A multi-level formalism to capture the characteristics of hybrid phenotypes in Epithelial-Mesenchymal Plasticity

Kishore Hari, Shubham Tripathi, Vaibhav Anand, Mohit Kumar Jolly, Herbert Levine

November 26, 2024

Email: h.levine@norteastern.edu

Abstract

Epithelial-mesenchymal plasticity (EMP) is a cell-fate switching program that enables cells to adopt a spectrum of phenotypes ranging from epithelial (E) to mesenchymal (M) and hybrid E/M states. Hybrid E/M phenotypes are considered optimal for cancer metastasis due to their association with metastatic initiation and traits such as stemness, drug resistance, and collective migration. Boolean models of gene regulatory networks (GRNs) underlying EMP have provided valuable insights into the dynamics of E and M phenotypes. However, these models fail to capture hybrid phenotypes effectively, as Boolean frameworks allow only binary gene expression, while hybrid E/M cells exhibit partial expression of both E and M genes. To address this limitation, we modified a threshold-based Boolean formalism to incorporate intermediate gene expression levels. The resulting multilevel model reveals new hybrid steady states with partial expression of both E and M genes, expanding the phenotypic landscape beyond that captured by traditional Boolean models. These hybrid states exhibit lower frustration and higher abundance, indicating enhanced stability. By resolving degeneracy in the dynamics, we further demonstrate that hybrid states identified by the multilevel model are more stable. These findings suggest that adding minimal complexity to Boolean models can uncover previously hidden qualitative features of phenotypic landscapes governed by GRNs.

1 Introduction

Metastasis, the primary cause of cancer mortality [1], is the mechanism by which cancer cells spread to multiple organs. Carcinomas are tumors that initially have an epithelial phenotype, enabling the cancer cells to adhere to each other and to their surroundings. To undergo metastasis, these cells must detach from the primary tumor site and migrate through the bloodstream to a suitable secondary site; they can accomplish this by acquiring a mesenchymal phenotype. This migration concludes with the cells invading and colonizing the secondary site, eventually regaining the epithelial phenotype [2]. This possibility of reversible transitions between epithelial and mesenchymal phenotypes, known as epithelial-mesenchymal plasticity (EMP), is a crucial enabler of metastasis [3].

Among the spectrum of states that are accessible to cells exhibiting EMP, hybrid phenotypes containing both epithelial and mesenchymal properties are the most conducive to metastasis [4–6]. Hybrid E/M phenotype has been implicated to be important in collective migration of cancer cells [7, 8], a mechanism of migration that has been shown to be most efficient in forming metastases [9, 10]. Most importantly, hybrid E/M phenotypes have also been shown to have enhanced metastatic site colonization capabilities such as stemness, as well as natural drug resistance [6, 11] and ability to withstand immune attack [12].

Gene regulation plays an important role in the cell-fate transitions underlying EMP [13] and mathematical models of gene regulatory networks (GRNs) have been widely used to understand tumor progression [14, 15] and the emergence of various phenotypes. including hybrid E/M phenotypes [16]. These GRNs are simulated to calculate the dynamics and steady-state behavior of the genes involved, and the steady states are assumed to represent the experimentally observed phenotypes. These formalisms can be broadly classified as continuous versus discrete, based on the expression levels allowed for the genes of the network. Continuous formalisms often involve ordinary differential equations that describe the production and degradation kinetics of proteins, where the gene expressions are real positive numbers. Discrete formalisms, on the other hand, limit the expression levels of genes to a few levels. The most popular discrete simulation formalism is the Boolean approach, where each gene can only have two expression levels: high and low. While continuous models are potentially more accurate in describing gene expression dynamics, these models also require knowledge of a large number of kinetic parameters in addition to network information. Boolean models on the other hand are simpler and therefore require a fewer number of parameters [17]. Furthermore, the sigmoidal nature of biomolecular interactions (e.g., transcription factor binding to DNA) introduces ultrasensitivity, thereby providing qualitative validation to the Booleanization of gene expression.

Studies simulating EMP networks using continuous ODE-based formalisms show the existence of stable hybrid phenotypes with partial expression of epithelial and mesenchymal nodes [18, 19], a property observed in transcriptomic data as well (see below). Boolean models on the other hand, which only allow two gene expression levels (high/low), cannot capture this type of expression pattern. Instead, they exhibit a different kind of stable hybrid states with high expression of a subset of epithelial and mesenchymal genes [15, 20]. The question then arises as to whether one can remedy this deficiency while maintaining many of the useful aspects of Boolean approaches. Here, we tackle this issue by expanding the state space of boolean simulations by allowing partial gene expression. We show that the modified multi-level Boolean formalism is capable of generating a larger number of steady states than the traditional formalism. Crucially, the hybrid states generated in this new formalism allow for partial expression of E and M genes. Using a further theoretical modification to the simulation formalism, we show that the hybrid states generated by the multi-level formalism can be much more robust than those arising in the traditional Boolean formulation.

2 Results

2.1 Partial expression of genes in transcriptomics

The simple Boolean model assumes the dynamics

$$s_{i_{(t+1)}} = \begin{cases} -1, & if \quad \Sigma_j J_{ji} s_j < 0 \\ 1, & if \quad 0 < \Sigma_j J_{ji} s_j \\ s_{i_t}, & if \quad \Sigma_j J_{ji} s_j = 0 \end{cases} \tag{1}$$

where $s_i(t)$ is the expression level of the ith gene/node (please note that we use "gene" and "node" interchangeably from here on) in the network that can take two values: -1 and 1. J_{ij} is the edge from

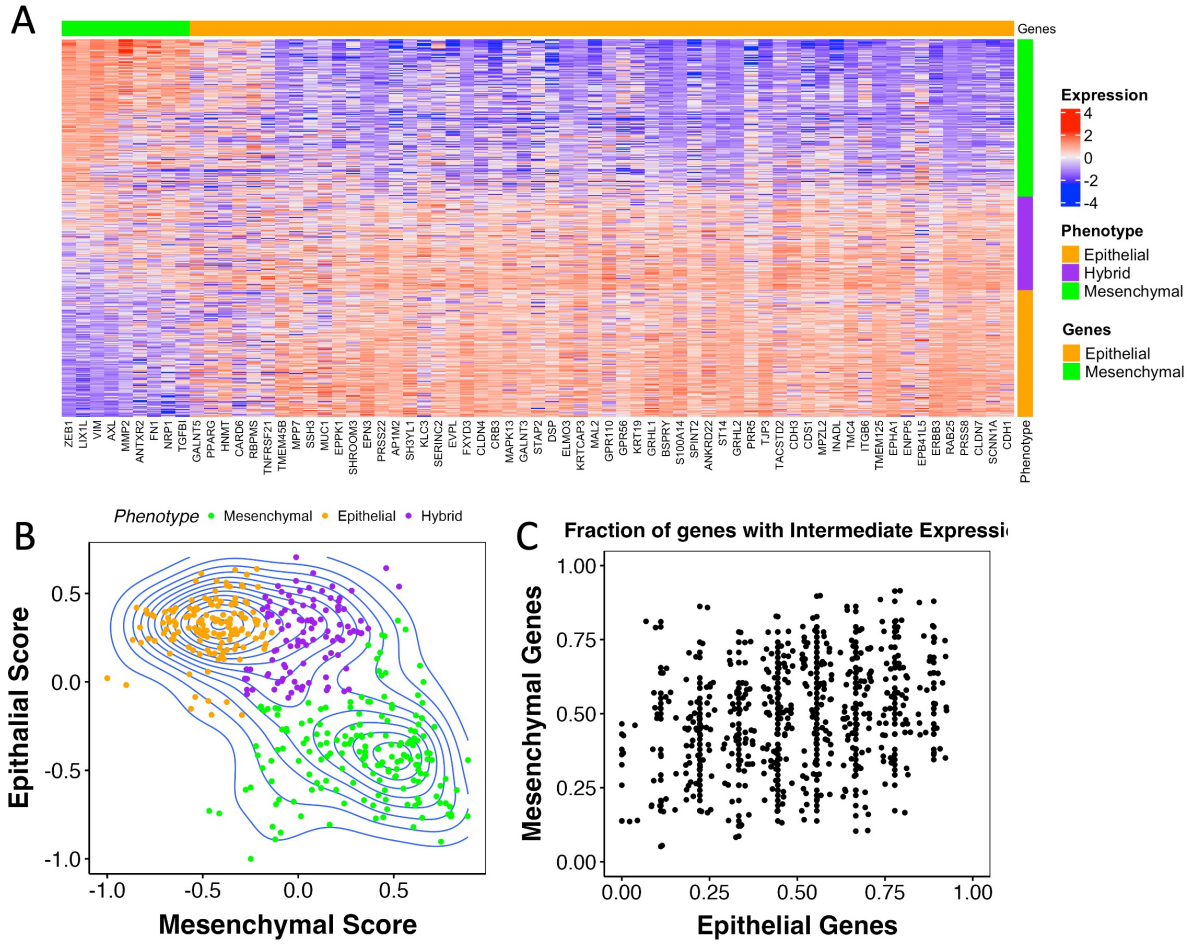

Figure 1: **Partial expression of genes in transcriptomics.** **A** Log normalized expression levels of CCLE carcinoma cell lines for a 76 gene set for EMP [21]. **B** Scatterplot depicting the Epithelial and Mesenchymal scores calculated from the expression matrix in **A**. Each point is a sample colored by its corresponding phenotypic classification. The contours represent the density of points at each coordinate, with smaller circles indicating higher density. **C** Scatterplot showing the fraction of Epithelial and Mesenchymal genes with intermediate expression levels (between 25th and 75th quantile).

the jth node to the ith node, and takes the value -1 for inhibiting edge, 1 for activating edge and 0 for no edge. Models of this form with a variety of different networks (i.e. the set of J_{ij}'s) have been used for EMP [14, 20, 22, 23]. These models have successfully captured the properties of the two dominant phenotypes (E versus M) and the transitions which will occur, for example, upon over-expression of some of the state specific factors.

Boolean models have also exhibited the presence of hybrid phenotypes, steady-states of the dynamical process which cannot be classified as either E or M. This possibility was initially suggested in studies of very small networks treated as sets of coupled ODE's [18, 24] and subsequently identified in a variety of cell lines [25, 26] and in-patient data [27, 28]. As mentioned above, these hybrid cells are particularly relevant for metastatic disease, as their plasticity enables secondary tumor growth and drug resistance [29]. From these studies, the basic profile of a hybrid E/M state was one which when subjected to EMP induction retained some intermediate degree of epithelial markers even as they expressed some intermediate degree of mesenchymal ones. These properties were then seen in larger networks, through use of the RACIPE computational framework [19] which scans through relevant parts of parameter space. Furthermore, these states were quite stable, as evidenced by their persistence for very long times in laboratory experiments [25], despite often having low abundance in heterogeneous EMP populations [26].

However, Boolean models were unable to capture hybrid states with the correct properties. Specifically, the framework does not consider the relative strengths of the network edges and booleanizes the expression levels of the nodes. As a consequence, hybrid steady states captured by Boolean models fully express subsets of both epithelial and mesenchymal genes and fully suppress other epithelial and mesenchymal genes. However, in transcriptomic data, we see that none of the samples show extreme (very high or very low) expression of genes in hybrid states. As an example, **Figure 1A** shows the expression levels of EMP genes [21] in carcinoma cell lines obtained from the CCLE (Cancer Cell line expression) database [30]. We classified each sample into Epithelial, Mesenchymal, and Hybrid phenotypes using scores calculated from the expression matrix (**Figure 1B**) and found a higher abundance of E and M samples than hybrid samples. We further looked into the expression levels of each gene, defining intermediate levels of expression to range from 25th to 75th percentile of their corresponding expression vectors across as samples. Notably, most of the samples have significant numbers of epithelial and mesenchymal genes showing partial expression (**Figure 1C**).

To capture such partial expressions, we applied a simple extension to the Boolean model in equation 1 to expand the number of levels of expression a node can have. We defined a "multi-level" formalism, allowing additional expression levels such that each node can take four values: $-1, -0.5, 0.5, and 1$. We define the corresponding update rules as follows:

$$S_{i_{(t+1)}} = \begin{cases} -1, & if \ \frac{\Sigma_j J_{ji} s_j}{d_i} < -0.5 \\ -0.5, & if \ -0.5 \leq \frac{\Sigma_j J_{ji} s_j}{d_i} < 0 \\ 0.5, & if \ 0 < \frac{\Sigma_j J_{ji} s_j}{d_i} \leq 0.5 \\ 1, & if \ 0.5 < \frac{\Sigma_j J_{ji} s_j}{d_i} \\ S_{i_t}, & if \ \frac{\Sigma_j J_{ji} s_j}{d_i} = 0 \end{cases} \tag{2}$$

Here, d_i is the in-degree of ith node. From here on, we refer to the model defined by Equation 1 as the two-level model, allowing two levels of expression: -1 and 1. Similarly, we refer to the model defined by Equation 2 as the four-level model. We compared the steady-state space obtained by simulating EMP GRNs using two and four-level formalisms to understand the composition and stability of hybrid and non-hybrid phenotypes.

2.2 Multilevel models capture stable hybrid states with partial expression of Epithelial and Mesenchymal nodes

We began by analyzing an EMP network with 26 nodes and 100 edges (26N100E). This network has two mutually inhibiting groups or "teams" of nodes corresponding to Epithelial (micro-RNAs, CDH1, CLDN7, GRHL2), and Mesenchymal (FOXC2, GSC, KLF8, SNAI1, SNAI2, TCF3, TGFbeta, Twist1, Twist2, Zeb1, and Zeb2) phenotypes respectively [15]. The interactions within a team are predominantly positive, while those across teams are predominantly negative. We have previously shown that this configuration of mutually inhibiting teams is a design principle of EMP networks, and it destabilizes hybrid steady states, leading to their low frequency. In addition to the teams, this network has genes known as PSFs (phenotype stability factors [25]): NP63, and Ovol2, which do not align with any team

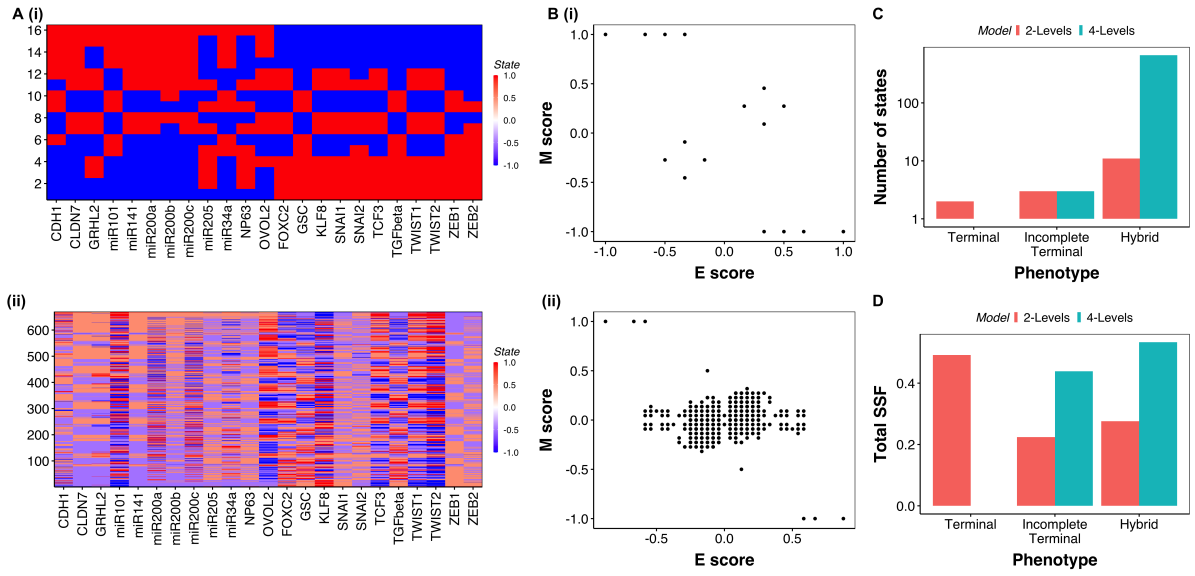

Figure 2: **Multilevel formalism leads to an increased number of hybrid states A** Heatmaps depicting the configurations of steady states of the 23N90E EMP network simulated using i) two-level and ii) four-level threshold-based formalism. Each row corresponds to one steady state, each column corresponds to a node labeled at the bottom. Red corresponds to positive expression level and blue corresponds to a negative expression level. **B** 2-D contour plots of the steady states obtained using i) two-level and ii) four-level threshold based formalism, on the E and M score axes. Each point is a steady state. **C** Barplot depicting the comparison of number of steady states (y axis) of each category (Hybrid, Terminal, and Incomplete terminal, x-axis) between two-level formalin (red) and four-level formalism (green). **D** Same as **C**, but for the total frequency of all steady states of a category.

and therefore might stabilize hybrid phenotypes. Due to these features, this network can indeed exhibit stable, low-frequency, hybrid phenotypes. For ease of analysis, we removed the signal nodes (nodes that do not have inputs from any other nodes in the network), as these nodes create inconsequential degeneracy in the steady states. The resultant network has 23 nodes and 90 edges (23N90E).

Using Equations 1 and 2, we simulated the network starting from 10^7 random initial conditions (RIC) till each RIC converged to a steady state. Each RIC can be thought of as analogous to a cell in a heterogeneous population, with the steady state being analogous to the phenotype of that cell. To classify these states/phenotypes, we define epithelial and mesenchymal scores as the average expression of nodes in the epithelial and mesenchymal teams, respectively, for a given steady state (**See methods**). We further defined phenotypic score as the difference between the epithelial and mesenchymal score for a given steady state. While the epithelial and mesenchymal scores range between $[-1, 1]$, the phenotypic score ranges between $[-2, 2]$. The epithelial (mesenchymal) states, defined as the states with the high expression level (1) of only epithelial (mesenchymal) nodes, and low expression level (-1) of all mesenchymal (epithelial nodes) have $+2$ and -2 scores respectively. We classify the Epithelial and Mesenchymal steady states, with an absolute phenotypic score of 2, as "Terminal" steady states. We further defined a class of steady states called the "Incomplete terminal states", with either epithelial score or the mesenchymal score equal to -1, i.e., the steady states with incomplete expression levels of either only epithelial or only mesenchymal nodes. The remaining steady states that could not be classified as Terminal or Incomplete terminal were classified as "Hybrid" steady states. Note that the hybrid states have a low magnitude of phenotypic scores as compared to Terminal and Incomplete terminal states (**Figure 2B**).

The four-level model leads to a larger number of unique steady states (600 in the four-level model vs. 16 in the two-level model, **Figure 2A**). We asked which class of phenotype shows the largest increase in the number of steady states. Based on the characterization of phenotypes described previously, there can only be two terminal states in both two-level and four-level models. Hence, the increase in the number of states can only be due to either hybrid or incomplete terminal steady states. While the number of terminal and incomplete terminal states in the two-level model was comparable to that of hybrid states (5 vs 11), the four-level model results in a much larger number of hybrid states (compare the centers of **Figure 2B, i and ii**). Out of the 600 steady states observed in the four-level model, the largest

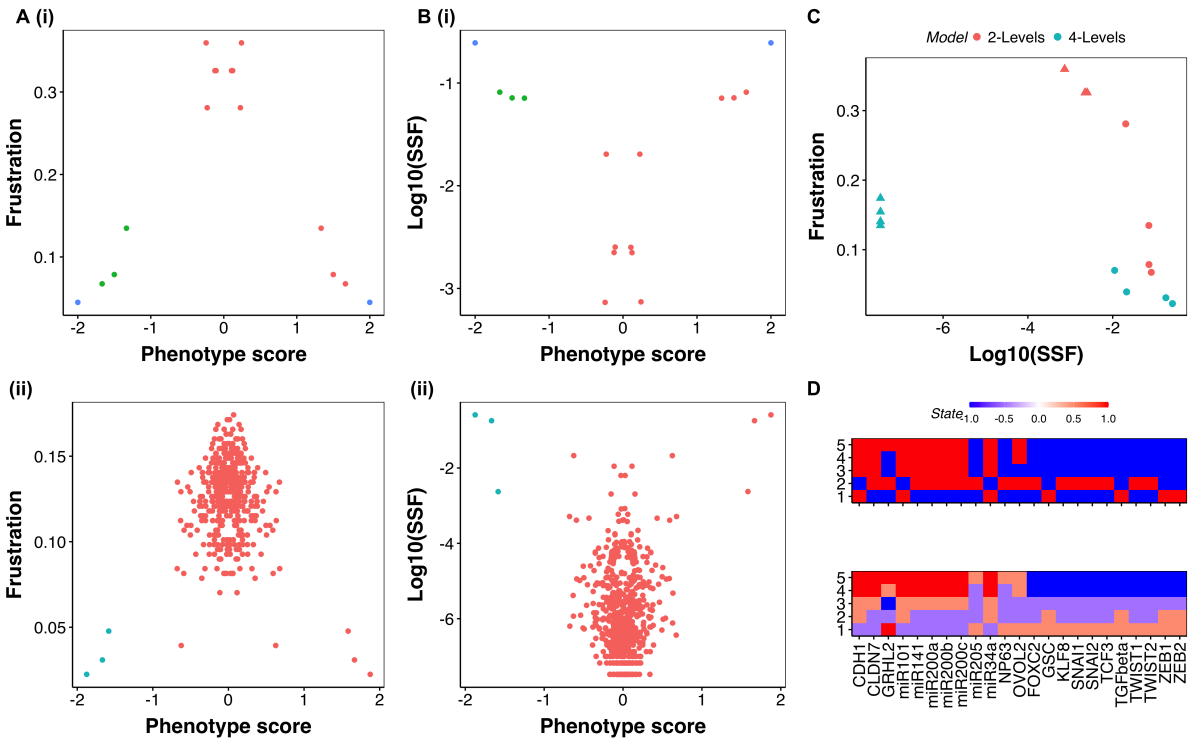

Figure 3: **Four-level model leads to high SSF, low frustration hybrid states with partial expression of nodes A** Scatter plots depicting the Frustration of the steady states of the 23N90E network simulated using i) two-level and ii) four-level formalism. **B** Same as A, but for steady state frequency (SSF). **C** Frustration and SSF of most frequent (circles) and least frequent (triangles) hybrid states from two-level (red) and four-level models (green). **D** Heatmaps depicting the composition of most frequent hybrid states from two-level (top) and four-level (bottom) formalisms.

fraction were hybrid states, and a much smaller fraction(¡1%) was incomplete terminal states (**Figure 2C**). Importantly, the increase in the number of steady states does not scale with the state space (2^n to 4^n), suggesting that the newly appeared hybrid states are not just an artifact of the increase in the state space.

From our simulations, we could also obtain the frequency of each steady state, defined as the number of initial conditions converging to the given steady state. We further calculated the sum of the steady state frequency (SSF) for all hybrid states, the indicator of the abundance of hybrid phenotype in the digital population of cells, and found that while the number of unique hybrid steady states showed a 15 fold increase from two-level to four-level model, the frequency of hybrid phenotype showed only a two-fold increase. This observation indicates significant heterogeneity in the stability of hybrid states and that fact that in general these states are much less stable than the others.

We further probed the stability of the steady states using the SSF and the the concept of frustration. Frustration of a steady state measures how well a steady state configuration agrees with the network structure, calculated as a weighted fraction of frustrated edges in the network for a given steady state configuration. Each edge in the network models has one of three weights: $-1, 1, or\ 0$, and a frustrated edge has the opposite sign as that of the product of the signs of the expression levels of the nodes participating in the edge. See methods for the detailed formulation of the frustration metric. In the two-level model, the states with a high magnitude of the phenotypic score (> 1) show low frustration, and vice-versa (**Figure 3A, i**). The four-level model leads to decrease in the overall frustration of the steady states across all ranges of the phenotypic score (**Figure 3A, ii**). It is still the case that the states with high phenotypic scores, i.e., the terminal and incomplete terminal states, maintain low frustration. While the majority of the hybrid states (with low phenotypic scores, colored red) show high frustration, we find some hybrid states in the four-level model with low phenotypic scores but nonetheless lower relative frustration (0.06 against a maximum frustration of 0.2) as compared to the two-level formalism (0.25 against a maximum of 0.4, **Figure 3A, i vs ii**). Thus, the four-level formalism allows for a larger number of low-frustration hybrid states. The terminal and incomplete terminal states have a higher SSF

(**Figure 3B**). A large fraction of the hybrid states have very low SSF; however, a few hybrid states with relatively higher frequency are found in both two-level and four-level formalisms, with both low (< 1) and high (> 1) magnitude of phenotypic scores.

We then compared the frustration and stability of the high-frequency hybrid states obtained from two-level and four-level models. The high-frequency hybrid states had lower frustration in four-level model, and had a comparable, albeit slightly higher SSF as that of the two-level model (**Figure 3C**). The composition of these states revealed further interesting results. One of the states in two-level formalism showed a higher expression of multiple nodes from the E and M teams. The rest of the high-frequency hybrid states were closer to E or M states, with a small fraction of the other team being expressed; perhaps these should not be thought of as true hybrid states. In contrast, the four-level model exhibits hybrid states with multiple nodes from both E and M teams having intermediate expression levels (-0.5, 0.5). However, the most stable hybrid states remained to be the ones closer to E or M states (**Figure 3D**, again indicating that these perhaps are not true hybrids. Thus, we find that multilevel formalism gives rise to a new type of stable hybrid state, with a partial expression of epithelial and mesenchymal nodes that are not seen in two-level formalism.

2.3 Eliminating degeneracy highlights the stability of the hybrid states obtained using the multilevel formalism

Our results so far have shown that increasing the state space by employing a multi-level formalism increases the incidence of low-frustration, high-SSF "true" hybrid states. Our previous results analyzing EMP networks have shown that the network structure, involving "teams" of E and M nodes, does not tend to favor hybrid states [15], which we found was the primary reason for their low-stability and plasticity. Therefore, the emergence of hybrid states has to be driven by events that happen outside the direct dictates of network topology. One example that we have seen before for such a behavior is the degeneracy in steady states introduced by the signal molecules in the network; these molecules are not regulated by any other node in the network and therefore retain their original expression level irrespective of the expression of their downstream genes.

We can use the threshold-based formalism studied here to study the possibility of a similar degeneracy arising during the dynamics of the non-signaling nodes. Specifically, consider again the models defined by equations 1 or 2. Here, when the sum $\Sigma J_{ij} s_j$ becomes zero for a node, the node retains its expression level from the previous step of the simulation. This can potentially introduce degeneracy in the state of the network irrespective of the network structure. We asked if the stability of hybrid states in two-level and multi-level formalisms are dependent on the degeneracy introduced by the dynamics. We therefore modified the two-state threshold-based formalism in the following manner to eliminate the emergence of degeneracy:

$$s_{i_{(t+1)}} = \begin{cases} -1, & if \;\; \Sigma_j J_{ji} s_j < 0 \\ 1, & if \;\; 0 < \Sigma_j J_{ji} s_j \\ 0, & if \;\; \Sigma_j J_{ji} s_j = 0 \end{cases}$$

This formalism, referred to as the "turn-off" formalism from here on, sets the expression of unregulated nodes to zero, making them inconsequential in the next step of the simulation. Therefore, the unregulated nodes are given a fixed state, eliminating their noise capability. One immediate implication of this formalism is that the expression of signal nodes (nodes with zero in-degree) is always set to zero. As signal nodes introduce degeneracy in Epithelial and Mesenchymal states (thereby creating microstates in the landscape), this formalism is expected to eliminate this degeneracy. We analyzed the 26N100E network (23N90E network in the previous figures plus signal nodes) with and without the turn-off formalism. In the two-level formalism, the 26N100E network gives rise to 114 unique steady states, out of which 8 are degenerate terminal states. Indeed, we see that applying the turn-off condition selectively to the signal nodes of the 26N100E network reduces the number of steady states from 114 to 20 (**Figure 4A, B**), with a single Epithelial and a single mesenchymal state (as opposed to 4 states each without node turn-off). We further extended this restriction to all nodes and found that the number of steady states decreased to four (**Figure 4C**). Among these four states, the Epithelial and Mesenchymal states have a total frequency of 0.999, a significant increase from the 0.5 (0.75 including incomplete terminal states) without turn-off.

Thus, we can conclude that the hybrid states obtained in the two-level formalism largely emerge due to the degeneracy in the dynamics. Nodes like NP63, which have conflicting interactions with the "team"

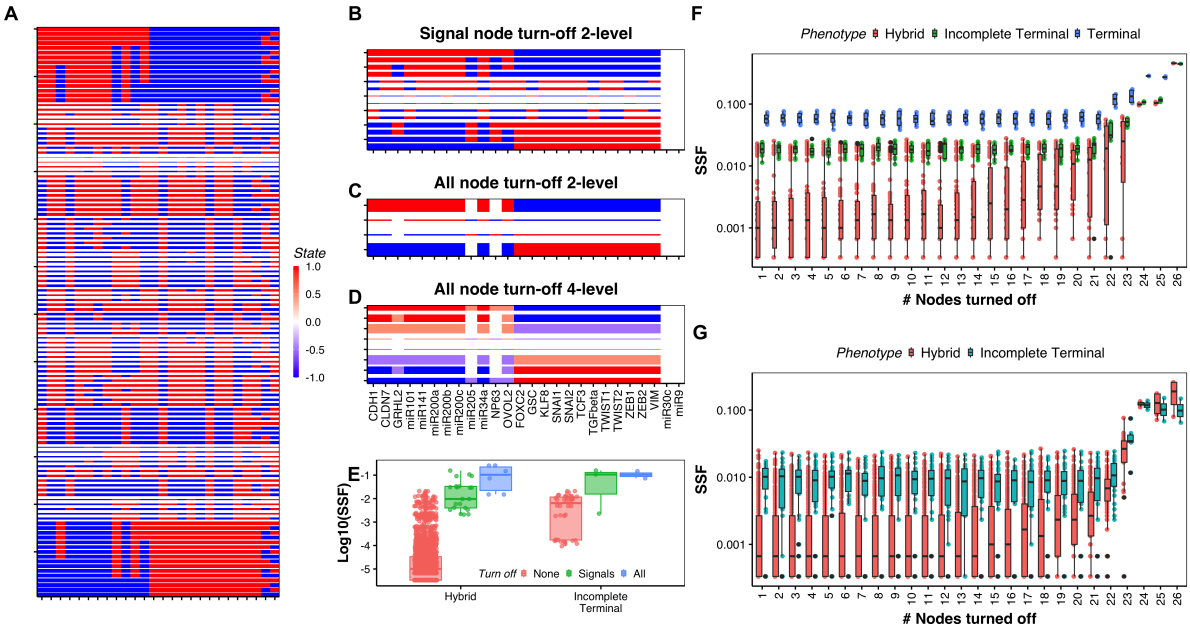

Figure 4: **Node turn-off highlights the stability of hybrid states in multi-level formalism A** Heatmap depicting the configuration of steady states of the 26N100E network (23N90E + signal nodes), simulated with the two-level formalism. Red corresponds to negative expression and blue represents positive expression. The height of each row is proportional to the log of SSF of that steady state. **B** Steady states obtained by applying turn-off to signal nodes with the two-level formalism. The color scale is the same as that of A. **C** Steady states obtained by applying turn-off to all nodes with the two-level formalism. The color scale is the same as that of A. **D** Steady states obtained by applying turn-off to all nodes with the four-level formalism. The color scale is the same as that of A. **E** Boxplots depicting the SSF of steady states of different classes from four-level formalism - without turn-off (red), turn-off applied to signal nodes (green) and turn-off applied to all nodes (blue). **F** Boxplot depicting the evolution of steady state frequency distribution for the sequential turn-off experiment with to optimally increase total frequency of hybrid states. **G** Same as **F**, but to optimally reduce frustration

structure of the network (activates the Epithalial gene GRHL2, and also the Mesenchymal gene SNAI2; inhibited by the Epithelial gene OVOL2 but activated by GRHL2) are often the cause of increase in degeneracy and hybridness in the steady states. We find that the turn-off formalism leads to the NP63 node being turned off in the dominant steady states of the 26N100E EMP network.

The turn-off formalism also reduces the number of states that emerge from four-level formalism to 8 (**Figure 4D**). As most states emergent from the multilevel formalism are weakly stable hybrid states, the turn-off formalism readily eliminates these states. However, the most dominant hybrid states see a nearly 100-fold increase in their frequencies (from the order of 10^{-3} without turn-off to 10^{-1} with the turn-off condition applied to all the nodes - **Figure 4E**). These hybrid states are composed of partial expressions (-0.5 or 0.5) from nearly all the nodes in the network. Furthermore, all six states maintain a coherent expression among the nodes of the same team/phenotype (i.e., all positive or all negative); these are what he have referred to above as "true" hybrid states. Thus, by reducing the degeneracy allowed in the simulation, turn-off formalism can greatly concentrate the steady states of a network. Adding the relaxation of multilevel formalism to the state space allows for the concentration of hybrid states, establishing the stability of these "true" hybrid states that cannot be captured in the two-level formalism.

As we have just shown, applying turn-off to signal nodes alone retained multiple weakly stable states, while applying turn-off to all nodes eliminated most of the weakly stable states in both two and four level models. But we also noticed that nodes like NP63 were turned-off in all steady states despite not being signal nodes. We were therefore interested in asking if it is possible to identify a minimum fraction of nodes to apply turn-off to that can achieve the same results as applying turn-off to all nodes. With both two-level and four-level model, we carried out a "sequential-turn-off" experiment. The simulations are carried out to a maximum of N iterations (N = 26 for the current network). At the first iteration,

we apply turn-off one node at a time, and identify the node, N_1 that gives the maximum "hybridness" calculated as the total frequency of hybrid states. In the next iteration, we apply turn-off to two nodes at a time, to identify the next node N_2 that, when turned off along with N_1 gives the maximum hybridness.

Interestingly we observed an increase in the mean SSF of the terminal and partial terminal phenotypes in both two-level and four-level models (**Figure 4F, G**) along with that of hybrid states. Note that the algorithm only aimed to enhance the hybrid state frequency. Another interesting observation is that while the increase in mean frequency for hybrid states in both two-level and four-level formalisms happened at a similar fraction of nodes turned off, the frequency rose sharply in four-level formalism, while that in two-level formalism saw a gradual rise. We also noticed that from 24 nodes onwards, we found more stable hybrid states than incomplete terminal states in the four-level model. Together, these results highlight that the hybrid states obtained with the four-level formalism are more stable than those obtained from two-level formalism.

2.4 Multilevel formalism reduces phenotypic score of the steady state space

So far we have explored the two-level and four-level formalisms. Since increasing the number of positive and negative levels by one each showed marked improvements in explaining biological data, we explored the effect of further expanding the state space. The corresponding update rules are given below:

$$S_{i_{(t+1)}} = \begin{cases} -1, & if \ \frac{\Sigma_j J_{ji} s_j}{d_i} < -\frac{n-1}{n} \\ -\frac{k}{n}, & if \ -\frac{k}{n} \leq \frac{\Sigma_j J_{ji} s_j}{d_i} < -\frac{k-1}{n} \quad \forall k \in n-1, n-2, ..., 0 \\ \frac{k}{n}, & if \ \frac{k-1}{n} < \frac{\Sigma_j J_{ji} s_j}{d_i} \leq \frac{k}{n} \quad \forall k \in 1, 2, ..., n-1 \\ 1, & if \ \frac{n-1}{n} < \frac{\Sigma_j J_{ji} s_j}{d_i} \\ S_{i_t}, & if \ \frac{\Sigma_j J_{ji} s_j}{d_i} = 0 \end{cases} \tag{3}$$

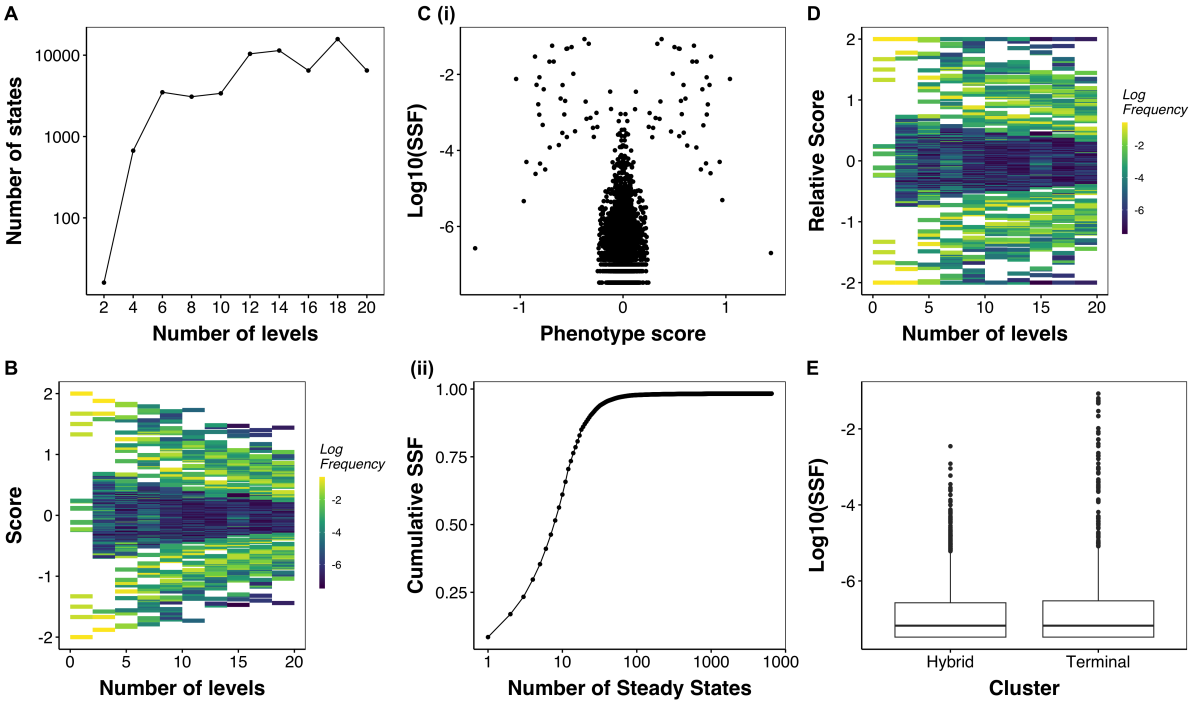

Figure 5: **Increased number of levels in the multi-level model narrows down the range of phenotypic score A** For multilevel formalisms of different numbers of levels, the number steady states. **B** Segment plot depicting the phenotypic score of the steady states with increasing number of levels in the multi-level model. **C** (i) 2-D contour plot of phenotypic score vs SSF for steady states of 23N90E network simulated using 20-level formalism. (ii) Scatterplot depicting the cumulative SSF (y-axis) of the n most frequent steady states (x-axis). **D** Same as B, but the y axis is normalized as $y_{relative} = (\frac{y-min(y)}{max(y)} - 0.5)*4$ at each resolution. **E** Steady state frequency of states clustered by the normalized phenotypic score.

While we observed a significant increase in the number of steady states between 2-level and 4-level formalisms, the number seems to saturate as the number of levels increased, with remaining around 700 after ten levels (**Figure 5A**). As the number of levels increases, we notice a consistent decrease in the range of phenotypic scores of the steady states. While the maximum score of two-level formalism was 2, the same for the twenty-level formalism was close to 1 (**Figure 5B**). Correspondingly, none of the states have a maximum expression of all Epithelial or Mesenchymal nodes. Therefore, the multilevel model does not capture any true "terminal" state. However, states with high stability still appear, with more than 80% of the state space converging to the first 10 steady states ordered by frequency (**Figure 5C**).

To capture the effectively terminal states in the multi-level formalisms, we adopted a new system of steady-state classification. Note that the narrow range of phenotypic scores is a property of the experimental transcriptomic data as well (Figure 1), despite the presence of samples in the dataset characterized as fully epithelial and mesenchymal phenotypes. An appropriate phenotypic classification for this scenario is often obtained using methods such as PCA or clustering together with subject knowledge/assumptions (e.g., the data contains samples of all three phenotypes), allowing the samples to be classified relatively. Therefore, we wanted to investigate a *relative* classification as opposed to an absolute classification. To do so, we clustered the data and classified the clusters into phenotypes based on their phenotypic score. Applying the previous classification with re-scaling the phenotypic scores also would not yield a reasonable classification, because the steady states near the boundaries of such re-scaled samples (phenotypic scores near 2 or -2) have low frequency (**Figure 5D**).

We clustered the steady states into three clusters based on the node expressions. We classified the cluster with a high Epithelial score and low mesenchymal score as Epithelial, a high Mesenchymal score and low Epithelial score as Mesenchymal, and the remaining cluster as Hybrid. The Epithelial and Mesenchymal states had higher mean frequency than the hybrid states (**Figure 5E**).

3 Discussion

Biomolecular interactions are non-linear. The key event of transcriptional regulation - binding of a transcription factor (TF) to the promoter region of the target gene, has switch-like sigmoidal dynamics such that below a threshold concentration of the TF, regulation is much weaker as compared to above the threshold [31]. This threshold behavior allows for the "booleanization" of the expression levels of biomolecules, where "low" expression represents functionally inactive concentrations and "high" expression represents functionally active concentrations. Boolean networks have been used in multiple contexts to understand the dynamics of biological systems. Despite their simplicity, boolean models are able to capture many intricate details of the steady states and dynamics of biological networks. Boolean network models are not concerned with kinetic features of the regulation such as production and degradation rates, the molecular binding strengths, or the dose-dependent dynamics of biomolecular interactions. Instead, the inputs required are the network topology, i.e., the pairwise regulation of transcription factors, and regulatory logic in case of multiple regulators to the same gene [32]. It is therefore important to ask, what additional qualitative properties of the phenotypic landscape are missed by the lack of complexity of boolean models. In this manuscript, we explore the effects of a multi-valued extension of boolean formalism in the specific context of Epithelial-Mesenchymal Plasticity (EMP).

The theory underlying multi-valued networks has been developing in the recent past. One such approach involves the discretization of ODE systems that allow multiple values of expression for each node. An example is the modeling formalism known as DSGRN [33], a comprehensive treatment of the switching system that allows multiple expression levels for a given node based on the out-degree of the node. Using a generic system of ODEs, they define the range of expression levels each node can have and divide the range into multiple levels such that each level indicates active regulation of as many downstream genes. Such approaches have found success in describing the phenotypic landscape of small GRNs, but scale combinatorially with the network size. We analyzed the EMP networks using a similar approach of adding multiple levels to nodes based on their out-degree and found that the results are qualitatively similar to their boolean counterparts (data not shown). Another approach involves adding intermediate, fractional levels of expression such that the highest expression level remains 1. Several studies have aimed to understand the implications of expanding the state space in this manner [34–37], with the discovery of some interesting properties such as the multi-level formalisms seeing lesser instances of limit cycles, capturing steady states using synchronous update scheme which in classic boolean formalism could only be captured by asynchronous update schemes. From a biological standpoint, complexity has often followed context-specific functionality. A classic example is seen in

signaling networks where proteins have different functions at different levels of phosphorylation [38]. Consequently, the selected nodes with the appropriate biological background allow multi-level expression.

Our current extension of boolean formalism aims to explain the biological observation of hybrid phenotypes, while maintaining a formalism that does not require detailed biological information of each node in the network, and therefore can be applied to GRNs out of the current context of EMP. Using the multi-level boolean model that allows two additional levels of expression (-0.5, 0.5), we capture hybrid steady states with a partial expression of E and M nodes. We find that these states have comparable stability to the hybrid states obtained using the classic boolean model, but have lower frustration due to their partial expression. Using another theoretical modification of the boolean formalism, we find that the hybrid states are predominantly generated due to the flexible regulation of unregulated nodes during the simulation. When the effect of such nodes is nullified in the two-level boolean model, hybrid states disappear completely. However, the multi-level model is capable of retaining the hybrid states, thereby indicating a robust mechanism for the emergence of hybrid states.

Expanding the state space also increases the total number of steady states in the network, majority of the increase coming from hybrid states. However, we find that the increase in the number of steady states is not purely due to the increase in the state space. As we further increase the number of levels, the increase in the number of steady states diminishes, saturating after the six-level model for the EMP network simulated here. However, the phenotypic score of the steady states keeps decreasing with increasing number of levels. The states with extreme phenotypic scores cease to be the most dominant states as well. Reduced phenotypic score is a property observed in both ODE-based simulations as well as transcriptomic data normalized according to the maximum expression across samples. It is important to note that the increase in the number of levels pushes the original boolean system towards a continuous regime and the interpretation of the expression levels as "functional" levels before and after the regulation threshold gets weaker. Therefore, at lower levels, the steady states of the multi-level formalism should be treated as discrete, while at higher levels, they can be treated as a continuous system with relative classification.

The development of the multi-level formalism revolves is rooted in explaining hybrid E/M phenotypes. However, it can have applications across multiple contexts. We have previously shown that team-based structure is found in GRNs of multiple cell-fate decision contexts. And in some of these contexts, hybrid states play a significant role. In the developmental context, the hybrid characteristics are often associated with progenitor cell types. High entropy has been utilized as the signature of the differentiability of a sample in transcriptomic data [39]. Hybrid states, especially the ones captured by the multi-level model have high entropy. Similarly, in T cell differentiation, hybrid cells expressing markers of more than one class of T cells have been suggested to be precursors of different T cell types [40]. Thus, the multilevel boolean formalism presented here can meaningfully be applied to a broader set of cell-fate decision networks.

4 Methods

4.1 Threshold-based models

In this manuscript, we used two kinds of models to simulate the EMP network. The first is a class of multi-level threshold-based formalisms described by the equation below:

$$s_{i_{(t+1)}} = \begin{cases} -1, & if \quad \frac{\Sigma_j J_{ji} s_j(t)}{d_i} < -\frac{n-1}{n} \\ -\frac{k}{n}, & if \quad -\frac{k}{n} \leq \frac{\Sigma_j J_{ji} s_j(t)}{d_i} < -\frac{k-1}{n} \quad \forall k \in n-1, n-2, ..., 0 \\ \frac{k}{n}, & if \quad \frac{k-1}{n} < \frac{\Sigma_j J_{ji} s_j(t)}{d_i} \leq \frac{k}{n} \quad \forall k \in 1, 2, ..., n-1 \\ 1, & if \quad \frac{n-1}{n} < \frac{\Sigma_j J_{ji} s_j(t)}{d_i} \\ S_{i_t}, & if \quad \frac{\Sigma_j J_{ji} s_j(t)}{d_i} = 0 \end{cases} \tag{4}$$

Where n is the number of levels defining the state space's granularity. We refer to these models as $2n-level$ models, with the two-level model being the classic boolean formulation. $s_i(t)$ is the state of the ith node at time t, J_{ij} is the regulation of node i by node j, and d_i is the indegree of the ith node. Note that the expression $E = \Sigma_j J_{ji} s_j(t) \in [-d_i, d_i]$. As the expression levels allowed are fractional, division with d_i ensures that $E' = \frac{E}{d_i} \in [-1, 1]$.

The turn-off formalism is obtained by modifying the above equation as follows:

$$s_{i_{(t+1)}} = \begin{cases} -1, \quad if \quad \frac{\Sigma_j J_{ji} s_j(t)}{d_i} < -\frac{n-1}{n} \\ -\frac{k}{n}, \quad if \quad -\frac{k}{n} \leq \frac{\Sigma_j J_{ji} s_j(t)}{d_i} < -\frac{k-1}{n} \quad \forall k \in n-1, n-2, ..., 0 \\ \frac{k}{n}, \quad if \quad \frac{k-1}{n} < \frac{\Sigma_j J_{ji} s_j(t)}{d_i}) \leq \frac{k}{n} \quad \forall k \in 1, 2, ..., n-1 \\ 1, \quad if \quad \frac{n-1}{n} < \frac{\Sigma_j J_{ji} s_j(t)}{d_i} \\ 0, \quad if \quad \frac{\Sigma_j J_{ji} s_j(t)}{d_i} = 0 \end{cases} \tag{5}$$

4.2 Phenotypic score

For simulations, the phenotypic score of a steady state S is defined as:

$$\begin{aligned} Escore &= \Sigma_{i \in E} \frac{S_i}{nE} \\ Mscore &= \Sigma_{j \in M} \frac{S_j}{nM} \\ Score &= Escore - Mscore \end{aligned} \tag{6}$$

where E and M here represent the set of Epithelial and mesenchymal nodes respectively, and nE and nM are the number of nodes in each category.

4.3 Frustration

For a given state S, we define the frustration of an edge as follows:

$$F_{ij} = \begin{cases} 0 \quad if \quad J_{ji} s_j s_i > 0 \\ |s_i s_j| \quad if \quad J_{ji} s_j s_i > 0 \end{cases} \tag{7}$$

The net frustration of the state is defined as the sum of the frustrations of the individual edges

$$F_S = \frac{\Sigma_{i,j} F_{ij}}{E} \tag{8}$$

where E is the total number of edges in the network.

4.4 Transcriptomic data analysis

We obtained bulk-RNAseq data of carcinoma cell-lines from CCLE [30]. We extracted the expression matrix corresponding to an EMT geneset [21] from the cell-line data. We then log-normalized the expression matrix, followed by a z-score normalization. We then calculated the phenotypic scores using the formula in equation 6, using the z-score in place of S_i and further normalised the E and M scores to adjust the range between -1 and 1. For each gene, we define the "intermediate expression" as ranging from 25th to 75th quantile of it's expression vector across all samples. For each sample, we calculate the fraction of Epithelial genes and mesenchymal genes having intermediate levels of expression.

5 Data and Code availability

The raw network files and the code used for network simulation, analysis and figure generation can be found at https://github.com/askhari139/MultilevelPaper. The simulations are performed using Julia and the corresponding code can be found at https://github.com/Boolean.jl.

6 Funding

MKJ was supported by the Ramanujan Fellowship (SB/S2/RJN-049/2018) awarded by the Science and Engineering Research Board (SERB), Department of Science and Technology, Government of India. MKJ was also supported by Param Hansa Philanthropies. KH and HL were supported by the Center for Theoretical Biological Physics, NSF PHY-2019745.

7 Conflict of Interest

The authors declare no competing financial or non-financial interests.

8 Author contributions statement

HL and ST designed the research. KH, ST, and VA carried out the simulations. All authors discussed the results and participated in the preparation of the manuscript.

References

1. Dillekås, H., Rogers, M. S. & Straume, O. Are 90% of deaths from cancer caused by metastases? *Cancer Medicine* **8,** 5574–5576. ISSN: 2045-7634. http://dx.doi.org/10.1002/cam4.2474 (Aug. 2019).
2. Welch, D. R. & Hurst, D. R. Defining the Hallmarks of Metastasis. *Cancer Research* **79,** 3011–3027. ISSN: 1538-7445. http://dx.doi.org/10.1158/0008-5472.CAN-19-0458 (June 2019).
3. Huang, Y., Hong, W. & Wei, X. The molecular mechanisms and therapeutic strategies of EMT in tumor progression and metastasis. *Journal of Hematology & Oncology* **15.** ISSN: 1756-8722. http://dx.doi.org/10.1186/s13045-022-01347-8 (Sept. 2022).
4. Jolly, M. K., Mani, S. A. & Levine, H. Hybrid epithelial/mesenchymal phenotype(s): The 'fittest' for metastasis? *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer* **1870,** 151–157. ISSN: 0304-419X. http://dx.doi.org/10.1016/j.bbcan.2018.07.001 (Dec. 2018).
5. Kröger, C. *et al.* Acquisition of a hybrid E/M state is essential for tumorigenicity of basal breast cancer cells. *Proceedings of the National Academy of Sciences* **116,** 7353–7362. ISSN: 1091-6490. http://dx.doi.org/10.1073/pnas.1812876116 (Mar. 2019).
6. Pastushenko, I. & Blanpain, C. EMT Transition States during Tumor Progression and Metastasis. *Trends in Cell Biology* **29,** 212–226. ISSN: 0962-8924. http://dx.doi.org/10.1016/j.tcb.2018.12.001 (Mar. 2019).
7. Saxena, K., Jolly, M. K. & Balamurugan, K. Hypoxia, partial EMT and collective migration: Emerging culprits in metastasis. *Translational Oncology* **13,** 100845. ISSN: 1936-5233. http://dx.doi.org/10.1016/j.tranon.2020.100845 (Nov. 2020).
8. Quan, Q. *et al.* Cancer stem-like cells with hybrid epithelial/mesenchymal phenotype leading the collective invasion. *Cancer Science* **111,** 467–476. ISSN: 1349-7006. http://dx.doi.org/10.1111/cas.14285 (Jan. 2020).
9. Aceto, N. *et al.* Circulating Tumor Cell Clusters Are Oligoclonal Precursors of Breast Cancer Metastasis. *Cell* **158,** 1110–1122. ISSN: 0092-8674. http://dx.doi.org/10.1016/j.cell.2014.07.013 (Aug. 2014).
10. Yamamoto, A., Doak, A. E. & Cheung, K. J. Orchestration of Collective Migration and Metastasis by Tumor Cell Clusters. *Annual Review of Pathology: Mechanisms of Disease* **18,** 231–256. ISSN: 1553-4014. http://dx.doi.org/10.1146/annurev-pathmechdis-031521-023557 (Jan. 2023).
11. Jolly, M. K. *et al.* Hybrid epithelial/mesenchymal phenotypes promote metastasis and therapy resistance across carcinomas. *Pharmacology & Therapeutics* **194,** 161–184. ISSN: 0163-7258. http://dx.doi.org/10.1016/j.pharmthera.2018.09.007 (Feb. 2019).
12. Sahoo, S. *et al.* Immunosuppressive Traits of the Hybrid Epithelial/Mesenchymal Phenotype. *Frontiers in Immunology* **12.** ISSN: 1664-3224. http://dx.doi.org/10.3389/fimmu.2021.797261 (Dec. 2021).
13. Debnath, P., Huirem, R. S., Dutta, P. & Palchaudhuri, S. Epithelial–mesenchymal transition and its transcription factors. *Bioscience Reports* **42.** ISSN: 1573-4935. http://dx.doi.org/10.1042/BSR20211754 (Dec. 2021).
14. Steinway, S. N. *et al.* Network Modeling of TGF-b Signaling in Hepatocellular Carcinoma Epithelial-to-Mesenchymal Transition Reveals Joint Sonic Hedgehog and Wnt Pathway Activation. *Cancer Research* **74,** 5963–5977. ISSN: 1538-7445. http://dx.doi.org/10.1158/0008-5472.CAN-14-0225 (Oct. 2014).

15. Hari, K., Ullanat, V., Balasubramanian, A., Gopalan, A. & Jolly, M. K. Landscape of epithelial–mesenchymal plasticity as an emergent property of coordinated teams in regulatory networks. *eLife* **11.** ISSN: 2050-084X. http://dx.doi.org/10.7554/eLife.76535 (Oct. 2022).

16. Rashid, M., Hari, K., Thampi, J., Santhosh, N. K. & Jolly, M. K. Network topology metrics explaining enrichment of hybrid epithelial/mesenchymal phenotypes in metastasis. *PLOS Computational Biology* **18** (ed Faeder, J. R.) e1010687. ISSN: 1553-7358. http://dx.doi.org/10.1371/journal.pcbi.1010687 (Nov. 2022).

17. Barbuti, R., Gori, R., Milazzo, P. & Nasti, L. A survey of gene regulatory networks modelling methods: from differential equations, to Boolean and qualitative bioinspired models. *Journal of Membrane Computing* **2,** 207–226. ISSN: 2523-8914. http://dx.doi.org/10.1007/s41965-020-00046-y (Sept. 2020).

18. Lu, M., Jolly, M. K., Levine, H., Onuchic, J. N. & Ben-Jacob, E. MicroRNA-based regulation of epithelial–hybrid–mesenchymal fate determination. *Proceedings of the National Academy of Sciences* **110,** 18144–18149. ISSN: 1091-6490. http://dx.doi.org/10.1073/pnas.1318192110 (Oct. 2013).

19. Huang, B. *et al.* Interrogating the topological robustness of gene regulatory circuits by randomization. *PLOS Computational Biology* **13** (ed Tang, C.) e1005456. ISSN: 1553-7358. http://dx.doi.org/10.1371/journal.pcbi.1005456 (Mar. 2017).

20. Font-Clos, F., Zapperi, S. & La Porta, C. A. M. Topography of epithelial–mesenchymal plasticity. *Proceedings of the National Academy of Sciences* **115,** 5902–5907. ISSN: 1091-6490. http://dx.doi.org/10.1073/pnas.1722609115 (May 2018).

21. Byers, L. A. *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. *Clinical Cancer Research* **19,** 279–290. ISSN: 1557-3265. http://dx.doi.org/10.1158/1078-0432.CCR-12-1558 (Jan. 2013).

22. Mendik, P. *et al.* Translocating proteins compartment-specifically alter the fate of epithelial-mesenchymal transition in a compartmentalized Boolean network model. *npj Systems Biology and Applications* **8.** ISSN: 2056-7189. http://dx.doi.org/10.1038/s41540-022-00228-7 (June 2022).

23. Tripathi, S., Kessler, D. A. & Levine, H. Biological networks regulating cell fate choice are minimally frustrated. *Physical Review Letters* **125,** 088101 (2020).

24. Tian, X.-J., Zhang, H. & Xing, J. Coupled reversible and irreversible bistable switches underlying TGFβ-induced epithelial to mesenchymal transition. *Biophysical journal* **105,** 1079–1089 (2013).

25. Jolly, M. K. *et al.* Stability of the hybrid epithelial/mesenchymal phenotype. *Oncotarget* **7,** 27067–27084. ISSN: 1949-2553. http://dx.doi.org/10.18632/oncotarget.8166 (Mar. 2016).

26. George, J. T., Jolly, M. K., Xu, S., Somarelli, J. A. & Levine, H. Survival outcomes in cancer patients predicted by a partial EMT gene expression scoring metric. *Cancer research* **77,** 6415–6428 (2017).

27. Yu, M. *et al.* Circulating breast tumor cells exhibit dynamic changes in epithelial and mesenchymal composition. *science* **339,** 580–584 (2013).

28. Puram, S. V. *et al.* Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. *Cell* **171,** 1611–1624 (2017).

29. Lüönd, F. *et al.* Distinct contributions of partial and full EMT to breast cancer malignancy. *Developmental cell* **56,** 3203–3221 (2021).

30. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483,** 603–607. ISSN: 1476-4687. http://dx.doi.org/10.1038/nature11003 (Mar. 2012).

31. Mirny, L. A. Nucleosome-mediated cooperativity between transcription factors. *Proceedings of the National Academy of Sciences* **107,** 22534–22539. ISSN: 1091-6490. http://dx.doi.org/10.1073/pnas.0913805107 (Dec. 2010).

32. Rozum, J. C., Campbell, C., Newby, E., Nasrollahi, F. S. F. & Albert, R. Boolean Networks as Predictive Models of Emergent Biological Behaviors. https://arxiv.org/abs/2310.12901 (2023).

33. Cummins, B., Gedeon, T., Harker, S. & Mischaikow, K. DSGRN: Examining the Dynamics of Families of Logical Models. *Frontiers in Physiology* **9.** ISSN: 1664-042X. http://dx.doi.org/10.3389/fphys.2018.00549 (May 2018).

34. Kadelka, C., Li, Y., Kuipers, J., Adeyeye, J. O. & Laubenbacher, R. Multistate nested canalizing functions and their networks. *Theoretical Computer Science* **675,** 1–14. ISSN: 0304-3975. http://dx.doi.org/10.1016/J.TCS.2017.01.031 (May 2017).

35. Bull, L. Nonbinary Representations in the NK and NKCS Models. *Complex Systems* **31,** 87–101. ISSN: 0891-2513. http://dx.doi.org/10.25088/ComplexSystems.31.1.87 (Mar. 2022).

36. Bull, L. Evolving Multi-valued Regulatory Networks on Tunable Fitness Landscapes. *Complex Systems* **32,** 289–307. ISSN: 0891-2513. http://dx.doi.org/10.25088/ComplexSystems.32.3.289 (Dec. 2023).

37. Galofre, J. G., Millán, M. P., Rial, A. G., Laubenbacher, R. & Dickenstein, A. Beyond Boolean networks, a multi-valued approach. https://arxiv.org/abs/2404.16760 (2024).

38. Mbodj, A., Junion, G., Brun, C., Furlong, E. E. M. & Thieffry, D. Logical modelling of Drosophila signalling pathways. *Molecular BioSystems* **9,** 2248. ISSN: 1742-2051. http://dx.doi.org/10.1039/C3MB70187E (2013).

39. Teschendorff, A. E. & Enver, T. Single-cell entropy for accurate estimation of differentiation potency from a cell's transcriptome. *Nature Communications* **8.** ISSN: 2041-1723. http://dx.doi.org/10.1038/ncomms15599 (June 2017).

40. Duddu, A. S. *et al.* Multistability and predominant double-positive states in a four node mutually repressive network: a case study of Th1/Th2/Th17/T-reg differentiation. http://dx.doi.org/10.1101/2024.01.30.575880 (Feb. 2024).