PNAS RESEARCH ARTICLE | BIOPHYSICS AND COMPUTATIONAL BIOLOGY

Minimal frustration underlies the usefulness of incomplete regulatory network models in biology

Shubham Tripathi[a,b,c,1], David A. Kessler[d,1], and Herbert Levine[b,c,1,2]

Contributed by Herbert Levine; received September 21, 2022; accepted November 16, 2022; reviewed by Gábor Balázsi and Réka Albert

Regulatory networks as large and complex as those implicated in cell-fate choice are expected to exhibit intricate, very high-dimensional dynamics. Cell-fate choice, however, is a macroscopically simple process. Additionally, regulatory network models are almost always incomplete and/or inexact, and do not incorporate all the regulators and interactions that may be involved in cell-fate regulation. In spite of these issues, regulatory network models have proven to be incredibly effective tools for understanding cell-fate choice across contexts and for making useful predictions. Here, we show that minimal frustration—a feature of biological networks across contexts but not of random networks—can compel simple, low-dimensional steady-state behavior even in large and complex networks. Moreover, the steady-state behavior of minimally frustrated networks can be recapitulated by simpler networks such as those lacking many of the nodes and edges and those that treat multiple regulators as one. The present study provides a theoretical explanation for the success of network models in biology and for the challenges in network inference.

cell-fate choice | gene regulatory networks | frustration | network inference | sloppiness

Significance

In spite of their limited accuracy, gene regulatory networks have provided an immensely successful framework for modeling the cell-fate decision-making process. Here, we show that this success is a consequence of the minimal frustration and the resultant low-dimensional dynamics of the underlying biological networks regulating cell-fate choice. While the minimal frustration property allows for the construction of incomplete/inexact network models of varying sizes that can nevertheless predict cellular behavior, this property also makes it exceedingly difficult to infer precise regulatory networks from gene expression profiles. Thus, the present study addresses fundamental questions concerning network models of cell-fate choice.

Biological network models that describe the regulatory relationship between different molecular players or between higher-level biological entities (such as signaling pathways or cell types) have been extremely useful in systems biology (1, 2) to model and understand the features of cell-fate regulation (3). With the advent of high-throughput molecular profiling techniques, network-based models and approaches have become nearly indispensable (4, 5). Identifying features that distinguish biological networks from random networks has been an area of active research. Previous studies have argued that biological networks present a scale-free degree distribution (6, 7), are hierarchically organized (8), and exhibit recurrence of certain patterns called motifs with a higher probability than expected by random chance (9). However, these and other analyses of topological differences have provided little insight into the functional differences between actual biological networks and random networks, i.e. those differences that enable biological networks to effectively regulate cell fates.

Two functional behaviors of biological regulatory networks stand out. First, physics would suggest that even systems with a relatively small number of independent variables are expected to exhibit exceedingly complex behaviors (10, 11). However, cell-fate regulation, successfully modeled using large and complex networks, is a macroscopically simple process (12–18). Different cell fates are characterized by distinct expression patterns or activity levels of sets of genes (including transcription factors, micro-RNAs, etc.) (19). The typical approach to modeling the establishment of distinct cell fates is to simulate the dynamics of a regulatory network using a methodology of choice (ordinary differential equation-based modeling or rule-based modeling, among others), identify the steady states of network dynamics, and then map each steady state or each group of similar steady states to a distinct cell fate. While the set of cell types—specific gene expression patterns seen in biology—is fairly limited, dynamical models of the size and complexity of biological regulatory networks should, in general, be capable of exhibiting a far more diverse set of expression patterns at steady state. Is there then a universal feature of regulatory networks in biology that restricts the set of gene expression patterns commonly seen?

Second, nearly all network descriptions of cell-fate regulation involve models that are inexact and/or incomplete—such network models do not incorporate all of the genes involved or all the interactions between the chosen genes and often treat multiple biomolecules as a single regulator. This is a consequence of the limited resolution of current experimental techniques, limited data availability, noise in the collection and interpretation of data from high-throughput experiments, the high context dependence of biological assays, and, in many cases, choices made to simplify the modeling task.

Author contributions: S.T., D.A.K., and H.L. designed research; S.T. performed research; S.T., D.A.K., and H.L. analyzed data; and S.T., D.A.K., and H.L. wrote the paper.

Reviewers: R.A., Pennsylvania State University, and G.B., Stony Brook University.

The authors have no competing interest.

[1]S.T., D.A.K., and H.L. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: h.levine@northeastern.edu.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2216109120/-/DCSupplemental.

Published December 29, 2022.

For example, gene networks of widely different sizes have been used to usefully model the regulation of choice between epithelial and mesenchymal fates (13, 14, 20–22). While none of these network models can claim to be more exact than the others, all can claim to recapitulate the gene expression patterns associated with epithelial and mesenchymal cell fates and to provide useful insights into the regulation of the underlying biological process. The success of these incomplete and inexact regulatory network models raises the following question: Is there a universal feature of regulatory networks in biology that allows us to recapitulate the observed biological behavior and make useful predictions without the need to know and incorporate the exact network structure?

Our previous work (23) answered, in part, the first question. We identified minimal frustration as a key property of biological regulatory networks across contexts and showed, within a Boolean modeling framework (14), that biological networks exhibit certain steady states with exceptionally low frustration. These states are the ones that are most frequently encountered when simulating network dynamics and correspond to the gene expression patterns seen in biology. Such low-frustration states are not seen in the case of random networks that have the same topological features as the biological network. While minimally frustrated biological networks can still exhibit steady states with nonbiological gene expression patterns, such steady states are rarely dynamically encountered.

In the present study, we extend our analysis to ordinary differential equation-based models of biological regulatory networks. We show that provided the network is minimally frustrated as defined previously (23) (and discussed below), the steady-state network behavior is simple and largely one-dimensional, in spite of the complex and multidimensional nature of the network model. This property underlies the suitability of large biological networks for describing a macroscopically one-dimensional process such as cell-fate regulation. We then go on to answer the second question posed above and show that the behavior modeled by a minimally frustrated network can be recapitulated by much smaller, simpler network models either lacking many of the regulators and interactions present in the original network or combining multiple regulators into single nodes. Thus, the present study builds upon the analysis in ref. 23 to establish minimal frustration as a key feature of biological regulatory networks and helps explain the success of necessarily incomplete systems biology models in modeling cell-fate regulation.

Modeling Regulatory Dynamics

Specifying regulatory networks—A regulatory network involved in cell-fate regulation can be specified using a directed graph. A node in such a graph may correspond to a transcription factor, a micro-RNA, an epigenetic modifier, or any other regulatory factor. Each directed edge in the graph is signed—either activating or inhibiting—depending on the type of the regulatory relationship between the regulators. Mathematically, a regulatory network with N nodes can be described with an $N \times N$ connection matrix J such that $J_{ij} = +1$ if the edge $i \leftarrow j$ is activating and $J_{ij} = -1$ if the edge is inhibitory; $J_{ij} = 0$ if there is no edge from j to i. Provided that the rules governing how the different inputs to a node combine are available, the network dynamics may be simulated either within a discrete modeling framework, (a Boolean framework being the most commonly used (24)), or a continuous framework involving ordinary differential equations (ODEs).

Boolean modeling and definition of frustration— In a Boolean modeling framework, the state of an N-node network is specified by a sequence of N binary variables $\{s_1, s_2, ..., s_N\}$, with $s_i = +1$ if the regulatory species represented by node i is active and/or highly expressed, and $s_i = -1$ otherwise. The various inputs to a given node, as described by the connection matrix J, may combine additively or via more complex logic-based rules. Here, we consider the simple case wherein the inputs to any given node combine additively and independently (14). In such a scenario, the discrete-time network dynamics are given as (23)

$$s_i(t+1) = \begin{cases} +1 & \sum_j J_{ij}s_j > 0, \\ -1 & \text{if } \sum_j J_{ij}s_j < 0, \\ s_i(t) & \sum_j J_{ij}s_j = 0. \end{cases} \quad [1]$$

The network state is updated in an asynchronous fashion: at any given point in time, a node is chosen at random and its state updated using Eq. **1**. Note that simulating the dynamics of biological regulation using Eq. **1** requires only the network connection matrix J—there are no other parameters involved. One can identify the stable states of such dynamics as any network state $\{s_i\}$ wherein $s_i(t+1) = s_i(t)$ for all i. For every network state, one can define frustration as the fraction of network edges that are not satisfied in that state, i.e., the fraction of edges for which $J_{ij}s_is_j < 0$ (23, 25). In the case wherein the inputs to a given node combine via logic-based rules, frustration of a network state may be similarly defined. However, the precise mathematical definition is more complex in such a scenario (see ref. 23).

Throughout this manuscript, we refer to a network as being minimally frustrated if, within the Boolean modeling framework described above, the network exhibits certain steady states with frustration significantly lower than that of the steady states exhibited by random networks with similar topological features (i.e., random networks with the same number of nodes and edges, the same number of activating and inhibitory edges, and the same in-degree and out-degree for each node). Such random networks can be generated from the original biological network by repeatedly choosing a pair of network edges at random and swapping their targets (*SI Appendix*, section S1*A*). The random networks thus obtained may be more or less modular, as quantified by the directed Louvain modularity (26), than the corresponding biological network (*SI Appendix*, Fig. S1).

ODE-based modeling—In an ODE-based modeling framework, the regulatory network state is described by a continuous N-dimensional vector $\{y_1, y_2, ..., y_N\}$, where y_i describes the expression or activity level of the regulator represented by node i. Given a connection matrix J, the network dynamics (in continuous time) can be described using a set of ordinary differential equations of the form (27)

$$\frac{dy_i}{dt} = g_i \prod_{j, J_{ij} \neq 0} H^S(y_j, \lambda_{ij}, \Theta_{ij}, n_{ij}) - k_i y_i. \quad [2]$$

Here, H^S is the shifted Hill function: $H^S(y_j, \lambda_{ij}, \Theta_{ij}, n_{ij}) = \lambda_{ij} + (1 - \lambda_{ij})\frac{1}{1+(y_j/\Theta_{ij})^{n_{ij}}}$. Note that $\lambda_{ij} > 1$ if the edge $i \leftarrow j$ is activating (i.e., if $J_{ij} = +1$) and $\lambda_{ij} < 1$ if the edge $i \leftarrow j$ is inhibitory (i.e., if $J_{ij} = -1$). Once again, we assume that the inputs to any given node act independently of one another. The system of ODEs in Eq. **2** associates two kinetic parameters with each network node: g_i, the production rate, and k_i, the degradation rate of the regulator represented by node i. Three kinetic parameters are associated with each

network edge: λ_{ij}, the maximum fold change in the production rate of node i that node j can cause, Θ_{ij}, the threshold parameter of the Hill function, and n_{ij}, the Hill coefficient. The system of ODEs in Eq. **2** describes a general setup to model the dynamics of a system of regulatory nodes that can activate or inhibit one another. Such a system can be defined for any given connection matrix J. A more specialized setup with different equations explicitly modeling different modes of biological regulation (e.g., transcriptional regulation, microRNA-mediated regulation, ubiquitination-mediated regulation, etc.) may be chosen to model specific biological systems of interest.

For a given network connection matrix J, the dynamics in an ODE-based framework will, of course, depend on the choice of the kinetic parameters involved in Eq. **2**. The choice of an appropriate parameter set will vary with the biological context and, in general, can be exceedingly difficult. Here, we analyze generic, statistical features of the dynamics for a fixed connection matrix J and an ensemble of kinetic parameter sets generated using the random circuit perturbation (RACIPE) approach (27) (*SI Appendix*, section S1*B*). RACIPE generates an ensemble of kinetic parameter sets in a systematic fashion such that the ensemble is representative of all biologically relevant possibilities. This approach ensures that our analysis is not restricted to the dynamical behavior under a fixed parameter set fitted to some given (arbitrarily chosen) experimental context. More importantly, it allows us to capture the heterogeneity in dynamical behavior that is inherent in biological systems. In fact, each parameter set in the ensemble generated by RACIPE can be interpreted as modeling a different cell in a population, with the variation in the ensemble capturing the cell-to-cell variation in a population. RACIPE can capture the variation arising from intrinsic (different genetic and epigenetic backgrounds) as well as extrinsic (different signaling environments) sources. Note that given the deterministic dynamics for each parameter set generated by RACIPE and the focus on steady-state behavior, the present study does not address how biological and random networks behave under stochastic gene expression noise or in response to fast environmental fluctuations.

Results

Steady-State Dynamics of Biological Regulatory Networks Are Simple. We analyzed features of the set of steady states exhibited by multiple biological regulatory networks taken from the literature (15, 16, 22) for an ensemble of kinetic parameter sets (as described in *SI Appendix*, section S1*B*) and compared these features with those obtained for random networks with similar topological features. Fig. 1 *A*–*C* shows that in the case of biological networks, most of the variation in the steady states is one-dimensional—along the first principal component. This suggests that while these networks are complex—they involve many biomolecular regulators and numerous interactions among them—their behavior at steady state is simple and can be sufficiently described by a single order parameter (e.g., the first principal component). There is no need to specify the expression/activity levels of all of the network nodes to describe the network steady state. In contrast, in the case of random networks, the behavior at steady state is much more complex: a much smaller fraction of the total variance in the set of steady states is captured by the first principal component than in the case of the biological network sharing similar topological features. Thus, describing the steady state in the case of random networks would require specifying the expression/activity levels of many or all of the network nodes, and restricting the description to the first (or even the first few) principal component(s) would be uninformative (*SI Appendix*, Fig. S2). Note that RACIPE was run ab initio for each biological network and each random network instance. Thus, the ensemble of kinetic parameters for which the network behavior is simulated is distinct in each case.

Analyzing the underlying structure in the set of steady states obtained in different cases, we find that the distribution of the first principal component in the case of the biological networks analyzed here is largely bimodal (Fig. 1 *D*–*F*). This indicates that the steady states obtained for the ensemble of parameter sets in the biological case cluster into two distinct groups. This observation is confirmed visually by hierarchical clustering of the set of steady states obtained (*Top-Left* plot in Fig. 1 *G*–*I*), wherein we see two sets of steady states with distinct activity patterns of the network nodes (also, see *SI Appendix*, Fig. S3). No such discernible pattern can be seen in the case of random networks (plots other than the *Top-Left* one in Fig. 1 *G*–*I*). The readily evident clustering of the steady states into two distinct groups (Fig. 1 *D*–*I*) indicates that most of the steady states of these networks can be mapped to one of two phenotypic states with distinct gene expression patterns. This is consistent with the role of these networks in establishing two distinct cell fates: epithelial cells and mesenchymal cells in the case of the epithelial–mesenchymal transition (EMT) network (22), neuroendocrine cells and mesenchymal cells in the case of the small cell lung cancer (SCLC) network (16), and stem cells and differentiated cells in the case of the pluripotency network (15).

Recall that the only input to RACIPE is the connection matrix J. The approach does not take any other experimental data as input, generating the ensemble of kinetic parameters in an unbiased fashion so as to capture the range of possible network behaviors. Thus, the structure in the steady states obtained using RACIPE (shown in Fig. 1 *D*–*I*) is an intrinsic property of biological connection matrices that is seen to be absent in random networks.

Fig. 1 demonstrates the capability of the biological networks analyzed here to robustly establish cell types with biological gene expression patterns. Biological networks exhibit this behavior for a broad range of kinetic parameter sets in a manner that is dependent on the network connection matrix J. While these networks can still exhibit certain steady states that cannot be uncontroversially mapped to one of the two groups that correspond to biological phenotypic states (Fig. 1 *D*–*F*) and with expression patterns not seen in canonical cell types, such steady states are infrequently encountered when simulating network dynamics starting from random initial conditions. Such steady states with aberrant expression patterns are seen at a higher frequency in the case of the SCLC network (see Fig. 1 *E* and *H*) as has been noted elsewhere (16). The noncanonical expression patterns corresponding to such aberrant steady states, while suppressed in healthy cells, have been reported in cancer cells (16, 28). Phenotypes with noncanonical expression patterns (and high frustration) have also been associated with transitions between phenotypic states with canonical gene expression patterns (29).

Minimal Frustration Underlies the Simple Steady-State Dynamics of Biological Regulatory Networks. We have previously demonstrated that biological regulatory networks taken from the literature (including the ones analyzed in Fig. 1), within a Boolean modeling framework, exhibit certain steady states with frustration lower than that of steady states exhibited by random networks with similar topological features, i.e., biological regulatory networks are minimally frustrated (23). In the previous

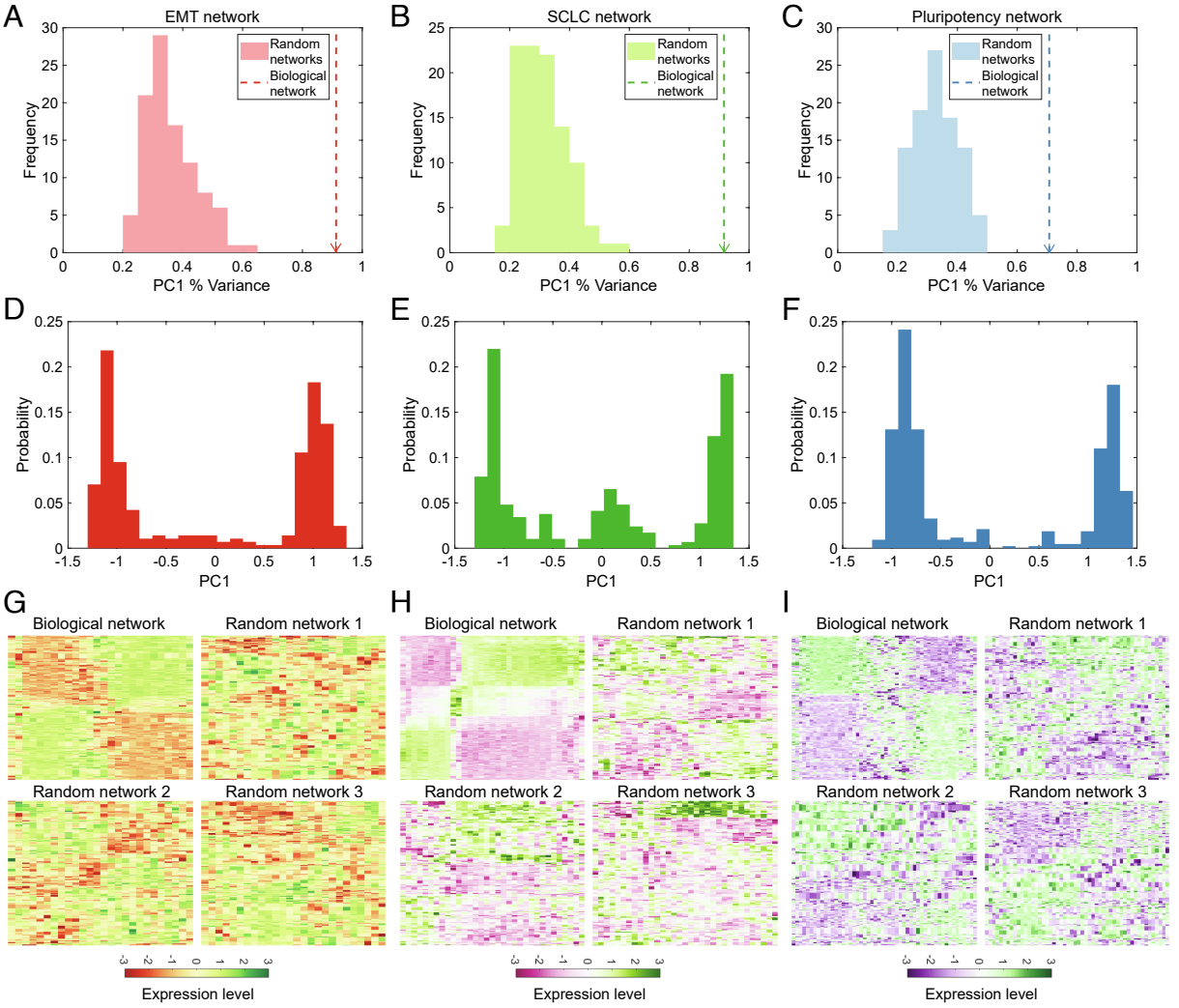

Fig. 1. Steady-state dynamics of complex biological networks are simple and largely one-dimensional. (*A*–*C*) Distribution of the percentage variance explained by the first principal component (PC1 % variance) in the case of random networks (histograms) with the same number of nodes and edges and similar topological features as the biological network (dashed vertical lines with arrows). In each case, PC1 is the principal component that explains the greatest percentage of the variance in the steady states. See *SI Appendix*, Fig. S2 for contribution from the other principal components. (*D*–*F*) Distribution of PC1 projection of the steady states obtained for the three biological networks. See *SI Appendix*, Fig. S3 for the projection of the steady states along PC1 and PC2. (*G*–*I*) Expression levels of the different network nodes in the various steady states obtained for biological and random networks. Different steady states are shown along the rows, while the network nodes are shown along the columns of the heatmaps, with the colors indicating the expression levels. Both rows and columns were hierarchically clustered to obtain the heatmap in each case (*SI Appendix*, section S1C). *Left* column (*A*, *D*, and *G*): epithelial-mesenchymal transition (EMT) network (22); *Middle* column (*B*, *E*, and *H*): small cell lung cancer (SCLC) network (16); right column (*C*, *F*, and *I*): pluripotency network (15).

section *Steady-State Dynamics of Biological Regulatory Networks Are Simple*, we have shown that, within an ODE-based modeling framework, the steady states exhibited by biological networks are simple and largely one-dimensional. In both cases, we argue that the reported biological network behavior underlies the ability of large and complex networks to describe cell-fate regulation. To determine whether the two features—minimal frustration within a Boolean modeling framework and simple, largely one-dimensional steady-state dynamics within an ODE-based modeling framework—are directly related, we simulated the evolution of a population of random networks with the same topological features as the EMT network (22) subject to different selection pressures (see *SI Appendix*, section S1*D* for the detailed methodology). Under selection for networks for which the steady-state dynamics are largely one-dimensional (as quantified by the percentage of variance in the set of steady states explained by the first principal component) (Fig. 2*A*), we obtained networks that were minimally frustrated (Fig. 2*B*). Reciprocally, selection for the low-frustration property (Fig. 2*C*) led to the emergence of networks with largely one-dimensional steady-state dynamics (Fig. 2*D*). Moreover, under selection for low frustration, we obtained networks that exhibited steady-state gene expression patterns very similar to the biological case (compare Fig. 2*F* and the *Top-Left* plot in Fig. 1*G*). It has previously been suggested that a large, complex regulatory network can exhibit low-dimensional gene expression patterns if the network has a modular topology (30). However, we did not see an increase in the modularity of networks in the population while selecting for networks with largely one-dimensional steady-state dynamics (*SI Appendix*, Fig. S4). Emergence of networks with largely one-dimensional steady-state dynamics was also observed for the case of simulations involving growth in network size (emulating increase in the size and complexity of the regulatory network involved in a certain cell-fate choice process over the course of biological evolution) under selection for minimal frustration (*SI Appendix*, Fig. S5).

The behavior in Fig. 2 shows that the minimal frustration property and the property of exhibiting simple, largely one-dimensional steady-state dynamics over an ensemble of parameter

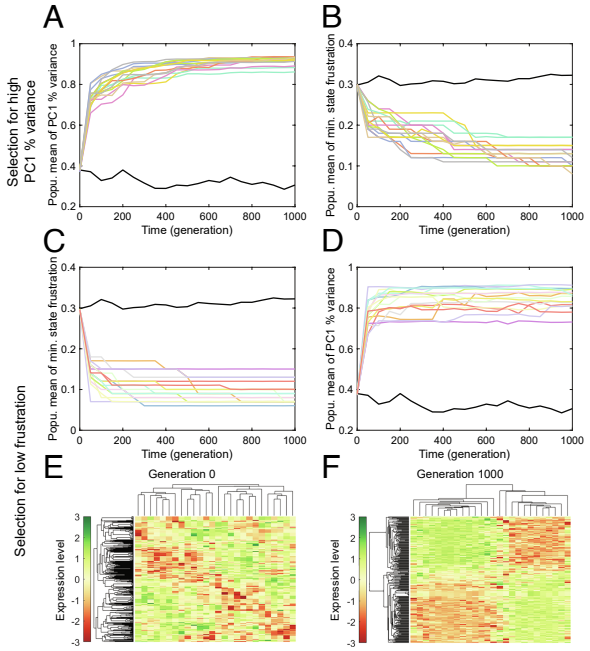

Fig. 2. Selection for networks with simple, one-dimensional steady-state dynamics automatically selects for minimal frustration, and vice versa. Results of an evolution simulation wherein: (*A* and *B*) networks for which a larger fraction of the variance in steady-state behavior can be explained by a single principal component (i.e., networks that exhibit more one-dimensional steady-state dynamics) are selected for at each generation, and (*C*–*F*) networks with minimally frustrated steady states are selected for at each generation. (*E* and *F*) Expression levels of network nodes at steady state for a network randomly chosen from the population at generation 0 (*E*) and at generation 1,000 (*F*). *A* and *C* show the mean population PC1 score, and *B* and *D* show the mean population frustration score (defined in *SI Appendix*, section S1*D*). In *E* and *F*, different steady states are shown along the rows while the network nodes are shown along the columns of the heatmap. Expression levels are indicated by the color (see adjacent color bars). Both rows and columns were hierarchically clustered to obtain the heatmaps. In panels *A*–*D*, the black curve shows the behavior for an evolution simulation in the absence of any selection pressure. The other colors indicate independent simulation runs with selection. Details of the simulation setup are provided in *SI Appendix*, section S1*D* .

sets are, in fact, equivalent—selection for one automatically selects for the other. Since a Boolean model, defined here simply by the connection matrix J, can be built into a corresponding ODE-based model by including a suitable set of kinetic parameters and mathematical expressions, we may conclude that the minimal frustration property within the Boolean framework underlies the simple steady-state dynamics seen in the ODE-based framework. An approach for directly obtaining a Boolean modeling framework starting from an ODE-based model would be helpful for verifying whether simple steady-state dynamics in an ODE-based model can underlie minimal frustration within a Boolean modeling framework. Such an approach will be investigated in a future study.

Simplicity of Steady-State Dynamics Is Preserved Under Node and Edge Deletions. Robustness of functional behavior to genomic and environmental perturbations is a well-known feature of biological systems (31). To determine whether the functional characteristic of biological regulatory networks highlighted here—simple, largely one-dimensional steady-state dynamics—is robust to node and edge deletion, we deleted nodes (Fig. 3*A*) and edges (Fig. 3*B*) in the EMT network (22) one by one (following the approach detailed in *SI Appendix*, section S1*E*), and reported the percentage of variance in the steady states that is explained by the first principal component (corrected for the number of nodes in the network) at each step. Fig. 3 *A* and *B* shows that the steady-state dynamics remain largely one-dimensional even as nodes and edges are successively deleted from the EMT network, and this behavior is maintained over the deletion of a large fraction of nodes and edges. Similar behavior is observed in the case of a minimally frustrated network (Fig. 3 *C* and *D*) obtained at the end of the evolution simulation subject to selection for low frustration shown in Fig. 2*C* as well as for other biological networks (*SI Appendix*, Fig. S6). The shape of the distribution of the first principal component can also withstand the deletion of a large fraction of network nodes and edges (Fig. 3 *E* and *F* and *SI Appendix*, Fig. S7). Note that the change in the variance along the first principal component depends on the order in which the nodes or edges are deleted (compare the red (blue) plots with the pink (light blue) plots in Fig. 3 *A*–*D*), indicating that certain nodes and edges in the network are more important than others in maintaining the simple, one-dimensional network dynamics.

Networks Lacking Multiple Nodes and Edges Can Recapitulate Biological Expression Patterns. Until now, we have shown that in spite of their large size and complexity, minimally frustrated networks such as biological networks taken from the literature exhibit fairly simple, one-dimensional steady-state dynamics. Based on this observation, we hypothesized that "simpler" network models should be capable of recapitulating the steady-state behavior exhibited by a larger, more complex network provided the larger network is minimally frustrated. Here, by

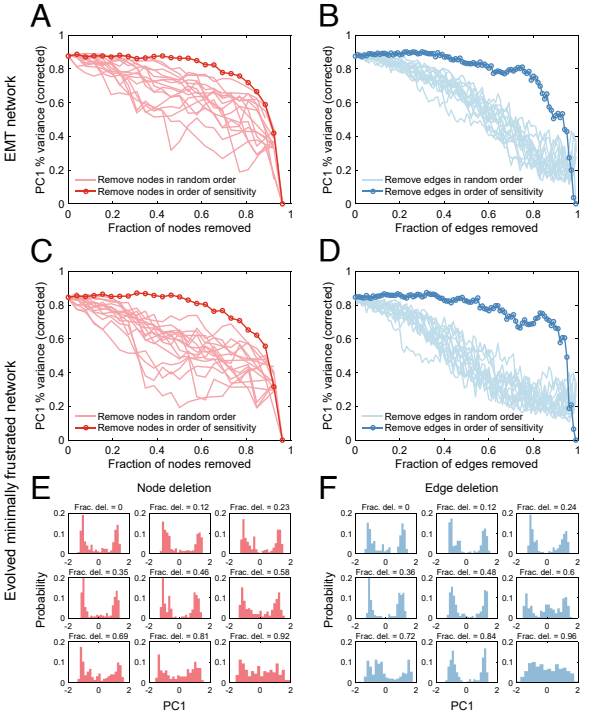

Fig. 3. The simple, one-dimensional steady-state behavior as characterized by a large fraction of the variance in the steady states being explained by the first principal component (PC1) is preserved under node and edge deletion in the case of the 26-node, 100-edge epithelial-mesenchymal (EMT) network (22) (*A* and *B*) as well as in the case of a minimally frustrated network obtained via an evolution simulation (*C*–*F*). (*E* and *F*) Change in the distribution of the first principal component (PC1) projection of the steady states during node (*E*) and edge (*F*) deletion. To generate the curves labeled "Remove nodes/edges in order of sensitivity", we defined "sensitivity" so as to capture the dependence of the obtained steady states on specific model parameters: higher sensitivity implies a larger change in steady states upon varying the model parameters (see *SI Appendix*, section S1*E* for a mathematical definition). In *A*, *C* (red curves), and *E*, the node with the lowest sensitivity was deleted first. In *B*, *D* (blue curves), and in *F*, the edge with the lowest sensitivity was deleted first. See *SI Appendix*, Fig. S7 for a quantitative comparison of the distributions obtained after node or edge deletion with the original distribution.

"simpler," we imply networks lacking multiple nodes and/or edges present in the original network. A different manner of network simplification is addressed in the next section.

In agreement with the above hypothesis, we find that networks lacking numerous nodes and edges present in the EMT network (22) (see *SI Appendix*, section S1*E* for how such networks were obtained) can still recapitulate the pattern of node expression/activity levels exhibited by the full 26-node, 100-edge EMT network (Fig. 4 *A* and *C*). Importantly, these smaller, simpler networks exhibit strikingly similar behavior in response to gene knockouts as the full EMT network (Fig. 4 *B* and *D* and *SI Appendix*, Fig. S8): the change in the distribution of the first principal component upon gene knockout in the simpler networks is qualitatively similar to the change in the case of the original network. Thus, given experimental data on the gene expression profiles seen in cells and even data on the effect of knocking out multiple genes, it is impossible to uniquely identify any one network model as the correct one. Instead, one can employ many useful network models of different sizes and varying complexities to model EMT, each missing a different subset of the nodes and edges present in the original network. Clearly, it is not necessary to know the exact network to recapitulate the overall biological behavior. In fact, the 26-node, 100-edge EMT network considered as the full EMT network in this section is itself not the "correct" network—while it captures many of the features of EMT regulation, it cannot claim to incorporate all the regulators that can affect EMT or even the entire set of interactions among the regulators it does incorporate. Note that useful network models of EMT cannot be simplified beyond a certain limit: exceedingly simple networks cannot adequately recapitulate biological behavior. This can be clearly seen in the heatmaps presented in Fig. 4*C*. Interestingly, this behavior is not limited to the EMT network. *SI Appendix*, Fig. S9 shows that for a minimally frustrated network obtained via the evolution simulation (Fig. 2 *C* and *D*), the steady-state node activity pattern can also be recapitulated by simpler networks with fewer nodes and edges.

Networks that Combine Regulators Can Recapitulate Biological Behavior. In the previous section, we analyzed the steady-state

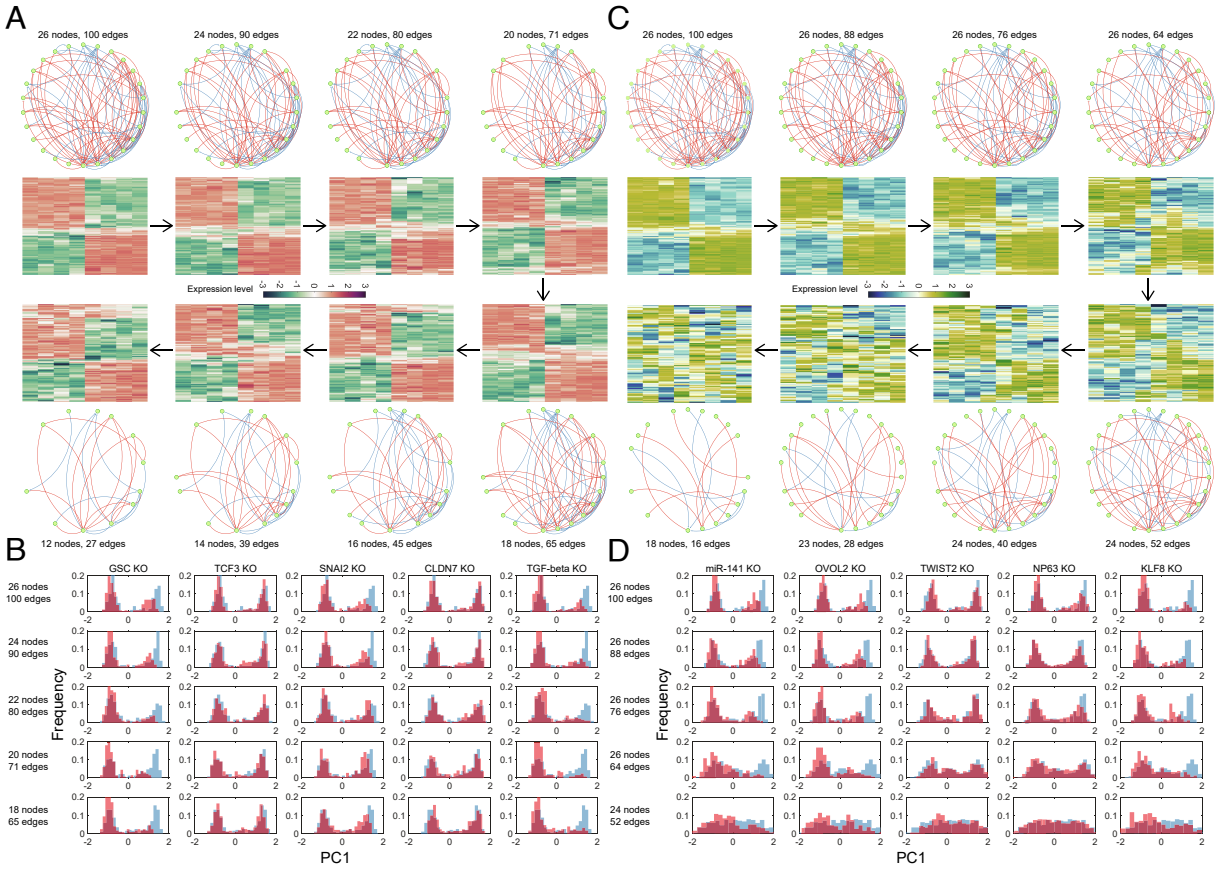

Fig. 4. The pattern of node expression/activity levels at steady state exhibited by the 26-node, 100-edge epithelial-mesenchymal (EMT) network (22) is recapitulated by simpler networks obtained upon node deletion (*A*) or edge deletion (*C*). Each heatmap shows the expression levels (indicated by the color) of the same eight nodes of interest across the different steady states obtained for each network. Different steady states are shown along the rows, while the network nodes are shown along the columns of the heatmaps. The heatmaps shown here were generated by hierarchically clustering the rows (i.e., the steady states), while the network nodes (i.e., the columns) are shown in the same order. The simpler networks obtained are shown alongside the corresponding heatmaps. (*B* and *D*) The simpler networks obtained upon node or edge deletion recapitulate the response of the larger, original EMT network to multiple gene knockouts. In each plot shown in *B* and *D*, the blue histogram shows the distribution of the first principal component in the control case, while the pink histogram shows the distribution obtained upon gene knockout. The principal component analysis was carried out for the eight nodes of interest. Each row in panels *B* and *D* shows the behavior for a fixed network (whose size is indicated in the figure), while each column shows the response to a given gene knockout (KO). See *SI Appendix*, Fig. S8 for a quantitative comparison of the distributions obtained after gene knockout with the distributions in the control case. The simpler networks analyzed here were obtained by successively deleting randomly chosen nodes (*A* and *B*) or edges (*C* and *D*) while ensuring that the eight nodes of interest are retained in the simpler networks. The same eight nodes were the nodes of interest in panels *A*–*D*: four epithelial state markers (CDH1, miR-200b, miR-200c, and miR-34a) and four mesenchymal state markers (VIM, ZEB1, SNAI1, and TWIST1). In each of the networks in *A* and *C*, nodes with the same name are shown in the same position. Descriptions of all networks shown are available online (*SI Appendix*, section S1*H*).

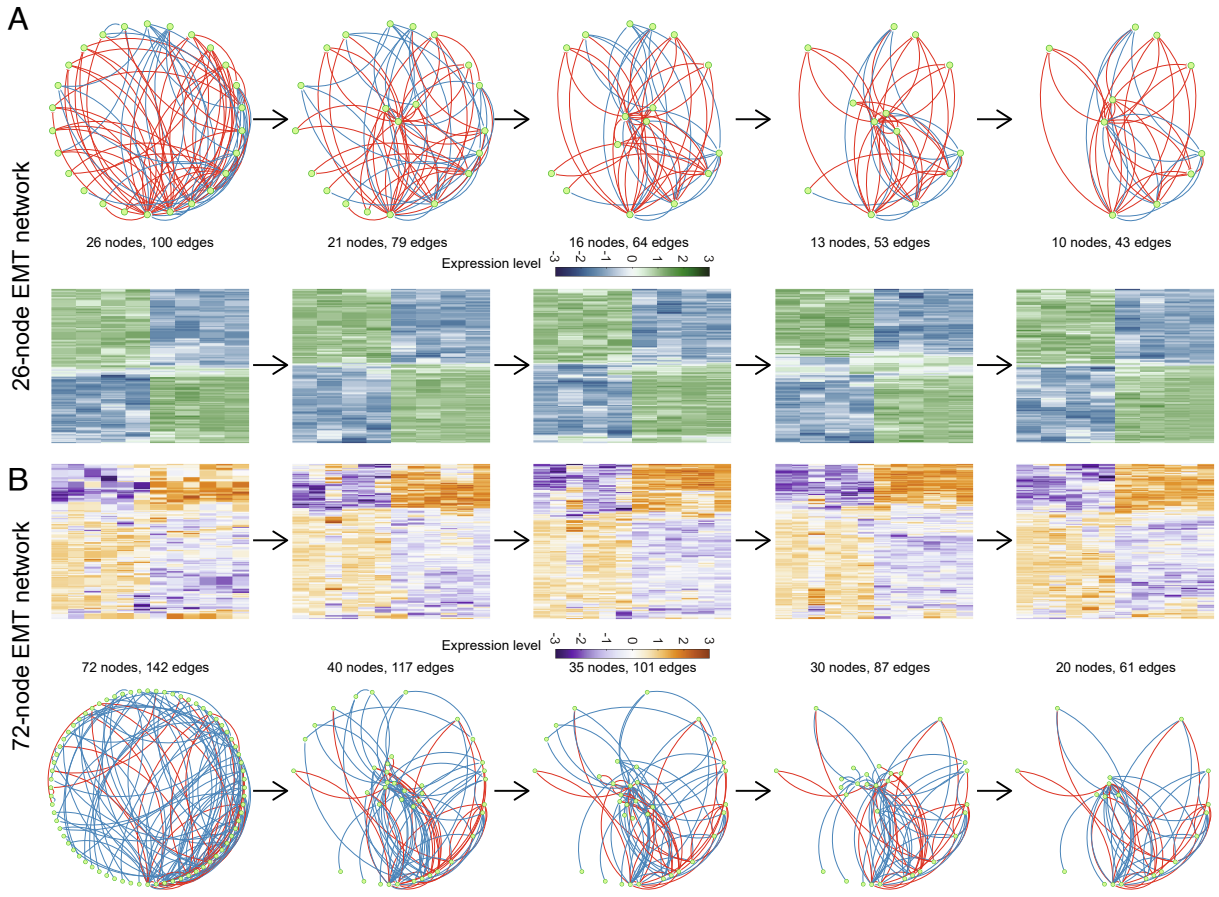

Fig. 5. (*A, Bottom*) Simpler networks obtained by repeatedly applying the coarse-graining procedure described in *SI Appendix*, section S1*F* to the 26-node, 100-edge epithelial–mesenchymal (EMT) network (22) recapitulate the steady-state expression patterns exhibited by the larger, original network. Each heatmap shows the expression levels (indicated by the color) of the same eight nodes of interest (same as the nodes of interest in Fig. 4) across the different steady states obtained for each network. (*A, Top*) The original 26-node, 100-edge EMT network is shown (*Left*) along with the simpler networks obtained via the coarse-graining procedure. (*B*, *Top*) Simpler networks obtained via the coarse-graining procedure applied to the 72-node, 142-edge EMT network (14) exhibit steady-state expression patterns similar to the pattern obtained for the larger, original network. Each heatmap shows the expression levels (indicated by the color) of the same twelve nodes of interest across the different steady states obtained for each network. The twelve nodes of interest were Ecadherin, KLF4, cateninmemb, miR200, GSK3, TrCP, cateninnuc, ZEB1, SNAI1, TWIST1, ZEB2, and FOXC2. (*B, Bottom*) The original 72-node, 142-edge EMT network is shown (*Left*) along with the simpler networks obtained via the coarse-graining procedure. In both *A* and *B*, the coarse-graining procedure was applied while ensuring that the nodes of interest are not combined with any other network node at any step. In all heatmaps, different steady states are shown along the rows while the network nodes are shown along the columns of the heatmaps. The heatmaps shown here were generated by hierarchically clustering the rows (i.e., the steady states). Each heatmap shows the network nodes (i.e., the columns) in the same order. In each of the networks in *A* and *B*, nodes with the same name are shown in the same position along the periphery, while nodes that represent multiple nodes combined into one are shown toward the center. Descriptions of all networks shown are available online (*SI Appendix*, section S1*H*.

behavior of simpler networks that lack many of the nodes and edges present in the larger, original network. Here, we consider simpler networks obtained by combining sets of nodes in the original network and treating each set as a single regulator. Such a network simplification was motivated by the observation that, in the literature, cell types and cell-state transitions have been described both in terms of the expression levels of individual genes as well as in terms of the overall activity levels of different pathways (often comprising numerous genes) (32, 33). We first developed a systematic procedure to combine sets of nodes into a single regulator (described in *SI Appendix*, section S1*F*). *SI Appendix*, Fig. S10 shows the sequence of networks obtained by repeatedly applying this "coarse-graining" procedure to the EMT network (22). Note that the network obtained at each step has both fewer nodes and edges and is thus simpler than the network in the previous step.

Consistent with the behavior in Fig. 4, we find that the simpler networks obtained by applying the above-mentioned coarse-graining procedure to the 26-node, 100-edge EMT network are able to recapitulate the steady-state expression patterns exhibited by the original network (Fig. 5*A*). *SI Appendix*, Fig. S11 shows two of the simplified networks obtained via the coarse-graining procedure. Note that the rightmost network shown therein groups together specific epithelial factors (miR-101, miR-200a, miR-141, CLDN7, OVOL2, GRHL2, miR-30c, and miR-9) into a single regulatory node and specific mesenchymal factors (FOXC2, ZEB2, SNAI2, TGF-beta, TWIST2, GSC, KLF8, TCF3, miR-205, and NP63) into another regulatory node, instead of treating each of these factors separately. This simplified network exhibits a steady-state expression pattern very similar to the one exhibited by the original, larger EMT network (compare the first and last panels in Fig. 5*A*). We also applied our coarse-graining procedure to a larger EMT network taken from the literature, one with 72 nodes and 142 edges (14). Like other biological networks analyzed in the present study, this network is also minimally frustrated (23). Once again, the simpler networks obtained by coarse graining reproduced the steady-state expression patterns of the key EMT-related genes (Fig. 5*B*) exhibited by the larger, original network. The grouping of different nodes in the simpler networks

is biologically interpretable (see the network on the right in *SI Appendix*, Fig. S12): genes that are in the same pathway are grouped together. The hypoxia stimulus, HIF1 gene, and LOXL gene are grouped to form a node representing the hypoxia pathway. Various factors involved in growth-factor signaling, including platelet-derived growth factor (PDGF) and receptor (PDGFR), epithelial growth factor (EGF) and receptor (EGFR), insulin-like growth factor (IGF1) and receptor (IGF1R), and fibroblast growth factor (FGF) and receptor (FGFR), are also grouped to form a single node representing the many signaling pathways known to drive EMT. The Notch pathway genes and factors (NOTCH, NOTCH intracellular domain (NOTCHic), DELTA, Jagged, and HEY1) are grouped together and so are the molecular players involved in the Wnt signaling pathway (TCF/LEF, Wnt, Frizzled, and AXIN2). We also applied the coarse-graining procedure to the SCLC network (16): the simpler networks thus obtained are shown in *SI Appendix*, Fig. S13. Once again, we observe that the simpler networks obtained recapitulate the gene expression patterns exhibited by the larger, original network.

SI Appendix, Fig. S9 shows that the ability of simpler, coarse-grained networks to recapitulate the steady-state gene expression patterns exhibited by the larger network is not limited to the case of biological networks taken from the literature but extends to minimally frustrated networks obtained via the evolution simulation in Fig. 2 *C*–*F*.

A Data-Driven Example. So far, we have analyzed the behavior of previously published biological networks that were constructed by aggregating information from the literature and from biological databases using a variety of methods (14–16, 22). For our final example, we turn to a network constructed directly from gene expression data.

Specifically, we study the regulation of MYC-pathway activation in breast tumors. Terunuma et al. obtained the bulk gene expression profiles from 61 breast tumor samples and identified a subset of tumors that showed a MYC-activated phenotypic state (34). This phenotype was associated with elevated levels of the oncometabolite 2-hydroxyglutarate, DNA hypermethylation, and poor disease prognosis. We used the gene expression data from this study as an input to the GRNBoost2 algorithm (35) and obtained a regulatory network that may be involved in the regulation of tumor cell-fate choice between high MYC-activation and low MYC-activation states (see *SI Appendix*, section S1*G* for the detailed methodology). We chose GRNBoost2 for network inference here since it is a popular and efficient algorithm, and a recent benchmarking study that compared multiple network inference methods recommended GRNBoost2 as a method of choice (36). Moreover, we believe that the choice of the network inference method is unlikely to change the conclusion here. The inferred regulatory network returned by GRNBoost2 consisted of 138 nodes and 451 edges. We have previously shown that such an inferred network is minimally frustrated, just like the biological networks taken from the literature (23). Here, we report that the inferred regulatory network modeled using ODEs exhibits steady states that vary mostly along the first principal component (Fig. 6*B*; red dot), once again consistent with the behavior seen in the case of biological networks taken from the literature (Fig. 1). This was not true for the case of networks inferred by using randomly shuffled gene expression patterns as inputs to our network inference methodology (Fig. 6*B*; blue dots). Importantly, simulation of dynamics of the inferred network using RACIPE (*SI Appendix*, section S1*B*) recapitulated the gene expression patterns seen in patient tumor samples (Fig. 6 *A* and *D*). As in the case of biological networks taken from the literature, the steady-state gene expression patterns exhibited by the large inferred network could be recapitulated by simpler networks obtained by node deletion, edge deletion, or via the coarse-graining procedure (Fig. 6*D*). Interestingly, the simpler networks obtained via the coarse-graining procedure better preserve the steady-state expression patterns of the nodes of interest than do the simpler networks obtained via node or edge deletion.

The low-dimensional dynamics in Fig. 6 are not altogether surprising given that the gene expression profiles of breast tumor samples in the dataset used for network inference are inherently low dimensional (evident from Fig. 6*A*). Thus, it would be reasonable that the inferred gene regulatory network must also have low-dimensional dynamics. The analysis in Fig. 6 does directly indicate that there was no role played by any possible systematic bias in the methodologies used to construct the other biological networks analyzed in the present study (the 26-node (22) and the 72-node (14) EMT networks, the SCLC network (16), and the pluripotency network (15)) that give rise to our demonstrated results. Instead, networks inferred from biological data in a fully automated manner with minimal human intervention (such as the MYC pathway activation network in Fig. 6) also exhibit similar low-dimensional dynamical behavior.

Additionally, the set of genes whose expression levels were used as inputs to GRNBoost2 to obtain the network analyzed in Fig. 6 included both genes that are up-regulated as well as those that are down-regulated in response to MYC activation (37) (see *SI Appendix*, section S1*G* for details). The fact that the inferred network involving these genes has features of minimally frustrated networks would suggest that the regulatory connections among the direct and indirect targets of MYC are organized so as to have low frustration: genes activated by MYC typically activate other genes activated by MYC while repressing MYC targets that are inhibited by MYC induction. Conversely, genes repressed by MYC typically activate other genes repressed by MYC while repressing MYC targets that are activated upon MYC induction. This observation further supports our hypothesis that gene regulatory networks in biology are minimally frustrated.

Discussion

Large, complex networks involving numerous molecular players and various regulatory relationships among them are frequently used to describe and model cell-fate choice, especially as high-throughput assays have become increasingly commonplace. From a dynamical systems perspective, such networks would be expected to exhibit complex behaviors in very high-dimensional spaces (10, 11). However, in most cases, cell-fate choice appears to be a macroscopically simple and low-dimensional process (12) and where a given cell falls on the spectrum between distinct cell fates can be specified with a single order parameter. For example, pseudotime, which is essentially a one-dimensional order parameter, has been widely used to order cells along a cell-fate trajectory based on their gene expression patterns (38, 39). Multiple studies have described metrics, a single number in each case, to specify where a sample lies on the epithelial-mesenchymal spectrum on the basis of the gene expression in that sample (40–42). Another one-dimensional metric has been developed to assess the stemness of leukemia samples (43). In the present study, we have shown that large, complex networks can exhibit largely one-dimensional steady-state behavior provided the network is minimally frustrated (as defined in ref. 23). Importantly, we demonstrate that minimally frustrated networks can be simplified—one can obtain multiple smaller networks that can recapitulate the behavior exhibited by the larger network. Since,

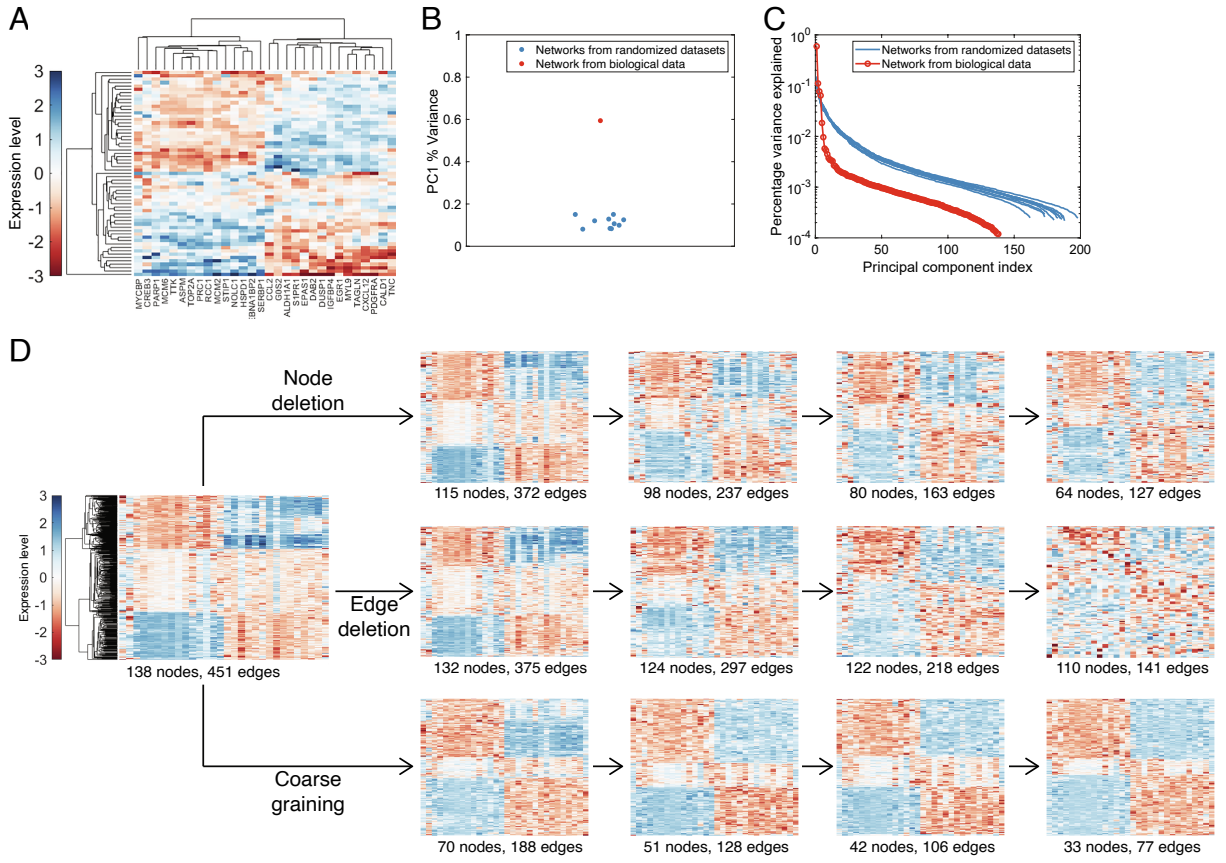

Fig. 6. A network inferred from a gene expression dataset (34) exhibits the same behavior as biological networks taken from the literature and other minimally frustrated networks. (*A*) Expression of 30 MYC-associated genes in 61 breast tumor samples obtained by Terunuma et al. (34). Different tumor samples are shown along the rows while the genes are shown along the columns. Color indicates the expression level. (*B*) Fraction of variance in the steady states explained by the first principal component (PC1) in the case of the network inferred from breast tumor samples (34) (red dot) and from randomized gene expression profiles (blue dots). Principal components are sorted in decreasing order of the fraction of variance explained. (*C*) The percentage of variance in the steady states explained by the different principal components in the case of the network inferred from a biological dataset (34) (pink curve) and from randomized gene expression datasets (blue curves). (*D*, *Left*) The network inferred from the expression profiles of breast tumor samples (34) exhibits steady states with gene expression patterns similar to those seen in the breast tumor samples in *A*. (*D, Right*) Simpler networks obtained via node deletion (*Top Row*), edge deletion (*Middle Row*), or coarse graining (*Bottom Row*) recapitulate the expression patterns exhibited by the larger inferred network. Each heatmap shows the expression levels of the same 30 nodes of interest across the different steady states. The simpler networks were obtained by deleting nodes and edges in random order while retaining the 30 nodes of interest at each stage. While applying the coarse-graining procedure, the 30 nodes of interest were not combined with any other network node at any stage. In all heatmaps, different steady states are shown along the rows, while the network nodes are shown along the columns of the heatmaps. The heatmap in *A* was obtained by hierarchically clustering both the rows (i.e., the patient expression profiles) and the columns (i.e., the genes of interest). The heatmaps in *D* were generated by hierarchically clustering the rows (i.e., the steady states): each heatmap shows the network nodes (i.e., the columns) in the same order as in *A*. The names of the 30 nodes of interest are listed below the heatmap in *A*.

as shown in ref. 23, cell-fate networks are minimally frustrated, it is possible to develop network models that can recapitulate the observed behavior without the need to incorporate all of the details of biological regulation.

The minimal frustration property described here is closely related to the near-monotone nature of biological networks characterized previously by Sontag et al. (44, 45). In fact, for the system of ODEs in Eq. **2**, the species graph (a signed digraph), as defined in ref. 44, is given by the connection matrix J defined in *Modeling Regulatory Dynamics* (ignoring self-edges, same as the case in ref. 44). A system is said to be near-monotone if there exists a spin assignment $\{\sigma_i\}$ for the *species graph* such that the number of edges for which $J_{ij}\sigma_i\sigma_j < 0$ is close to zero (closer to zero as compared to the case of random networks). Thus, the definition of near-monotonicity is the same as the definition of minimal frustration (provided that all self-edges are ignored). Monotone systems are unlikely to exhibit chaotic behaviors, and their response to perturbations is robust and predictable. The near-monotonicity of biological networks has been implicated in their dynamically stable behavior (45). However, monotonicity analysis relies solely on the existence of a minimally inconsistent spin assignment and does not consider whether that spin assignment has any special biological meaning. The analysis in ref. 23 showed that the minimally inconsistent spin assignments correspond to steady states of Boolean dynamics that are biologically important and exhibit very large basins of attraction as compared to the more inconsistent spin assignments. In predicting that the dynamics of minimally frustrated, or near-monotone, biological networks are largely one-dimensional, with steady states corresponding to biological gene expression patterns having very large basins of attraction, the present analysis provides far more functional insight than the prediction of monotone dynamics for Eq. **2**. Interestingly, the mathematical framework for characterizing the monotonicity of dynamical systems, along with the tools for analyzing network state spaces and controllability in the context of monotone systems/subsystems (46, 47), provides a promising methodology for extending the concept of minimal frustration beyond regulatory networks to signaling and other biochemical networks. This idea will be explored in a future study.

As genome-wide transcriptional profiling became commonplace, first via microarray analysis and then via RNA-seq, it was recognized that gene expression datasets are effectively low dimensional, as evidenced by the extensive covariation in the gene expression levels (48, 49). The observed low dimensionality has been attributed to the coregulation of genes within regulatory modules, with the number of modules in the underlying network determining the dimensionality of the gene expression dataset (30). Our analysis suggests that the regulatory network need not be modular to generate gene expression datasets that are low dimensional: The randomized networks in our analysis can be more or less modular as compared to the biological network with similar topological features (*SI Appendix*, Fig. S1). The steady-state expression profiles obtained from the biological networks are however always more one-dimensional as compared to their random counterparts (Fig. 1). While selection for networks that exhibit low-dimensional steady-state expression patterns automatically results in the emergence of minimally frustrated networks (Fig. 2 *A* and *B*), such a selection pressure does not result in networks with higher modularity (*SI Appendix*, Fig. S4). These results clearly establish that low-dimensional steady-state gene expression is a consequence of the minimal frustration property of the underlying regulatory network, independent of the network modularity. Note that a different scenario wherein a large network can exhibit low-dimensional behavior is when the network consists of a very small "core" subnetwork and numerous downstream genes that are regulated by the core subnetwork with no feedback from the downstream genes to this subnetwork. If this was the case with the networks analyzed here, biological networks should exhibit a very different flow hierarchy (50) than random networks. We have previously shown that this is not the case (Fig. S1 in ref. 23): the hierarchical structure in our studied biological networks is comparable to that of random networks with similar topological properties.

The results presented here and in our previous study (23) raise the following question: what kind of evolutionary selection pressure could have resulted in minimally frustrated regulatory networks? We posit that while macroscopically simple cell-fate decision processes can be regulated by small gene networks such as by a toggle switch with only two genes (51), increasing complexity in higher organisms would have favored the emergence of larger, more complex regulatory networks so as to accommodate a broader range of inputs to the cell-fate decision-making process, which must nevertheless remain decisive. These larger networks would also allow for various interactions between numerous cellular processes in higher organisms. Large networks are also more robust to the loss of one or more genes as compared to smaller networks (which may become dysfunctional upon mutation in a single gene) and are more likely to exhibit parameter-independent dynamical behavior. Emergence of the minimal frustration property as smaller networks grew into larger, complex networks over the course of evolution would help ensure that the large network can still regulate a cell-fate decision-making process that is macroscopically simple. This is illustrated by the rudimentary simulation of network growth in *SI Appendix*, Fig. S5. The importance of maintaining macroscopically simple cell-fate decision-making processes, even in higher organisms, is a more difficult question and beyond the scope of the present study.

Sethna et al. have previously shown that systems biology models are sloppy—the dynamical behavior of these models is dominated by a small number of combinations of kinetic parameters (52, 53). This property makes dynamical models with poorly constrained kinetic parameters sufficient for recapitulating biological behavior and for making useful predictions. In the present study, we have shown that regulatory network models can recapitulate experimental behavior and make useful predictions even if the network connection matrix J is poorly constrained: one need not have a complete and exact description of the regulatory network underlying a physiological process to obtain a useful model of the process. Interestingly, while the sloppy parameter sensitivities reported by Sethna et al. are not limited to systems biology models and seem to extend to multiparameter models in general (54, 55), the behavior reported in the present study is restricted to minimally frustrated networks. In any case, it would be interesting to compare the sloppiness property of biological regulatory networks with that of random networks exhibiting similar topological properties.

Just as parameter fits even to comprehensive time series data fail to yield precise estimates of the underlying kinetic parameters due to the sloppy parameter sensitivities of systems biology models (53), our analysis suggests that collecting gene expression profiles at increasingly higher resolution and from more and more cells (56) is unlikely to yield more accurate biological regulatory networks. Pratapa et al. (36), benchmarking twelve different network inference algorithms on a variety of simulated and experimental gene expression datasets, found low stability in network prediction across datasets for the same biological process and little agreement between the predictions by different algorithms for the same dataset. This is unsurprising in light of our observation that the expression profiles generated by biological networks and other minimally frustrated networks can be recapitulated by various simpler networks that lack several of the nodes and edges present in the original network (Figs. 3 and 4), as well as by lower-resolution, coarse-grained networks that approximate the activity of several nodes by a single regulator (Fig. 5). Thus, the present study explains why the inference of gene regulatory networks from expression data remains a formidable challenge despite more than 20 years of research and one that is unlikely to benefit from higher-resolution experimental data (36). Instead of striving to obtain exact regulatory networks involved in establishing cell type-specific gene expression patterns by collecting higher-resolution data and employing advanced statistical techniques, efforts to understand cell-fate choice must focus on building imperfect, predictive network models with rapidly verifiable predictions, and on carrying out experiments that are optimally designed to constrain the network connection matrix.

Our analysis shows that in the case of biological regulatory networks and in the case of minimally frustrated networks in general, steady-state expression patterns are largely preserved under progressive coarse graining of the network, a simplification procedure during which sets of network nodes are combined into single regulators (Fig. 5). Such behavior was previously reported for the small cell lung cancer (SCLC) network (16) analyzed in Fig. 1 (57). This result provides a theoretical explanation for the popularity, despite the increasingly high resolution at which gene expression levels can be characterized by modern experimental techniques (56), of approaches involving the aggregation of genes into meaningful sets. One such approach is gene set enrichment analysis (58), whose popularity has endured the transition from microarray analysis to RNA-seq as the preferred method for transcriptome characterization. Previous studies have shown that gene pathways–based metrics that coarse-grain the information contained in the expression levels of the constituent genes retain crucial information about the biological sample (32). Note that

the strategy to coarse-grain a regulatory network described here is fairly simplistic and is proposed only as a sample strategy to demonstrate that the network behavior can be preserved under such a procedure. There exist numerous possibilities for vastly improving upon the present strategy; for example, one could introduce an objective function that a network coarse-graining procedure must optimize. Strategies for coarse-graining other biological network models including signaling networks and metabolic networks have been described elsewhere (59, 60) and could motivate improvements to the strategy introduced here for regulatory networks.

While previous studies have attributed the robustness of biological networks to the different topological features of these networks, our analysis posits a fairly simple explanation: the structure of these networks is far more complex and requires far more information to describe as compared to the dynamics exhibited by the underlying biological process. Consequently, loss of features such as nodes and edges from the network description is unlikely to significantly affect the behavior. Since simple steady-state dynamics can emerge from a complex regulatory network only if the network is minimally frustrated, we posit that the minimal frustration property of biological networks is responsible for their functional robustness.

Finally, we note that the minimal frustration property, along with the simple steady-state dynamics, will be seen in biological networks of sufficient size and complexity; the minimal frustration framework becomes less insightful in the case of small network models of cell-fate choice that have been simplified to remove the many redundancies characteristic of minimally frustrated networks. Furthermore, the present analysis is restricted to networks that are involved in cell-fate choice either between two phenotypic states or along a single phenotypic axis, such as the epithelial–mesenchymal axis in the case of the EMT networks (14, 22) we have analyzed. In the case of network models that couple two (or more) cell-fate choice processes (see ref. 61 for an example), preliminary analysis suggests that strong coupling between cell fates along with the minimal frustration property would be associated with coexistence of specific pairs of phenotypic states. Analysis of the coupling between genes involved in different cell-fate choice processes in biology will be required to determine whether minimal frustration is also a property of large networks formed by coupling networks regulating different cell-fate choice processes.

Data, Materials, and Software Availability. All study data are included in the article and/or *SI Appendix*.

ACKNOWLEDGMENTS. This work was supported by the NSF grant PHY-2019745.

Author affiliations: [a]PhD Program in Systems, Synthetic, and Physical Biology, Rice University, Houston, TX 77005; [b]Center for Theoretical Biological Physics, Northeastern University, Boston, MA 02115; [c]Department of Physics, Northeastern University, Boston, MA 02115; and [d]Department of Physics, Bar-Ilan University, Ramat-Gan 52900, Israel

1. T. Ideker, R. Nussinov, Network approaches and applications in biology. *PLOS Comput. Biol.* **13**, 1–3 (2017).
2. N. D. Price, I. Shmulevich, Biochemical and statistical network models for systems biology. *Curr. Opin. Biotechnol.* **18**, 365–370 (2007).
3. R. Albert, Network inference, analysis, and modeling in systems biology. *Plant Cell* **19**, 3327–3338 (2007).
4. T. Charitou, K. Bryan, D. J. Lynn, Using biological networks to integrate, visualize and analyze genomics data. *Genet. Sel. Evol.* **48**, 27–39 (2016).
5. F. J. Bruggeman, H. V. Westerhoff, The nature of systems biology. *Trends Microbiol.* **15**, 45–50 (2007).
6. R. Albert, Scale-free networks in cell biology. *J. Cell Sci.* **118**, 4947–4957 (2005).
7. M. Aldana, P. Cluzel, A natural class of robust networks. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 8710–8714 (2003).
8. H. Yu, M. Gerstein, Genomic analysis of the hierarchical structure of regulatory networks. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 14724–14731 (2006).
9. R. Milo *et al*., Network motifs: Simple building blocks of complex networks. *Science* **298**, 824–827 (2002).
10. J. E. Bailey, Complex biology with no parameters. *Nat. Biotechnol.* **19**, 503–504 (2001).
11. R. Palmer, "Broken ergodicity" in *Lectures in the Sciences of Complexity*, D. L. Stein, Ed. (Addison-Wesley, Reading, MA, ed. 1, 1989), pp. 275–300.
12. T. Enver, M. Pera, C. Peterson, P. W. Andrews, Stem cell states, fates, and the rules of attraction. *Cell Stem Cell* **4**, 387–397 (2009).
13. S. N. Steinway *et al*., Network modeling of TGFβ signaling in hepatocellular carcinoma epithelial-to-mesenchymal transition reveals joint Sonic hedgehog and Wnt pathway activation. *Cancer Res.* **74**, 5963–5977 (2014).
14. F. Font-Clos, S. Zapperi, C. A. M. L. Porta, Topography of epithelial-mesenchymal plasticity. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 5902–5907 (2018).
15. R. Chang, R. Shoemaker, W. Wang, Systematic search for recipes to generate induced pluripotent stem cells. *PLOS Comput. Biol.* **7**, e1002300 (2011).
16. A. R. Udyavar *et al*., Novel hybrid phenotype revealed in small cell lung cancer by a transcription factor network model that can explain tumor heterogeneity. *Cancer Res.* **77**, 1063–1074 (2017).
17. S. Li, X. Zhu, B. Liu, G. Wang, P. Ao, Endogenous molecular network reveals two mechanisms of heterogeneity within gastric cancer. *Oncotarget* **6**, 13607–13627 (2015).
18. O. Ríos *et al*., A Boolean network model of human gonadal sex determination. *Theor. Biol. Med. Model.* **12**, 26–44 (2015).
19. O. Hobert, Gene regulation by transcription factors and microRNAs. *Science* **319**, 1785–1786 (2008).
20. M. Lu, M. K. Jolly, H. Levine, J. N. Onuchic, E. Ben-Jacob, MicroRNA-based regulation of epithelial-hybrid-mesenchymal fate determination. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 18144–18149 (2013).
21. X. J. Tian, H. Zhang, J. Xing, Coupled reversible and irreversible bistable switches underlying TGFβ-induced epithelial to mesenchymal transition. *Biophys. J.* **105**, 1079–1089 (2013).
22. D. Jia *et al*., Testing the gene expression classification of the EMT spectrum. *Phys. Biol.* **16**, 025002 (2019).
23. S. Tripathi, D. A. Kessler, H. Levine, Biological networks regulating cell fate choice are minimally frustrated. *Phys. Rev. Lett.* **125**, 088101 (2020).
24. R. Albert, J. Thakar, Boolean modeling: A logic-based dynamic approach for understanding signaling and regulatory networks and for making useful predictions. *WIREs Syst. Biol. Med.* **6**, 353–369 (2014).
25. P. Anderson, The concept of frustration in spin glasses. *J. Less Common Met.* **62**, 291–294 (1978).
26. N. Dugué, A. Perez, Directed Louvain: Maximizing modularity in directed networks. *Univ. Res. Rep.* (2015).
27. B. Huang *et al*., Interrogating the topological robustness of gene regulatory circuits by randomization. *PLOS Comput. Biol.* **13**, 1–21 (2017).
28. I. Pastushenko *et al*., Identification of the tumour transition states occurring during EMT. *Nature* **556**, 463–468 (2018).
29. W. Wang, D. Poe, K. Ni, J. Xing, Cell phenotypic transition proceeds through concerted reorganization of gene regulatory network. arXiv [preprint] (2021). http://arxiv.org/abs/2107.03581.
30. G. Heimberg, R. Bhatnagar, H. El-Samad, M. Thomson, Low dimensionality in gene expression data enables the accurate extraction of transcriptional programs from shallow sequencing. *Cell Syst.* **2**, 239–250 (2016).
31. J. Stelling, U. Sauer, Z. Szallasi, F. J. Doyle, J. Doyle, Robustness of cellular functions. *Cell* **118**, 675–685 (2004).
32. Y. Drier, M. Sheffer, E. Domany, Pathway-based personalized analysis of cancer. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 6388–6393 (2013).
33. A. P. Deshmukh *et al*., Identification of EMT signaling cross-talk and gene regulatory networks by single-cell RNA sequencing. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2102050118 (2021).
34. A. Terunuma *et al*., Myc-driven accumulation of 2-hydroxyglutarate is associated with breast cancer prognosis. *J. Clin. Investig.* **124**, 398–412 (2014).
35. T. Moerman *et al*., GRNBoost2 and Arboreto: Efficient and scalable inference of gene regulatory networks. *Bioinformatics* **35**, 2159–2161 (2018).
36. A. Pratapa, A. P. Jalihal, J. N. Law, A. Bharadwaj, T. Murali, Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nat. Methods* **17**, 147–154 (2020).
37. S. Chandriani *et al*., A core MYC gene expression signature is prominent in basal-like breast cancer but only partially overlaps the core serum response. *PLOS ONE*, 1–13 (2009).
38. W. Saelens, R. Cannoodt, H. Todorov, Y. Saeys, A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37**, 547–554 (2019).
39. K. R. Moon *et al*., Manifold learning-based methods for analyzing single-cell RNA-sequencing data. *Curr. Opin. Syst. Biol.* **7**, 36–46 (2018).
40. J. T. George, M. K. Jolly, S. Xu, J. A. Somarelli, H. Levine, Survival outcomes in cancer patients predicted by a partial EMT gene expression scoring metric. *Cancer Res.* **77**, 6415–6428 (2017).
41. T. Z. Tan *et al*., Epithelial-mesenchymal transition spectrum quantification and its efficacy in deciphering survival and drug responses of cancer patients. *EMBO Mol. Med.* **6**, 1279–1293 (2014).
42. L. A. Byers *et al*., An epithelial-mesenchymal transition gene signature predicts resistance to EGFR and PI3K inhibitors and identifies Axl as a therapeutic target for overcoming EGFR inhibitor resistance. *Clin. Cancer Res.* **19**, 279–290 (2013).
43. S. W. Ng *et al*., A 17-gene stemness score for rapid determination of risk in acute leukaemia. *Nature* **540**, 433–437 (2016).
44. E. D. Sontag, Monotone and near-monotone biochemical networks. *Syst. Synth. Biol.* **1**, 59–87 (2007).
45. A. Ma'ayan, A. Lipshtat, R. Iyengar, E. Sontag, Proximity of intracellular regulatory networks to monotone systems. *IET Syst. Biol.* **2**, 103–112 (2008).
46. J. C. Rozum, R. Albert, Identifying (un)controllable dynamical behavior in complex networks. *PLOS Comput. Biol.* **14**, e1006630 (2018).

47. J. C. Rozum, R. Albert, Self-sustaining positive feedback loops in discrete and continuous systems. *J. Theor. Biol.* **459**, 36–44 (2018).
48. M. B. Eisen, P. T. Spellman, P. O. Brown, D. Botstein, Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U.S.A.* **95**, 14863–14868 (1998).
49. S. Bergmann, J. Ihmels, N. Barkai, Iterative signature algorithm for the analysis of large-scale gene expression data. *Phys. Rev. E* **67**, 031902 (2003).
50. E. Mones, L. Vicsek, T. Vicsek, Hierarchy measure for complex networks. *PLOS ONE* **7**, e33799 (2012).
51. D. Jia *et al*., Operating principles of tristable circuits regulating cellular differentiation. *Phys. Biol.* **14**, 035007 (2017).
52. K. S. Brown *et al*., The statistical mechanics of complex signaling networks: Nerve growth factor signaling. *Phys. Biol.* **1**, 184–195 (2004).
53. R. N. Gutenkunst *et al*., Universally sloppy parameter sensitivities in systems biology models. *PLOS Comput. Biol.* **3**, 1–8 (2007).
54. K. S. Brown, J. P. Sethna, Statistical mechanical approaches to models with many poorly known parameters. *Phys. Rev. E* **68**, 021904 (2003).
55. B. B. Machta, R. Chachra, M. K. Transtrum, J. P. Sethna, Parameter space compression underlies emergent theories and predictive models. *Science* **342**, 604–607 (2013).
56. V. Svensson, R. Vento-Tormo, S. A. Teichmann, Exponential scaling of single-cell RNA-seq in the past decade. *Nat. Protoc.* **13**, 599–604 (2018).
57. L. Chauhan, U. Ram, K. Hari, M. K. Jolly, Topological signatures in regulatory network enable phenotypic heterogeneity in small cell lung cancer. *eLife* **10**, e64522 (2021).
58. A. Subramanian *et al*., Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 15545–15550 (2005).
59. M. K. Transtrum, P. Qiu, Model reduction by manifold boundaries. *Phys. Rev. Lett.* **113**, 098701 (2014).
60. H. M. Rasool, S. H. Khoshnaw, Techniques of model reductions in biochemical cell signaling pathways. arXiv [preprint] (2021). http://arxiv.org/abs/2109.06566.
61. C. Li, G. Balazsi, A landscape view on the interplay between EMT and cancer metastasis. *Npj Syst. Biol. Appl.* **4**, 34 (2018).