npj | systems biology and applications

www.nature.com/npjsba

ARTICLE OPEN

Check for updates

The nonlinearity of regulation in biological networks

Santosh Manicka[1,3], Kathleen Johnson[2,3], Michael Levin[1] and David Murrugarra[2]✉

The extent to which the components of a biological system are (non)linearly regulated determines how amenable they are to therapy and control. To better understand this property termed "regulatory nonlinearity", we analyzed a suite of 137 published Boolean network models, containing a variety of complex nonlinear regulatory interactions, using a probabilistic generalization of Boolean logic that George Boole himself had proposed. Leveraging the continuous-nature of this formulation, we used Taylor decomposition to approximate the models with various levels of regulatory nonlinearity. A comparison of the resulting series of approximations of the biological models with appropriate random ensembles revealed that biological regulation tends to be less nonlinear than expected, meaning that higher-order interactions among the regulatory inputs tend to be less pronounced. A further categorical analysis of the biological models revealed that the regulatory nonlinearity of cancer and disease networks could not only be sometimes higher than expected but also be relatively more variable. We show that this variation is caused by differences in the apportioning of information among the various orders of regulatory nonlinearity. Our results suggest that there may have been a weak but discernible selection pressure for biological systems to evolve linear regulation on average, but for certain systems such as cancer, on the other hand, to simultaneously evolve more nonlinear rules.

npj Systems Biology and Applications (2023)9:10 ; https://doi.org/10.1038/s41540-023-00273-w

INTRODUCTION

How nonlinear is the regulation of the components of biological networks? That is, to what extent do the biochemical components of these networks non-independently interact (Fig. 1) in influencing downstream processes and network behavior. Research on the "nonlinearity" front has hitherto focused on its various manifestations in the dynamics of biological systems, such as chaos, bifurcation, multistability, synchronization, patterning, dissipation, etc.[1], but a characterization of regulatory nonlinearity among the components of the underlying systems that give rise to those phenomena is lacking. A more complete understanding of biological regulatory nonlinearity would not only yield insights into their design principles[2] but also have theoretical implications ranging from canalization to control[3,4] and practical implications for biomedical therapy, synthetic biology, etc.[1,5]. A good example of this concerns the mapping between molecular or genetic information and the resulting system-level anatomical structure and function of an organism. Advances in regenerative medicine and synthetic morphology require rational control of physiological and anatomical outcomes[6], but progress in genetics and molecular biology produce methods and knowledge targeting the lowest-level cellular hardware. There is no one-to-one mapping from genetic information to tissue- and organ-level structure; similarly, ion channels open and close post-translationally, driving physiological dynamics that are not readily inferred from proteomic or transcriptomic data. System-level properties in biology are often highly emergent, with gene-regulatory or bioelectric circuit dynamics connecting initial state information and transition rules to large-scale structure and function. Thus, the difficult inverse problem[7] of inferring outcomes and desirable interventions across scales of biology illustrates some of the fundamental questions about the directness or nonlinearity of encodings of information, as well as the importance of this question for practical advances in biomedicine and bioengineering that exploit the plasticity and robustness of cellular collectives. Many deep questions remain about the potential limitations and best strategies to bridge scales for prediction and control in developmental, evolutionary, and cell biology. To that end, we introduce here a formal characterization of the nonlinearity of models of biological regulatory networks, such as those often used to describe relationships between regulatory genes. Specifically, we consider a class of discrete models of biological regulatory systems called "Boolean models" that are known for their relative simplicity and tractability compared to continuous ordinary differential equation-based (ODE) models[8].

A Boolean network is a discrete network model characterized by the following features. Each node in a Boolean network can only be in one of two states, ON or OFF, which represents the expression or activity of that node. The state of a node depends on the states of other input nodes which are represented as a Boolean rule of these input nodes. Many of the available Boolean network models were created via literature search of the regulatory mechanisms and subsequently validated via experiments[9]. Some of the publicly available models were generated via network inference methods from time course data[3].

Previous studies have found that certain characteristic features of the biological Boolean models, such as the mean in-degree, output bias, sensitivity and canalization, tend to assume an optimal range of values that support optimal function[10,11]. Here we study a new but generic feature of complex systems termed "regulatory nonlinearity" that we broadly define as the degree to which the inputs to the components of a complex interact. To characterize the regulatory nonlinearity of Boolean networks we formalize an approach to generalizing Boolean logic by casting it as a form of probability, which was originally proposed by George Boole himself[12]. We leverage the continuous nature of the resulting polynomials to decompose a Boolean function using Taylor-series and reveal the distinct layers of its regulatory nonlinearity (Fig. 5). Various other methods, both discrete and continuous, of decomposing Boolean functions exist, such as

[1]Department of Biology, Tufts University, Medford, MA 02155, USA. [2]Department of Mathematics, University of Kentucky, Lexington, KY 40506, USA. [3]These authors contributed equally: Santosh Manicka, Kathleen Johnson. ✉email: murrugarra@uky.edu

Reed-Muller, Walsh spectrum, Fourier, discrete Taylor and fuzzy logic[13–16]. Our continuous Taylor decomposition method is distinct in that it offers a clear and systematic way to characterize nonlinearity.

Is the regulatory nonlinearity of biological systems special? To that end we specifically ask: (1) how well could biological Boolean models be approximated, that is, faithfully represented with only partial information containing lower levels of nonlinearity relative to that of the original?; (2) is there an optimal level of regulatory nonlinearity, characterized by maximum approximability, that these models may have been selected for by evolution?; and (3) do different classes of biological networks show characteristically different optimal levels of regulatory nonlinearity? To answer these questions, we first approximate the biological models by systematically composing the various nonlinear layers resulting in a sequence of model-approximations with increasing levels of nonlinearity. We then estimate the accuracy of these approximations by comparing the outputs of their simulations with that of the original unapproximated model. We then construct an appropriate random ensemble for each biological model and compare their mean accuracies for fixed levels of approximation. The main idea is that a biological model that is more approximable than expected for a particular level of nonlinearity would mean that the network may have been optimized for that level nonlinearity. Finally, we classify the biological networks into various categories and compare their approximabilities to identify any category-dependent effects.

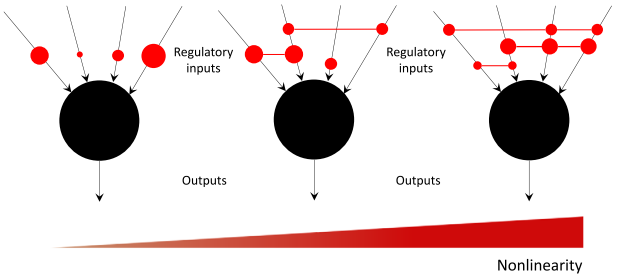


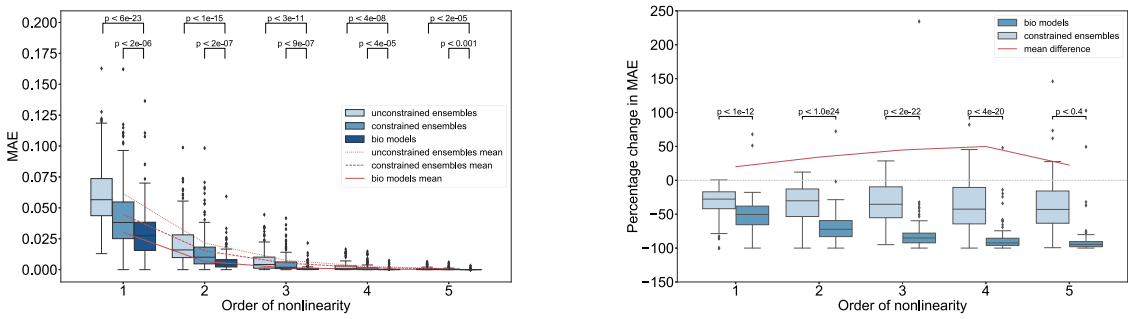
Fig. 1 An illustration of the concept of regulatory nonlinearity. Each black circle represents a generic biochemical component such as a gene, transcription factor, enzyme, etc., regulated by a set of inputs (also biochemical components) and generates a generic output such as concentration level, strength, etc. Non-zero interactions among the inputs are represented by red circles connected by red lines, with the total number of possible interactions for a node with k inputs equal to $\sum_{l=1}^{k} \binom{k}{l}$. The size of the red circles and the width of the connecting lines represents the weight of the interactions. Independent inputs are represented by unconnected red circles. The degree of nonlinearity would thus be expected to increase, though not necessarily linearly, from left to right, as the numbers and the strengths of the regulatory interactions increase. One could also visualize these local interactions in a broader network-context as "hypergraphs"[46].

RESULTS

The regulation of biological models tends to be less nonlinear than expected by chance

The mean approximation error (MAE; see "Methods") of biological models tends to be less than both the constrained and the unconstrained random ensembles across all approximation orders (Fig. 2a). At the linear order, for example, the mean MAE of biological models is about 0.025, whereas the mean MAE of the constrained ensembles is twice as large at about 0.05 ($p < 10^{-6}$) and even larger for the unconstrained ensembles at about 0.07 ($p < 10^{-23}$). This ordering of MAE extends to all approximation orders (Fig. 2a). Even though the mean MAE of the unconstrained ensembles tends to be remarkably low (0.07 at worst), the percentage change in MAE (PMAE; see "Methods") convey a more accurate sense of approximability. For example, the MAE of the biological models and the constrained ensembles are respectively about 50 and 25% lower than the baseline expectation for the linear order (Fig. 2b). This means that the biological models are about 25% more linearly approximable than expected by chance, an effect that amplifies with higher approximation orders going up to 50% for order 4, since the PMAE of biological models shrinks faster than the constrained ensembles (Fig. 2b).

A possible explanation for the observed trends is that the Taylor spectrums of the biological models tend to be lopsided and more clustered at the linear ends (see Fig. 4 for an example), dramatically reducing the load on the higher orders and thus resulting in a faster growth of approximability with higher orders. This effect may be less pronounced in constrained ensembles, and even less in unconstrained ensembles, as the apportioning in the corresponding Taylor spectrums may be less lopsided and more uniform in comparison. Based on these observations we hypothesize, but do not conclude, that biological systems may have been subjected to a slight but discernible selection pressure for developing less nonlinear regulatory rules. This hypothesis is further supported by the fact that the approximability



(a) (b)

Fig. 2 Biological models are more approximable than expected by chance. a The MAE of the biological models and the associated constrained and unconstrained ensembles for approximation orders 1 to 5; MAE values for orders 6 and above are negligible and not shown. **b** Percentage change in MAE for the biological models and the associated constrained ensembles computed with respect to the MAE of the corresponding unconstrained ensembles. Every point in the distributions corresponding to the random ensembles represents the average MAE of an ensemble of 100 random networks associated with each biological model. The p values indicate the statistical significance of the difference in mean MAEs between sets of random ensembles and the biological models for a given order of nonlinearity. Statistical analysis by Welch's unequal variances t-test.

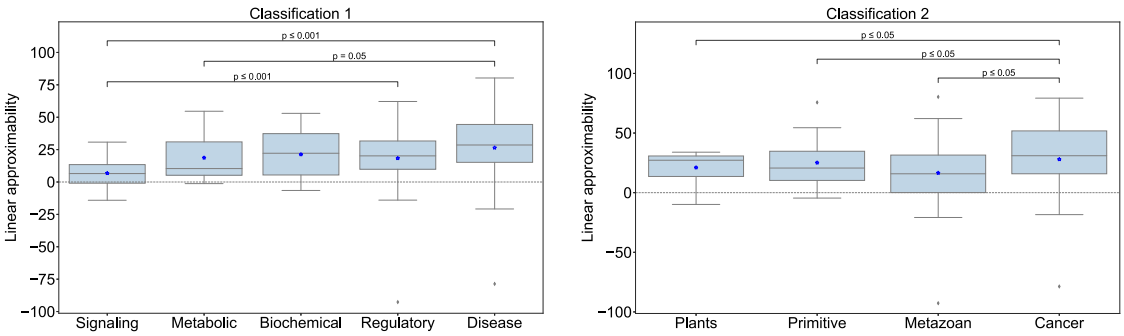


Fig. 3 The linear approximabilities of various categories of biological models. a Classification 1 (C1); **b** Classification 2 (C2). The categories in either classification are displayed in increasing order of variance. Each box represents the distribution of the linear approximabilities of the corresponding category. The *p* values indicate the statistical significance of the difference in the variance between pairs of categories; only the *p* values of significantly different pairs are shown. Statistical analysis by *F*-test of equality of variances.

systematically increases as more biological constraints are imposed, as one goes from unconstrained to constrained to the biological models (Fig. 2). This has implications not only for the feasibility of biomedical approaches to control emergent somatic complexity or guided self-assembly of novel forms[17], but also for models of anatomical homeostasis and evolvability: linearity implies easier control of its own complex processes by any biological system, and more efficient credit assignment during evolution.

Random models with minimal biological constraints tend to be naturally approximable

The MAE of the unconstrained random ensembles are remarkably low with values averaging around 0.07 and not exceeding 0.17 at the linear order; these values only decrease further with higher approximation orders (Fig. 2a). These observations suggest that even random Boolean models with minimal biological constraints tend to be considerably approximable. In other words, approximability may be a natural property of random Boolean ensembles to some extent and not necessarily a special property of biological models. Furthermore, the MAE has its inflection point at order 2 (Fig. 2a), meaning that the approximability starts saturating at that order. We hypothesize based on these observations that regulatory information beyond the second order may be inconsequential with regards to the statistics of network dynamics in general.

The above hypothesis has an analog in the realm of Boolean Ising models, where it was found that maximum entropy (MaxEnt) models with only pairwise interactions were sufficient to fit random multivariate Ising networks that are densely connected and whose state spaces satisfy certain entropy constraints, features that many biological systems share[18]. There are however important differences between our results and theirs: (1) even though the Taylor expansions resemble MaxEnt expressions, the latter fit global state-space distributions, whereas our Taylor polynomials are local formulations; and (2) whereas our Taylor polynomials are built on derivatives of the states, MaxEnt models are based on the raw values. Notwithstanding these potentially superficial differences, the analogous results are striking and calls for further research on whether the respective explanations are similar if not fundamentally the same. Besides, the above observations are reminiscent of the concept of model degeneracy or "sloppiness", where several models (defined by unique sets of parameters) explain the same biological phenomenon due to redundant parameters[19]. In our case, it's not the parameters but the higher-order relationships among them that are redundant, as they contribute minimally to the MAE beyond a certain order of approximation (Fig. 2a). Future research will determine whether there's a deep connection between the sloppiness of the parameters and the orders of relationships among them.

The approximability of a biological model depends on its class, with the cancer family displaying the most variability

Even though the nonlinearity of biological networks is less than expected on average, individual and category-dependent variations were observed. In the following, we focus on the "linear approximability" (PMAE corresponding to the linear order) of the biological models since it's hypothesized to be the cause of the observed approximabilities in the higher order; values are listed in Supplementary Table 1. First, there are a few networks in almost every category that are more nonlinear than expected, as evidenced by the negative linear approximability values (Fig. 3). Second, the disease networks in C1 and the cancer networks in C2 are the ones with the most linear approximability, with values ≥75% (Fig. 3). In other words, the cancer or disease pathways tend to be more optimized for linearity compared to the other categories. This makes sense since a more linear pathway is more amenable to control, which presumably works in favor of the agents of disease. On the other hand, the disease and cancer networks also display the highest variability in their linear approximability compared to other categories ($p < 0.05$ in all comparisons for the cancer models), with the PMAE values ranging between −92 and 80% (Fig. 3). In other words, disease and cancer models could be either relatively extremely more linear or nonlinear than expected, depending on the specific cases. Note that about 56% of the disease category comprises of non-cancer networks (31/55), which suggests that the effect is not significantly biased by cancer networks. These observations suggest that regulatory nonlinearity may offer an effective "entry point" to the agents of disease by virtue of its natural heterogeneity that they could leverage to their advantage perhaps as a means to evade treatment since there's no single level of nonlinearity to target. This heterogeneity may also have a connection to one of the hallmarks of cancer, namely genetic heterogeneity[20] where the cancerous cells within an individual display heterogeneous gene expression compared to the homogeneous expression in the healthy cells. In the case of nonlinearity, the heterogeneity manifests at the population level, raising the question of whether it may also be observed at the level of single cells within an individual. In other words, could the heterogeneity of nonlinearity be yet another hallmark of cancer?

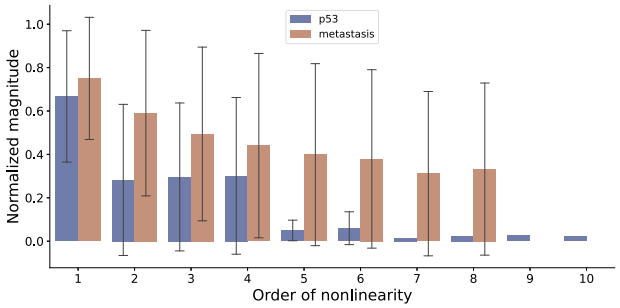


Fig. 4 A comparison of the spectrums of the magnitudes of nonlinearity of the models corresponding to the two extreme cases of approximability. The bars represent the mean absolute value of all the Taylor derivatives of the corresponding order, averaged over all the nodes of the corresponding model containing those terms. The maximum orders (in-degrees) of P53 and Metastasis are respectively 10 and 8. To compare the derivatives from different models they were normalized with respect to the maximum possible absolute value of a Taylor derivative of order $|\alpha|$ (Eq. (6)) of a Boolean function with output bias p, given by $(\min(p, 0.5) - \max(p - 0.5, 0))2^{|\alpha|}$ (proof provided in the "Methods" section). The error bars represent the standard deviation and not the confidence intervals of the means since they are not estimates. The errors appear large whenever the corresponding distribution of the magnitudes are bimodal with many values clustered close to 0.

The shape of the Taylor spectrum explains the extreme opposite characters of linear approximability of a pair of cancer models

Why are some models more linearly approximable and others less? The answer lies in the organization of the corresponding Taylor decompositions, as described above. To illustrate this in detail, we compared the Taylor decompositions of a pair of models chosen from among the most extreme outliers of linear approximability in either direction (Fig. 3). Those models are respectively the following: a linear model describing the role of the protein p53 in the regulation of cell-cycle arrest in breast cancer[21], henceforth referred to as the "P53" model; and a nonlinear model describing the role of mutations in the regulation of metastasis in lung cancer[22], henceforth referred to as the "Metastasis" model. P53 has a linear approximability of about 72%, and it consists of 16 nodes with a mean in-degree of 3.8 ± 2.4 and a mean output bias of 0.38 ± 0.14; and, Metastasis has a linear approximability of about −79%, and it consists of 32 nodes with a mean in-degree of 4.9 ± 2.5 and a mean output bias of 0.27 ± 0.26. Thus, while P53 is smaller and sparser than Metastasis, its nodes exhibit more output-uncertainty compared to Metastasis. According to the mean field theory of random Boolean networks[23], the opposite characters of the mean in-degree and the output bias of these models means that their dynamical behaviors could be expected to be similar (although with the caution that the theory was originally developed for infinite-sized and homogeneously connected networks, which is not the case here). However, we know that their linear approximabilities, which is another expression of dynamical behavior, are opposites. One explanation for this discrepancy lies in the distinct apportioning of nonlinearity in their respective Taylor decompositions (Fig. 4). Specifically, while the magnitude of nonlinearity, defined as the mean absolute value of the Taylor derivatives for a given order and normalized appropriately (see text of Fig. 4), tends to be clustered around the linear order for P53, they are relatively more spread out for Metastasis. Moreover, while the magnitude of the linear order for P53 is more than twice as large the next largest magnitude at lower orders the corresponding ratios for Metastasis are relatively smaller, thus explaining why P53 is more linearly approximable than Metastasis. This result is consistent with predictions based on a model of scaling of cellular control policies[24]. A more controllable (linear) network (P53) is optimal for cooperation with other cells toward collective (normal morphogenetic) goals. In contrast, a cell defecting from the collective and reverting to a more unicellular lifestyle (Metastasis) should exhibit a less predictable, controllable network due to pressures from parasites and competitors that independent unicellular organisms face. Methods for calculating controllability (e.g., linearity) are an important addition to recent efforts to solve the conundrum of interpretability of information structures in contexts ranging from machine learning to evolutionary developmental biology[25–27].

DISCUSSION

This paper introduces the concept of regulatory nonlinearity as a characteristic of Boolean networks. There are several other related characterizations of Boolean networks such as canalization[28], effective connectivity[11], symmetry[29] and controllability[30]. It has been previously reported that the levels of canalization (a measure of the extent to which a subset of inputs actually influence the outputs of a Boolean function) and the mean effective connectivity (a measure of collective canalization) are high in biological networks[3,11]. It has also been found that biological networks need few inputs to reprogram[31] and are relatively easier to control[4]. Our formulation of regulatory nonlinearity is related to these other measures in that more linearity implies more apportioning of influence to individual inputs rather than collective sets of inputs (Fig. 4). Hence, we hypothesize that regulatory nonlinearity may serve the purpose of controllability and epigenetic stability[32].

Our results further moot the possibility that regulatory nonlinearity may be a factor underlying more powerful dynamical phenomena such as memory[33] and computation, defined as the capacity for adaptive information-processing[34]. Even though there's increasing consensus that biological systems contain memory and perform computation, clarity is lacking as to what features of those systems enable it and what general principles underlie it[34,35]. Our framework of regulatory nonlinearity offers an approach to answering these questions. For example, one could consider a known dynamical model with a capacity for memory[33] or universal computation such as the elementary cellular automaton (ECA) driven by rule 110[36] and ask if there are unique properties of their Taylor spectrums that confer their respective capabilities. Present approaches to answering this question typically consists of characterization of the dynamical behavior and not the rules[33,37,38]. A characterization of the rules especially makes sense for ECA[39] since the structure is the same (lattice) and the only feature that distinguishes one ECA from the other is the rule.

Looking at such questions from an even broader perspective it becomes evident that they are only instances of the ultimate puzzle of complex systems, namely what connects the structure and the function of a system. Even though recent work has attempted to answer this question from the perspective of the rules or dynamical laws that govern the system[11,39,40], more tools are needed[41]. In that regard, our framework of regulatory nonlinearity could be a novel addition to this burgeoning toolkit in that it could also be applied to continuous differentiable models of biological networks such as those based on differential equations.

The main limitation of our formulation of approximability is that the approximation accuracy will necessarily increase with higher orders of approximation since the Taylor terms would be accumulated with each higher order (the highest order of approximation is exact). However, this does not affect the falsifiability of our framework since it's possible, for example, to construct networks with XOR-like functions that would be clearly less linearly approximable than the associated ensembles. The

Metastasis model is another example in that regard (Fig. 4). Furthermore, though our experiments control for regulatory nonlinearity they don't offer insights into how it dynamically interacts with other network features such as connectivity in generating the observed approximability. For example, the network structure may be expected to determine which states the network enters into at any given point that in turn determines which parts of the Taylor compositions are actually utilized and how. Moreover, it's known that certain characteristics of the local regulatory rules, such as "effective connectivity", dynamically modulate the network structure itself[42]. Thus, more research is required to investigate the extent to which the structure and output bias facilitates the high approximability of biological networks and vice versa, for which we already have produced preliminary results by way of the constrained models (Fig. 2). Lastly, our conclusions about the linearity of biological regulatory networks may be a reflection of a hidden bias built in the inference methods that produced the models in the first place. We leave it to future work to explore these realms.

METHODS

Probabilistic generalization of Boolean logic

Here we provide a continuous-variable formulation of a Boolean function by casting Boolean values as probabilities, thus transforming it into a pseudo-Boolean function[43]. Consider random variables $X_i : \{0,1\} \to [0,1], i = 1, \ldots, n$, with Bernoulli distributions. That is, $p_i = Pr(X_i = 1) = 1 - Pr(X_i = 0) = 1 - q_i$, for $i = 1, \ldots, n$. Let $X = X_1 \times \cdots \times X_n$ be the product of random variables and $f : X \to \{0,1\}$ a Boolean function. Let $R_0^f = \{x \in X : f(x) = 0\}$ and $R_1^f = \{x \in X : f(x) = 1\}$. Note that X is a disjoint union of R_0^f and R_1^f. Then:

$$Pr(f = 1) = Pr(R_1^f) = \sum_{x \in R_1^f} Pr(x) = \sum_{x \in R_1^f} \prod_{i=1}^{n} \hat{p}_i \tag{1}$$

where $\hat{p}_i = p_i$ if $x_i = 1$ and $\hat{p}_i = 1 - p_i$ if $x_i = 0$. Let $\hat{f}(p_1, \ldots, p_n) = \sum_{x \in R_1^f} \prod_{i=1}^{n} \hat{p}_i$. Thus, $\hat{f} : [0,1]^n \to [0,1]$ is a continuous-variable function. The following theorem shows that $\hat{f}$ is a generalization of f in the sense that $\hat{f}(x) = f(x)$ for all $x \in \{0,1\}^n$.

Theorem. *For discrete values of $x_i \in \{0,1\}, i = 1, \ldots, n$, we have $\hat{f}(x_1, \ldots, x_n) = f(x_1, \ldots, x_n)$.*

Proof. Let $z = (z_1, \ldots, z_n) \in \{0,1\}^n$. Since each z_i is either 0 or 1, we have that $p_i = 1$ if $z_i = 1$ or $p_i = 0$ if $z_i = 0$ for $i = 1, \ldots, n$. We want to show that $\hat{f}(p_1, \ldots, p_n) = f(z_1, \ldots, z_n)$. Since $X = R_0^f \cup R_1^f$, we have that either $z \in R_0^f$ or $z \in R_1^f$. If $z \in R_1^f$, then $f(z) = 1$ and $Pr(z) = \prod_{i=1}^{n} \hat{p}_i = 1$. Moreover, for any other $x \in R_1^f$ with $x \neq z$ we have that $Pr(x) = 0$. Thus, $\hat{f}(z) = \sum_{x \in R_1^f} Pr(x) = Pr(z) = 1$. Now if $z \in R_0^f$, then $\hat{f}(z) = 0$ because $\sum_{\emptyset} = 0$. Thus, $\hat{f}(x) = f(x)$ for all $x \in \{0,1\}^n$.□

Corollary. *If $p_i = 1/2$ for all $i = 1, \ldots, n$, then $\hat{f}(p_1, \ldots, p_n)$ is the output bias of f.*

Proof If $p_i = 1/2$, the $q_i = 1/2$. Then

$$Pr(f = 1) = Pr(R_1) = \sum_{x \in R_1^f} Pr(x) = \sum_{x \in R_1^f} \prod_{i=1}^{n} \hat{p}_i \tag{2}$$

□

Example. *Consider the AND, OR, XOR, and NOT Boolean functions given in Table 1. The continuous-variable generalization of f_1, f_2, f_3, and f_4 are: $\hat{f}_1 = x_1 x_2$, $\hat{f}_2 = (1 - x_1)x_2 + x_1(1 - x_2) + x_1 x_2 = x_1 + x_2 - x_1 x_2$, $\hat{f}_3 = (1 - x_1)x_2 + x_1(1 - x_2) = x_1 + x_2 - 2x_1 x_2$, and $\hat{f}_4 = 1 - x$.*

Note that the above expressions have previously been derived via other (not probability-based) means[15,16].

Table 1. Truth tables of basic Boolean functions.

x_1	x_2	f_1	f_2	f_3
0	0	0	0	0
0	1	0	1	1
1	0	0	1	1
1	1	1	1	0

x	f_4
0	1
1	0

Taylor decomposition of Boolean functions

Since $\hat{f}$ is a continuous-variable function, we can calculate its Taylor expansion. And since $\hat{f}$ is a square-free polynomial, its Taylor expansion is finite and simplified (any term containing multiple derivatives of the same variable is zeroed out), as described in Proposition using the standard multi-index notation. Let $a = (a_1, \ldots, a_n)$ where $a_i \in \{0,1\}$. We define:

$$|a| = a_1 + \cdots + a_n, \tag{3}$$

$$x^a = x_1^{a_1} x_2^{a_2} \cdots x_n^{a_n}, \tag{4}$$

and

$$\partial^a f = \partial_1^{a_1} \partial_2^{a_2} \cdots \partial_n^{a_n} f = \frac{\partial^{|a|} f}{\partial_1^{a_1} \partial_2^{a_2} \cdots \partial_n^{a_n}}. \tag{5}$$

Proposition. *For $p \in [0,1]^n$, we have:*

$$\hat{f}(x) = \sum_{|a| \leq n} \frac{\partial^a \hat{f}(p)}{a!} (x - p)^a. \tag{6}$$

Note that $\hat{f}(p)$ in Eq. (6) is the output bias of f as was seen in Corollary. A natural choice for p is $p = (1/2, \ldots, 1/2)$ as it represents an unbiased selection for each variable and it also gives the output bias of the function. Such unbiased choices are not available for the discrete case. Our continuous formulation thus offers such unique advantages over the discrete Taylor decomposition, as it's a natural generalization of the latter. The Taylor decomposition can be used to approximate a Boolean function by considering a subset of the terms. For example, a linear approximation consists of terms only up to $|a| \leq 1$, a bilinear approximation up to $|a| \leq 2$, etc., up until $|a| \leq n$ where it ceases to be an approximation and provides an exact decomposition of $\hat{f}$. A visual illustration is provided in Fig. 5. The approximation order of a Boolean network could therefore vary between its minimum and maximum in-degrees (number of inputs per node).

Example. *Consider the continuous generalizations of the AND, OR, XOR and NOT functions given in Example The corresponding Taylor expansions using Eq. (6) and using the derivatives shown in Table 2 with $p = (1/2, 1/2)$ are: $\hat{f}_1 = 0.25 + 0.5(x_1 - 0.5) + 0.5(x_2 - 0.5) + (x_1 - 0.5)(x_2 - 0.5)$, $\hat{f}_2 = 0.75 + 0.5(x_1 - 0.5) + 0.5(x_2 - 0.5) - (x_1 - 0.5)(x_2 - 0.5)$, $\hat{f}_3 = 0.5 - 2(x_1 - 0.5)(x_2 - 0.5)$, and $\hat{f}_4 = 0.5 - (x - 0.5) = 1 - x$.*

Note that $\hat{f}_1(1/2, 1/2) = 0.25, \hat{f}_2(1/2, 1/2) = 0.75, \hat{f}_3(1/2, 1/2) = 0.5$, and $\hat{f}_4(1/2) = 0.5$ in the above equations are the output biases of the AND, OR, XOR, and NOT functions respectively. Also note that both the AND and OR functions contain the linear and the second order terms in their Taylor decomposition while the XOR function only contains the second order term. This difference is because both the AND and OR functions are monotone while XOR is not since it requires both inputs to be known.

Approximability of a model

We considered a suite of Boolean network models of biochemical regulation from two sources, namely the cell collective[9] and

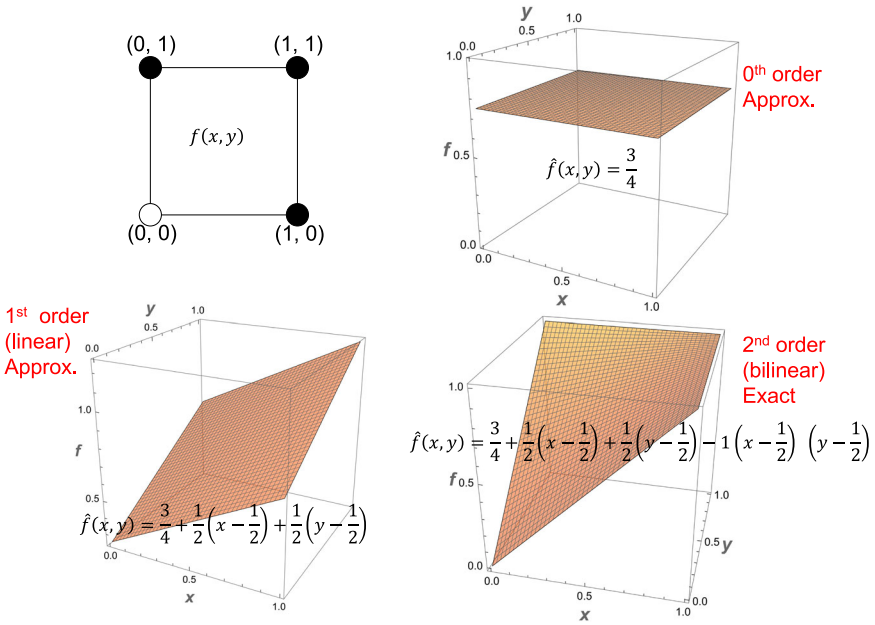


Fig. 5 The various approximations of a Boolean function in increasing order of nonlinearity. The logical OR function is represented as a 2D hypercube (top left) with the coordinate values representing input combinations and the color of the circles representing the corresponding outputs (white = 0, black = 1) and is approximated using Taylor decomposition as the 0th order approximation (top right) showing only the first term, the mean output bias; the 1st order approximation (bottom left) including the linear terms; and finally the 2nd order exact form (bottom right) including all the terms.

Table 2. Values of partial the derivatives in the Taylor decompositions of the generalizations of basic Boolean functions.

Derivative	f_1	f_2	f_3
∂_1	0.5	0.5	0
∂_2	0.5	0.5	0
$\partial_1\partial_2$	1	-1	-2

Derivative	f_4
∂_1	-1

reference[3]. This suite consists of 137 networks with the number of nodes ranging from 5 to 321. The mean in-degree of these models ranges from 1.1818 to 4.9375 with the variances ranging between 0.1636 and 9.2941, while the mean output bias is limited to the range [0.1625, 0.65625] with the variances between 0.0070 and 0.0933. For each biological model an ensemble of 100 randomized models was generated each of whose connectivity and output biases were set to be the same as the former, with only the logic rules randomized. This set, referred to as the "constrained ensemble", facilitates benchmarking the role of regulatory nonlinearity in the biological rules. To accurately assess the differences between the biological models and the constrained ensembles, we considered a baseline set known as the "unconstrained ensemble" relative to which those differences were computed. This ensemble preserved neither the connectivity nor the output bias but had them bootstrap-sampled from the corresponding distributions characterizing the associated biological models. Taylor decomposition was then applied to both the biological models and the random ensembles and all possible nonlinear approximations were computed by considering terms starting from the linear order accumulating up to the maximum possible nonlinear order. Both the biological models and the random ensembles were then simulated using a set of 1000 randomly chosen initial states iterated through 500 update steps for all orders of approximation; the same initial conditions were used for a given biological model and the random ensemble. The states of the variables were restricted to the interval [0,1] at every step during the simulations by simply resetting them to the nearest boundary of the interval whenever they crossed it. The mean approximation error (MAE) of each model is defined as the mean squared error between the exact Boolean states and the approximated probabilistic states at the end of the simulations; for the random ensembles a single average MAE was computed. It varies between 0.0 and 1.0. The "percentage change in MAE" (PMAE) of the biological models and the constrained ensembles is defined as the percentage difference between the respective MAE and the corresponding unconstrained MAE; it can vary between −100.0 and positive infinity. The "approximability" of a biological model is represented by the difference between the PMAE of the corresponding constrained ensemble and that of itself; it can vary between negative infinity and positive infinity. Hence, the more negative the PMAE is for biological models compared to that of constrained ensembles the more approximable they are deemed to be and the more unique their regulatory nonlinearity is. An illustration of how approximability is computed is provided in Fig. 6.

Classification of biological models

To identify any differences among the approximabilities of different types of biological networks we sought to classify them. Since there are multiple ways to classify biological networks, we chose two classifications so that: (1) they are as orthogonal as possible to each other; and (2) each classification has an appropriate number of (neither too few nor too many) categories. Classification 1 (C1) follows the "pathway ontology" (PW)[44] where the networks are grouped into five categories (Fig. 3a), namely biochemical ($n = 13$), signaling ($n = 22$), disease ($n = 55$), metabolic ($n = 14$) and regulatory ($n = 33$). According the definitions used in the PW ontology, a "signaling" network comprises mainly of extracellular signal transduction components such as growth factors, kinases, etc. A "regulatory" network, on the other hand, comprises intracellular transcriptional components such as genes,

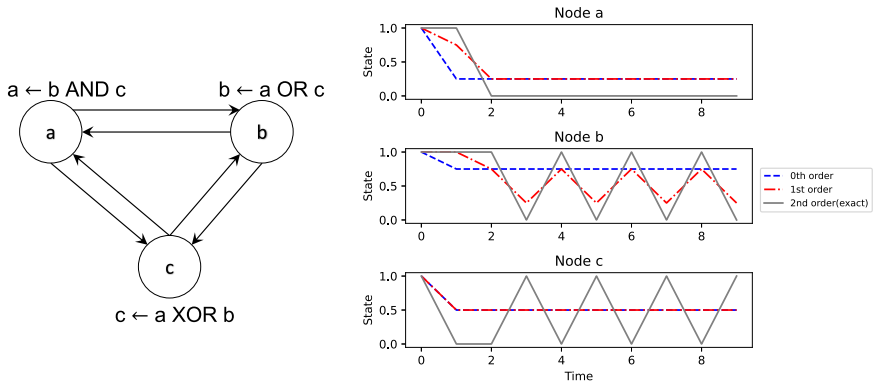


Fig. 6 An illustration of the relationship between nonlinearity and approximability using a simple 3-node Boolean network. The higher the order of nonlinearity of the network the more approximable it is with respect to the exact dynamics. Left: An example 3-node network utilizing all three 2-input Boolean functions. Right: The network is simulated from a single initial state of (1,1,1). Here the dynamics of the 0th order network (blue) least matches the exact Boolean trajectory (gray), while 1st order network (red) is a slightly better match. The 2nd order network (gray) is exact since it includes all the Taylor terms up to the maximum possible order of 2 (the maximum in-degree of any node). The closer the approximated average dynamics to the exact dynamics, the lower the MAE and the higher the approximability of the former. Notice also that while for node "a" the 0th order network is almost as good as the 1st order network, it is the opposite for node "c" in that its 1st order approximation is as poor as the 0th order; this is because "c" implements the XOR function whose 1st order derivatives are zero.

transcription factors, etc. The term "biochemical" here refers to

$$\frac{\partial \hat{f}(x_1,\ldots,x_i,\ldots,x_k)}{\partial x_i} = \sum_{x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_k\in\{0,1\}} \frac{\hat{f}(x_1,\ldots,x_{i-1},1,x_{i+1},\ldots,x_k) - \hat{f}(x_1,\ldots,x_{i-1},0,x_{i+1},\ldots,x_k)}{2^{k-1}} \tag{9}$$

networks that comprises a mix of signaling and regulatory components. "Metabolic" networks consist of components involved in the synthesis and conversion of biomolecules such as enzymes and lipids. Finally, "disease" networks consist of components involved in diseases such as cancer, anemia, pathogenic ailments and disorders such as cell cycle malfunction. Classification 2 was suggested by in-house expertise, where the networks are grouped into four categories (Fig. 3b), namely metazoan ($n = 85$), cancer ($n = 24$), primitive ($n = 19$) and plants ($n = 9$). The "metazonan" category refers to multicellular organisms and "primitive" refers to unicellular organisms. A given model could naturally belong in multiple categories within a classification but is assigned a unique category for the purpose of simplicity; we chose the categories according to the emphasis laid in the abstracts of the corresponding publications. More details are provided in the Supplementary Material (Supplementary Table 1).

Maximum absolute value of a Taylor derivative

Here we show that $\max(|\partial^\alpha \hat{f}|) = (\min(p, 0.5) - \max(p - 0.5, 0))2^{|\alpha|}$.

We begin with the definition of the derivative given by:

$$\frac{\partial \hat{f}(x_1,\ldots,x_i,\ldots,x_k)}{\partial x_i} = \lim_{h\to 0} \frac{\hat{f}(0.5,\ldots,h,\ldots,0.5) - \hat{f}(0.5,\ldots,0,\ldots,0.5)}{h} \tag{7}$$

Since $\hat{f}$ is a pseudo-Boolean function and hence a multilinear polynomial[45], we can compute this as a finite difference; the idea being that the derivative taken over any point on a line is the line itself:

$$\frac{\partial \hat{f}(x_1,\ldots,x_i,\ldots,x_k)}{\partial x_i} = \hat{f}(0.5,\ldots,1,\ldots,0.5) - \hat{f}(0.5,\ldots,0,\ldots,0.5) \tag{8}$$

Since multilinear interpolation can be formulated as weighted averaging[43], we can further rewrite it as follows (the weights are equal here since the non-binary values are all set to 0.5):

As can be seen, there are a total of 2^k terms, with half of them positively signed and half negative. This form of expression generalizes to derivatives taken over two or more variables. For example, the derivative taken over two variables, x_i and x_j, looks as follows:

$$\frac{\partial^2 \hat{f}(\ldots,x_i,\ldots,x_j,\ldots)}{\partial x_i \partial x_j} = \sum_{x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_k\in\{0,1\}} \sum_{x_1,\ldots,x_{j-1},x_{j+1},\ldots,x_k\in\{0,1\}} \left[\frac{\hat{f}(\ldots,x_{i-1},1,x_{i+1},\ldots,x_{j-1},1,x_{j+1},\ldots) - \hat{f}(\ldots,x_{i-1},1,x_{i+1},\ldots,x_{j-1},0,x_{j+1},\ldots)}{2^{k-2}}\right] - \left[\frac{\hat{f}(\ldots,x_{i-1},0,x_{i+1},\ldots,x_{j-1},1,x_{j+1},\ldots) - \hat{f}(\ldots,x_{i-1},0,x_{i+1},\ldots,x_{j-1},0,x_{j+1},\ldots)}{2^{k-2}}\right] \tag{10}$$

Following rearrangement of terms it becomes evident that this expression also contains 2^{k-1} positive terms and 2^{k-1} negative terms, with the only difference in the power of the denominator term. It can thus be concluded that any derivative $\partial^\alpha \hat{f}$ (in multi-index notation) has 2^{k-1} positive terms and 2^{k-1} negative terms.

A straightforward way to maximize the value of a derivative expressed in this form is by assigning as many instances of 1 as possible to the positive terms and as few instances of 1 as possible to the negative terms. For a Boolean function with k inputs and output bias p, this can be accomplished by assigning $\min(2^{k-1}, p2^k)$ ones and $\max(p2^k - 2^{k-1}, 0)$ ones respectively. Therefore:

$$\begin{aligned}\max(|\partial^\alpha \hat{f}|) &= \frac{\min(2^{k-1}, p2^k) - \max(p2^k - 2^{k-1}, 0)}{2^{k-|\alpha|}} \\ &= (\min(p, 0.5) - \max(p - 0.5, 0))2^{|\alpha|}; 1 \le |\alpha| \le k.\end{aligned} \tag{11}$$

Note that this formula only applies to a specific order of nonlinearity $|\alpha|$ independent of the other orders within the same Boolean function. In actuality, there are dependencies between the various orders within a Boolean function. That is, if a Boolean function were to be constructed such that the derivative of a

particular order $|a_1|$ is maximized then there's no guarantee that the derivative of another order $|a_2| \neq |a_1|$ could be simultaneously maximized. This is one of the limitations of the normalization for which the above formula is used.

Reporting summary

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

DATA AVAILABILITY

We used 137 Boolean network models of gene regulation from two sources, namely the cell collective[9] and reference[3]. Subsequently, we categorized these models using two types of classifications C1 and C2 as described in Section. We provide the details of the classification of each model along with their Pubmed ID and their linear approximation score in the Supplementary Information file (SI), Supplementary Table 1.

CODE AVAILABILITY

The code that we used for the approximations and the simulations are available through this link: https://gitlab.com/smanicka/boolion.

Received: 6 July 2022; Accepted: 23 March 2023;
Published online: 04 April 2023

REFERENCES

1. Kapitaniak, T. & Jafari, S. *Nonlinear Effects in Life Sciences* (Springer, 2018).
2. Savageau, M. A. Design principles for elementary gene circuits: elements, methods, and examples. *Chaos* **11**, 142–159 (2001).
3. Kadelka, C., Butrie, T.-M., Hilton, E., Kinseth, J. & Serdarevic, H. A meta-analysis of Boolean network models reveals design principles of gene regulatory networks. *arXiv* https://doi.org/10.48550/arXiv.2009.01216 (2020).
4. Borriello, E. & Daniels, B. C. The basis of easy controllability in Boolean networks. *Nat. Commun.* **12**, 5227 (2021).
5. Stoof, R. & Goñi-Moreno, Á. Modelling co-translational dimerization for programmable nonlinearity in synthetic biology. *J. R. Soc. Interface* **17**, 20200561 (2020).
6. Pezzulo, G. & Levin, M. Top-down models in biology: explanation and control of complex living systems above the molecular level. *J. R. Soc. Interface* **13**, 20160555 (2016).
7. Lobo, D., Solano, M., Bubenik, G. A. & Levin, M. A linear-encoding model explains the variability of the target morphology in regeneration. *J. R. Soc. Interface* **11**, 20130918 (2014).
8. Saadatpour, A. & Albert, R. A comparative study of qualitative and quantitative dynamic models of biological regulatory networks. *EPJ Nonlinear Biomed. Phys.* **4**, 1–13 (2016).
9. Helikar, T., Kowal, B. & Rogers, J. A cell simulator platform: the cell collective. *Clin. Pharmacol. Ther.* **93**, 393–395 (2013).
10. Daniels, B. C. et al. Criticality distinguishes the ensemble of biological regulatory networks. *Phys. Rev. Lett.* **121**, 138102 (2018).
11. Manicka, S., Marques-Pita, M. & Rocha, L. M. Effective connectivity determines the critical dynamics of biochemical networks. *J. R. Soc. Interface* **19**, 20210659 (2022).
12. Boole, G. *Collected Logical Works, Vol. I. Studies in Logic and Probability* (ed. Rhees, R.) (Open Court Publishing Company, 1952).
13. O'Donnell, R. *Analysis of Boolean Functions* (Cambridge University Press, 2014).
14. Yanushkevich, S. & Shmerko, V. Taylor expansion of logic functions: from conventional to nanoscale design. In *Int. TICSP Workshop on Spectral Methods and Multirate Signal Processing* (Citeseer, 2004).
15. Grieb, M. et al. Predicting variabilities in cardiac gene expression with a Boolean network incorporating uncertainty. *PLoS ONE* **10**, e0131832 (2015).
16. Betel, H. & Flocchini, P. On the relationship between fuzzy and Boolean cellular automata. *The. Comput. Sci.* **412**, 703–713 (2011).
17. Ebrahimkhani, M. R. & Levin, M. Synthetic living machines: a new window on life. *Iscience* **24**, 102505 (2021).
18. Merchan, L. & Nemenman, I. On the sufficiency of pairwise interactions in maximum entropy models of networks. *J. Stat. Phys.* **162**, 1294–1308 (2016).
19. Gutenkunst, R. N. et al. Universally sloppy parameter sensitivities in systems biology models. *PLoS Comput. Biol.* **3**, e189 (2007).
20. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).
21. Choi, M., Shi, J., Jung, S. H., Chen, X. & Cho, K.-H. Attractor landscape analysis reveals feedback loops in the p53 network that control the cellular response to dna damage. *Sci. Signal.* **5**, ra83–ra83 (2012).
22. Cohen, D. P. et al. Mathematical modelling of molecular pathways enabling tumour cell invasion and migration. *PLoS Comput. Biol.* **11**, e1004571 (2015).
23. Derrida, B. & Pomeau, Y. Random networks of automata: a simple annealed approximation. *Europhys. Lett.* **1**, 45 (1986).
24. Levin, M. Bioelectrical approaches to cancer as a problem of the scaling of the cellular self. *Prog. Biophys. Mol. Biol.* **165**, 102–113 (2021).
25. Linardatos, P., Papastefanopoulos, V. & Kotsiantis, S. Explainable AI: a review of machine learning interpretability methods. *Entropy* **23**, 18 (2020).
26. Manicka, S. & Levin, M. Minimal developmental computation: a causal network approach to understand morphogenetic pattern formation. *Entropy* **24**, 107 (2022).
27. Watson, R. A., Levin, M. & Buckley, C. L. Design for an individual: connectionist approaches to the evolutionary transitions in individuality. *Front. Ecol. Evol.* **64**, 64–87 (2022).
28. Kauffman, S. The large scale structure and dynamics of gene control circuits: an ensemble approach. *J. Theor. Biol.* **44**, 167–190 (1974).
29. Reichhardt, C. O. & Bassler, K. E. Canalization and symmetry in Boolean models for genetic regulatory networks. *J. Phys. A* **40**, 4339 (2007).
30. Gates, A. J. & Rocha, L. M. Control of complex networks requires both structure and dynamics. *Sci. Rep.* **6**, 1–11 (2016).
31. Müller, F.-J. & Schuppert, A. Few inputs can reprogram biological networks. *Nature* **478**, E4–E4 (2011).
32. Wagner, A. Does evolutionary plasticity evolve? *Evolution* **50**, 1008–1023 (1996).
33. Biswas, S., Manicka, S., Hoel, E. & Levin, M. Gene regulatory networks exhibit several kinds of memory: quantification of memory in biological and random transcriptional networks. *Iscience* **24**, 102131 (2021).
34. Mitchell, M. Ubiquity 2011, 1-7. *Ubiquity* **2011**, 1–7 (2011).
35. Chu, D., Prokopenko, M. & Ray, J. C. J. Computation by natural systems (2018).
36. Cook, M. et al. Universality in elementary cellular automata. *Complex Syst.* **15**, 1–40 (2004).
37. Wolfram, S. Statistical mechanics of cellular automata. *Rev. Mod. Phys.* **55**, 601 (1983).
38. Zenil, H. & Riedel, J. Asymptotic intrinsic universality and natural reprogrammability by behavioural emulation. In *Advances in Unconventional Computing* (ed Adamatzky, A.) 205–220 (Springer, 2017).
39. Manicka, S. V. S. *The Role of Canalization in the Spreading of Perturbations in Boolean Networks*. Ph.D. thesis, Indiana University (2017).
40. Marques-Pita, M. & Rocha, L. M. Canalization and control in automata networks: body segmentation in drosophila melanogaster. *PLoS ONE* **8**, e55946 (2013).
41. Rocha, L. M. On the feasibility of dynamical analysis of network models of biochemical regulation. *arXiv* https://doi.org/10.48550/arXiv.2110.10821 (2021).
42. Gates, A. J., Brattig Correia, R., Wang, X. & Rocha, L. M. The effective graph reveals redundancy, canalization, and control pathways in biochemical regulation and signaling. *Proc. Natl. Acad. Sci. USA* **118**, e2022598118 (2021).
43. Press, W. H., Flannery, B. P., Teukolsky, S. A. & Vetterling, W. T. *Numeric Recipes in C: The Art of Scientific Computing* (Cambridge University Press, 1992).
44. Petri, V. et al. The pathway ontology–updates and applications. *J. Biomed. Semant.* **5**, 1–12 (2014).
45. Hammer, P. L. et al. On the determination of the minima of pseudo-Boolean functions. *Studii si Cercetari Matematice* **14**, 359–364 (1963).
46. Battiston, F. et al. The physics of higher-order interactions in complex systems. *Nat. Phys.* **17**, 1093–1098 (2021).

ACKNOWLEDGEMENTS

D.M. was partially supported by a Collaboration grant (850896) from the Simons Foundation. M.L. gratefully acknowledges support via grant TWCF0606 of the Templeton World Charity Foundation. We thank Claus Kadelka for sharing a large repository of Boolean models curated by his group and the accompanying Python code for analyzing it. The authors thank the referees for their insightful comments that have improved the manuscript.

AUTHOR CONTRIBUTIONS

S.M. and D.M. conceived the study; S.M. and K.J. designed code; K.J. performed research and simulations; S.M., K.J. and D.M. analyzed data; M.L. helped bridge the theory to biology. All authors helped in the writing of the manuscript. All authors approved the final version of the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41540-023-00273-w.

Correspondence and requests for materials should be addressed to David Murrugarra.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.