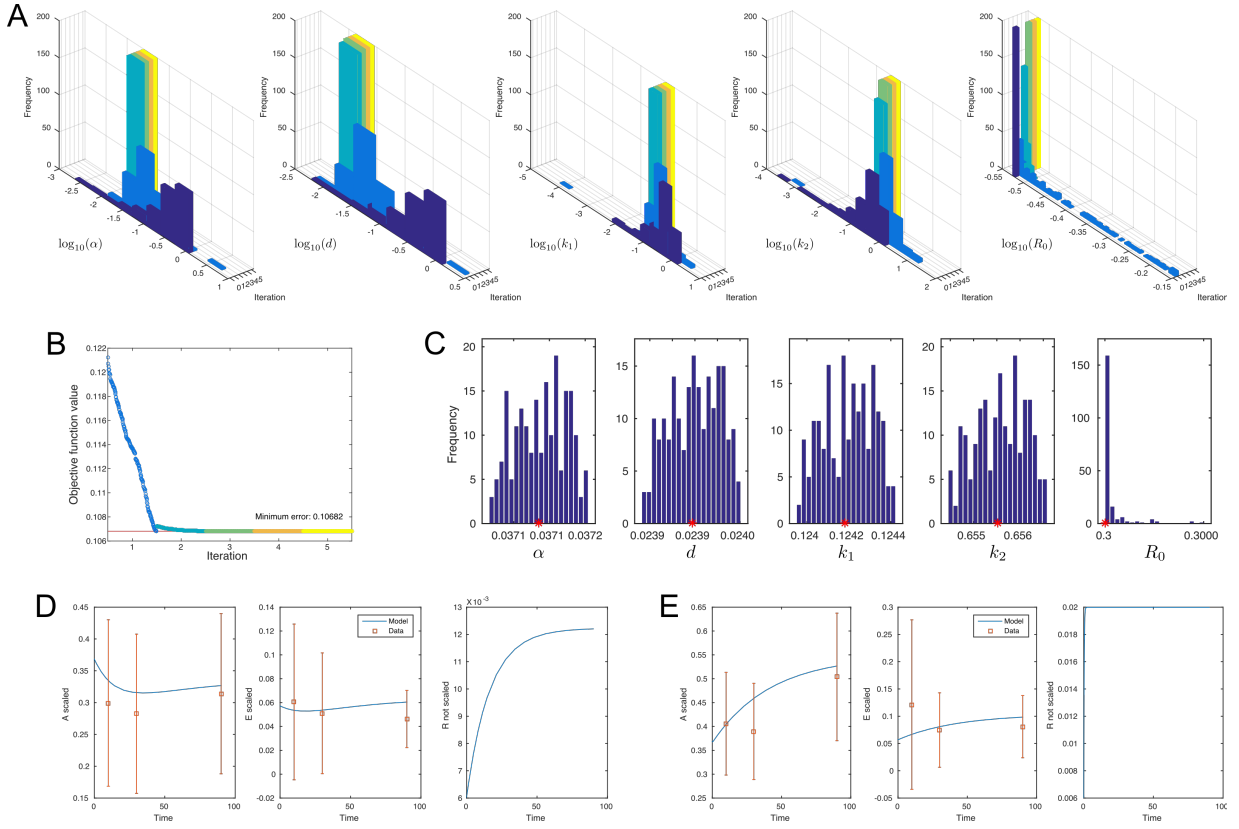
FIG. S10. Values of the AIC for the models ranked first to tenth in each of the five clusers depicted in Fig. 3 of the Main Text. The dashed red line indicates the cutoff for the models in Fig. 3.

[6] Delvenne JC, Schaub MT, Yaliraki SN, Barahona M (2013) in *Dynamics On and Of Complex Networks, Volume 2*, Modeling and Simulation in Science, Engineering and Technology, eds Mukherjee A, Choudhury M, Peruani F, Ganguly N, Mitra B (Springer New York), pp 221–242.

[7] Meilă M (2007) Comparing clusterings—an information based distance. *Journal of Multivariate Analysis* 98:873 – 895.

[8] Heinrich R, Neel BG, Rapoport TA (2002) Mathematical Models of Protein Kinase Signal Transduction. *Molecular Cell* 9:957–970.

[9] Huang CY, Ferrell JE (1996) Ultrasensitivity in the mitogen-activated protein kinase cascade. *Proc Natl Acad Sci USA* 93:10078–10083.

[10] von Kriegsheim A, et al. (2009) Cell fate decisions are specified by the dynamic ERK interactome. *Nature* 11:1458–1464.

[11] Marshall CJ (1995) Specificity of receptor tyrosine kinase signaling: transient versus sustained extracellular signal-regulated kinase activation. *Cell* pp 179–185.

[12] Purvis JE, Lahav G (2013) Encoding and Decoding Cellular Information through Signaling Dynamics. *Cell* 152:945–956.

[13] Chen WW, et al. (2009) Input–output behavior of erbb signaling pathways as revealed by a mass action model trained against dynamic data. *Molecular Systems Biology* 5.

[14] Liepe J, Filippi S, Komorowski M, Stumpf MPH (2013) Maximizing the information content of experiments in systems biology. *PLoS Comput Biol* 9:e1002888.

[15] Fujita KA, et al. (2010) Decoupling of receptor and downstream signals in the akt pathway by its low-pass filter characteristics. *Sci Signal* 3:ra56.

[16] Won JK, et al. (2012) The crossregulation between erk and pi3k signaling pathways determines the tumoricidal efficacy of mek inhibitor. *Journal of Molecular Cell Biology* 4:153–163.

[17] Fey D, Croucher DR, Kolch W, Kholodenko BN (2012) Crosstalk and signaling switches in mitogen-activated protein kinase cascades. *Frontiers in physiology* 3:355.

[18] Beguerisse-Díaz M, Desikan R, Barahona M (2016) Linear models of activation cascades: analytical solutions and coarse-graining of delayed signal transduction. *Journal of The Royal Society Interface* 13.

[19] Niepel M, et al. (2013) Profiles of basal and stimulated receptor signaling networks predict drug response in breast cancer lines. *Sci Signal* 6:ra84.

[20] Bellu G, Saccomani MP, Audoly S, D'Angiò L (2007) Daisy: A new software tool to test global identifiability of biological and physical systems. *Computer Methods and Programs in Biomedicine* 88:52–61.

[21] Beguerisse-Díaz M, Wang B, Desikan R, Barahona M (2012) Squeeze-and-breathe evolutionary Monte Carlo optimization with local search acceleration and its application to parameter fitting. *J R Soc Interface* 9:1925–1933.

[22] Nelder JA, Mead R (1965) A simplex method for function minimization. *The computer journal* 7:308–313.

[23] Burnham K, Anderson D (2003) *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (Springer New York).