BIOLOGICAL REGULATORY NETWORKS ARE LESS NONLINEAR THAN EXPECTED BY CHANCE

A PREPRINT

Santosh Manicka[*] Tufts University

Kathleen Johnson* University of Kentucky

Michael Levin Tufts University

David Murrugarra[†] University of Kentucky

December 22, 2021

ABSTRACT

Nonlinearity is a characteristic of complex biological regulatory networks that has implications ranging from therapy to control. To better understand its nature, we analyzed a suite of published Boolean network models, containing a variety of complex nonlinear interactions, with an approach involving a probabilistic generalization of Boolean logic that George Boole himself had proposed. Leveraging the continuous-nature of this formulation using Taylor-decomposition methods revealed the distinct layers of nonlinearity of the models. A comparison of the resulting series of model-approximations with the corresponding sets of randomized ensembles furthermore revealed that the biological networks are relatively more linearly approximable. We hypothesize that this is a result of optimization by natural selection for the purpose of controllability.

Keywords Nonlinearity · Biological networks · Boolean decomposition

1 Introduction

How nonlinear are biological regulatory networks? That is, to what extent do the biochemical components of these networks non-independently interact in influencing downstream processes? Research on this front has hitherto focused on the various manifestations of nonlinearity in the dynamics of biological systems, such as chaos, bifurcation, multistability, synchronization, patterning, dissipation, etc.[1], but a characterization of nonlinearity in the underlying systems that give rise to those phenomena is lacking. A more complete understanding of biological nonlinearity would have theoretical implications ranging from canalization to control [2, 3] and practical implications such as therapy, synthetic biology, etc. [1, 4]. 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 [5], 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 [6] 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

[*] 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.

*The authors have declared that no competing interests exist.

*S.M. contributed equally to this work with K.J.

†To whom correspondence should be addressed. E-mail: murrugarra@uky.edu

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 [7].

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 [8]. Some of the publicly available models were generated via network inference methods from time course data [2].

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 [9, 10]. Here we study a new but generic feature of complex systems, namely, nonlinearity. To characterize the 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 [11]. We leverage the continuous nature of these polynomials to decompose a Boolean function using Taylor-series and reveal its distinct layers of nonlinearity (Fig 1). Various other methods, both discrete and continuous, of decomposing Boolean functions exist, such as Reed-Muller, Walsh spectrum, Fourier and discrete Taylor [12, 13, 14]. Our continuous Taylor decomposition method is distinct in that it offers a clear and systematic way to characterize nonlinearity.

By characterizing the nonlinearity of networks in this way, we answer the following questions: 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?; and 2) is there an optimal level of nonlinearity that these models may have been selected for by evolution? 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. Finally, we 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.

Methods

Probabilistic generalization*

Here we provide a continuous-variable formulation of a Boolean function by casting Boolean values as probabilities. Consider random variables $X_i : \{0,1\} \to [0,1]$, $i = 1,\dots,n$, with Bernoulli distributions. That is, $p_i = Pr(X_i = 1) = 1 - Pr(X_i = 0) = 1 - q_i$, for $i = 1,\dots,n$. 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, $Pr(f = 1) = Pr(R_1^f) = \sum_{x \in R_1^f} Pr(x) = \sum_{x \in R_1^f} \prod_{i=1}^n \hat{p}_i$ where $\hat{p}_i = p_i$ if $x_i = 1$ and $\hat{p}_i = 1 - p_i$ if $x_i = 0$. Let $\hat{f}(p_1,\dots,p_n) = \sum_{x \in R_1^f} \prod_{i=1}^n \hat{p}_i$. 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 1.1. *For discrete values of $x_i \in \{0,1\}$, $i = 1,\dots,n$, we have $\hat{f}(x_1,\dots,x_n) = f(x_1,\dots,x_n)$.*

Proof. Let $z = (z_1,\dots,z_n) \in \{0,1\}^n$. 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 $i = 1,\dots,n$. We want to show that $\hat{f}(p_1,\dots,p_n) = f(z_1,\dots,z_n)$. 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 1.2. *If $p_i = 1/2$ for all $i = 1,\dots,n$, then $\hat{f}(p_1,\dots,p_n)$ is the output bias of f.*

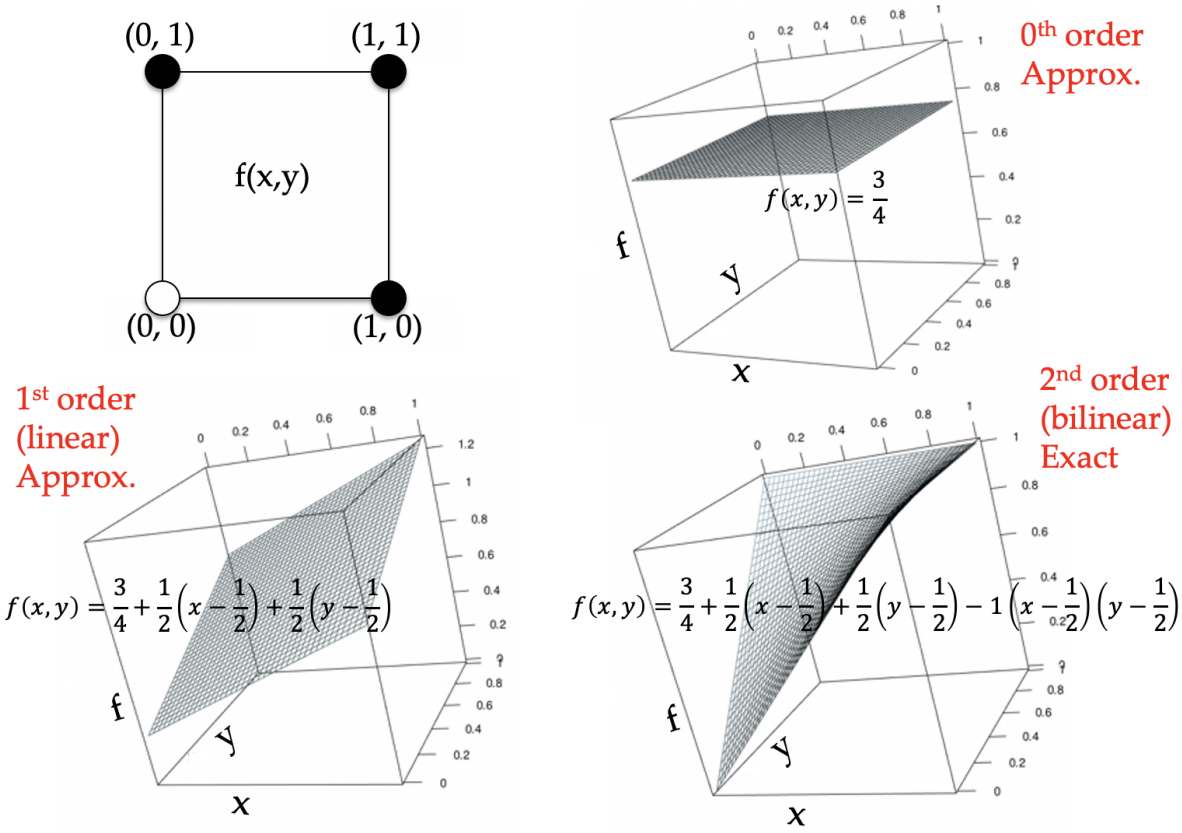

Figure 1: The various approximations of a Boolean function. The logical OR function is represented as a 2D hypercube (top left) with the coordinate values representing input combinations and the color of the circles representing the corresponding outputs (white=0, black=1) and is approximated using Taylor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

Example 1.3. *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:* $\hat{f}_1 = x_1x_2$, $\hat{f}_2 = (1-x_1)x_2 + x_1(1-x_2) + x_1x_2 = x_1 + x_2 - x_1x_2$, $\hat{f}_3 = (1-x_1)x_2 + x_1(1-x_2) = x_1 + x_2 - 2x_1x_2$, *and* $\hat{f}_4 = 1 - x$.

Note that the above expressions have previously been derived via other (not probability-based) means [14].

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

Table 1: Truth tables of basic Boolean functions.

Taylor Decomposition*

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 1.4 using the standard multi-index notation. Let $\alpha = (\alpha_1, \ldots, \alpha_n)$ where $\alpha_i \in \{0, 1\}$. We define $\mid \alpha \mid = \alpha_1 + \cdots + \alpha_n$, $x^\alpha = x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n}$, and $\partial^\alpha f = \partial_1^{\alpha_1} \partial_2^{\alpha_2} \cdots \partial_n^{\alpha_n} f = \frac{\partial^{|\alpha|} f}{\partial_1^{\alpha_1} \partial_2^{\alpha_2} \cdots \partial_n^{\alpha_n}}$.

Proposition 1.4. *For* $p \in [0,1]^n$, *we have*

$$\hat{f}(x) = \hat{f}(p) + \textstyle\sum_{1<|\alpha|\leq n} \partial^\alpha \hat{f}(p)(x-p)^\alpha. \tag{1}$$

Note that $\hat{f}(p)$ in Equation 1 is the output bias of f as was seen in Corollary 1.2. 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. Being a natural generalization of the discrete Taylor decomposition, it thus offers certain unique advantages over 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 $|\alpha| \leq 1$, a bilinear approximation up to $|\alpha| \leq 2$, etc. A visual illustration is provided in Figure 1. The approximation order of a Boolean *network* therefore varies between its minimum and maximum in-degrees.

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

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

Example 1.5. *Consider the continuous generalizations of the AND, OR, XOR and NOT functions given in Example 1.3 The corresponding Taylor expansions using Equation 1 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.

Simulation and analysis*

We considered a suite of Boolean network models of biochemical regulation from two sources namely the *cell collective* [8] and reference [2]. This suite consists of 138 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 we generated an associated ensemble of 100 randomized models, where the connectivity and the output bias of the nodes of the original model were preserved and the logic rules were randomly chosen under the above constraints. This approach helps avoid confounding the causes of any observed effects with network structure or output bias, thereby narrowing the focus on the nonlinearity of the Boolean functions. We applied the Taylor decomposition to both the biological models and the associated ensembles and computed all possible nonlinear approximations. We then simulated each biological model and the associated random ensemble using the same set of 1000 randomly chosen initial states iterated through 500 update steps for all orders of approximation. The states of the variables were restricted to the interval [0,1] at every step in the simulations. We then computed the mean squared error (MSE) between the exact Boolean state and the approximated state at the end of the simulations, the inverse of which could be defined as a measure of the approximation accuracy. For each random ensemble, we computed a single average MSE.

Results and Discussion

The central result is that the biological models are relatively more approximable for various degrees of nonlinearity when compared to a reference ensemble (Figure 2). The contrast is most prominent at the linear regime where the biological models are about 2% more accurately approximable ($p < 10^{-5}$) compared to their random counterparts. This suggests that the biological regulatory networks may have been optimized (presumably by evolution) for linearity in terms of the nonlinearity of the Boolean rules, given that the reference ensemble preserves the network structure and the output biases of the corresponding biological models. This has implications not only for the feasibility of biomedical approaches to control emergent somatic complexity or guided self-assembly of novel forms [15], 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. The main methodological contribution of this paper is the introduction of 'nonlinearity' as a new measure of characterization for Boolean networks. Our results hint at a connection with other existing measures such as canalization, effective connectivity, symmetry and controllability since it has been previously reported that the levels of canalization (a measure of the extent to which fewer inputs influence the outputs of a Boolean function) and the mean effective connectivity (a measure of collective

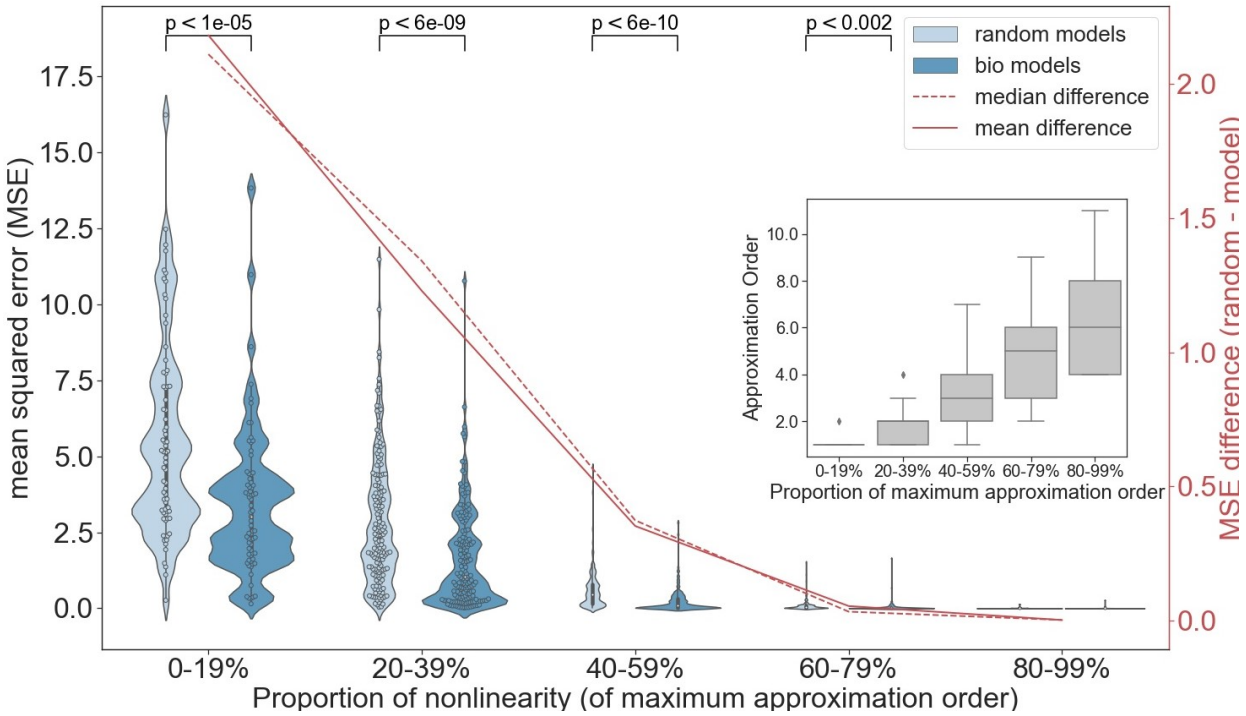

Figure 2: Biological models are more approximable for various degrees of nonlinearity compared to a reference random ensemble. Every point in the light blue plots represent the average approximation accuracy of an ensemble of 100 random networks associated with each biological model. The maximum order of approximation for a model is equal to its max in-degree. The inset shows the spread of the absolute approximation orders for every proportion bucket. See main text for details.

canalization) are high in biological networks [2, 10]. Furthermore, it has been found that biological networks need few inputs to reprogram [16] and are relatively easier to control [3]. Since less nonlinearity suggests more apportioning of influence to smaller sets of inputs, we hypothesize that the observed optimization of biological networks for linearity may serve the purpose of better controllability. The main limitation of our analysis is that the approximation accuracy will necessarily increase with higher orders of approximation for arbitrary Boolean networks (the highest order of approximation yields the exact function). However, this does not affect the falsifiability of our framework since it's possible to construct networks, say with XOR-like functions, that are less linearly approximable than the associated ensembles. Also, the notion of nonlinearity is limited to the local level of the Boolean rules in our framework, whereas its possible to conceive network-level measures of nonlinearity where the role of the network structure is included. 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.

References

[1] Tomasz Kapitaniak and Sajad Jafari. Nonlinear effects in life sciences, 2018.

[2] Claus Kadelka, Taras-Michael Butrie, Evan Hilton, Jack Kinseth, and Haris Serdarevic. A meta-analysis of boolean network models reveals design principles of gene regulatory networks. *arXiv preprint arXiv:2009.01216*, 2020.

[3] Enrico Borriello and Bryan C. Daniels. The basis of easy controllability in boolean networks. *Nature Communications*, 12(5227), 2021.

[4] Ruud Stoof and Ángel Goñi-Moreno. Modelling co-translational dimerization for programmable nonlinearity in synthetic biology. *Journal of the Royal Society Interface*, 17(172):20200561, 2020.

[5] Giovanni Pezzulo and Michael Levin. Top-down models in biology: explanation and control of complex living systems above the molecular level. *Journal of The Royal Society Interface*, 13(124):20160555, 2016.

[6] Daniel Lobo, Mauricio Solano, George A Bubenik, and Michael Levin. A linear-encoding model explains the variability of the target morphology in regeneration. *Journal of The Royal Society Interface*, 11(92):20130918, 2014.

[7] Assieh Saadatpour and Réka Albert. A comparative study of qualitative and quantitative dynamic models of biological regulatory networks. *EPJ Nonlinear Biomedical Physics*, 4(1):1–13, 2016.

[8] Tomáš Helikar, B Kowal, and JA Rogers. A cell simulator platform: the cell collective. *Clinical Pharmacology & Therapeutics*, 93(5):393–395, 2013.

[9] Bryan C Daniels, Hyunju Kim, Douglas Moore, Siyu Zhou, Harrison B Smith, Bradley Karas, Stuart A Kauffman, and Sara I Walker. Criticality distinguishes the ensemble of biological regulatory networks. *Physical review letters*, 121(13):138102, 2018.

[10] Santosh Manicka, Manuel Marques-Pita, and Luis M Rocha. Effective connectivity determines the critical dynamics of biochemical networks. *arXiv preprint arXiv:2101.08111*, 2021.

[11] George Boole. Collected logical works, vol. i. *Studies in Logic and Probability, R. Rhees (ed.), Open Court Publ. Co., LaSalle, Ill*, 1952.

[12] Ryan O'Donnell. *Analysis of boolean functions*. Cambridge University Press, 2014.

[13] SN Yanushkevich and VP Shmerko. Taylor expansion of logic functions: From conventional to nanoscale design. In *Int. TICSP Workshop on Spectral Methods and Multirate Signal Processing*. Citeseer, 2004.

[14] Melanie Grieb, Andre Burkovski, J Eric Sträng, Johann M Kraus, Alexander Groß, Günther Palm, Michael Kühl, and Hans A Kestler. Predicting variabilities in cardiac gene expression with a boolean network incorporating uncertainty. *PloS one*, 10(7):e0131832, 2015.

[15] Mo R Ebrahimkhani and Michael Levin. Synthetic living machines: A new window on life. *Iscience*, page 102505, 2021.

[16] Franz-Josef Müller and Andreas Schuppert. Few inputs can reprogram biological networks. *Nature*, 478(7369):E4–E4, 2011.