Theory in Biosciences (2020) 139:309–318
https://doi.org/10.1007/s12064-020-00332-4

REVIEW

Ingredients for robustness

Nihat Ay[1,2,3]

Received: 12 October 2020 / Accepted: 12 November 2020 / Published online: 2 December 2020

Abstract
A core property of robust systems is given by the invariance of their function against the removal of some of their structural components. This intuition has been formalised in the context of input–output maps, thereby introducing the notion of exclusion independence. We review work on how this formalisation allows us to derive characterisation theorems that provide a basis for the design of robust systems.

Keywords Robustness · Knockouts · Neutrality · Interaction order

Introduction: What are robust systems?

The robustness of a system, be it biological or artificial, always refers to a set of perturbations. A biological system has to cope with a number of perturbations, including a constantly changing environment (Wagner 2007 gives a general introduction to the robustness of biological systems). One way to do so is by shaping and controlling its environment in a way that reduces perturbations. This is often referred to as niche construction, which can take place in the ecological as well as the social domain (Flack et al. 2006; Krakauer et al. 2009). Another way to deal with perturbations is by an intrinsic adaptation, change of the system itself and not the environment, so that the system's function remains unchanged. It is this second way of coping with perturbations that we are addressing in this article. The setting in which we want to study this kind of robustness is kept very simple, as shown in Fig. 1. It is given by a (stochastic) map that receives n inputs and generates a, potentially high-dimensional, output. It is surprising how far-reaching this minimalistic setting is. It is general enough for studying the robustness of a number of biologically relevant mappings, such as the genotype-phenotype map, the genetic code, and neurons. However, addressing such application fields is beyond the scope of this article. Our aim is rather moderate. We review previous work on this subject and highlight the main insights in a more direct, instructive and conceptually complete way. The presented theory has been initially proposed by Ay and Krakauer (2007) and further developed in a number of works, including (Ay et al. 2007; Krakauer et al. 2010; Boldhaus et al. 2010; Rauh 2013; Rauh and Ay 2014). It touches upon important subjects related to robustness, such as niche construction, adaptation and neutrality, which have been addressed from various perspectives in a large number of publications. This article is by no means complete in providing a review of these publications. Instead, it restricts attention to the mentioned works in which connections to other works are outlined more thoroughly.

In order to study edge or node deletion, which we also call *knockout intervention* or simply *knockout*, we consider a single unit of the network, the output unit, together with all its parents $I := \{1, \dots, n\}$, which provide the input. We denote the input variables by $X_1, \dots, X_n$ and the output variable by Y. Correspondingly, the values of X_i are denoted by x_i and the values of Y by y (see Fig. 1). Throughout this article, we assume that x_i and y are elements of some non-empty and finite state spaces, $\mathcal{X}_i$ and $\mathcal{Y}$, respectively. For a subset J of I, we write $\mathcal{X}_J$ for the Cartesian product of the $\mathcal{X}_i$, $i \in J$, and denote its elements by $\boldsymbol{x}_J$.

We model the mechanism of Y as a stochastic map $\kappa(\boldsymbol{x};y)$ which assigns to each input $\boldsymbol{x} = (x_1, \dots, x_n)$ a random output y, that is $\kappa(\boldsymbol{x};y) \geq 0$ and $\sum_{y'} \kappa(\boldsymbol{x};y') = 1$ for all $\boldsymbol{x}$, y. Such a map is called a *Markov kernel*. In this article, we study robustness with respect to knockouts of some of the input nodes I. When a subset K of the input nodes is removed, typically the

✉ Nihat Ay
nay@mis.mpg.de

[1] Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany

[2] Leipzig University, Leipzig, Germany

[3] Santa Fe Institute, Santa Fe NM, USA

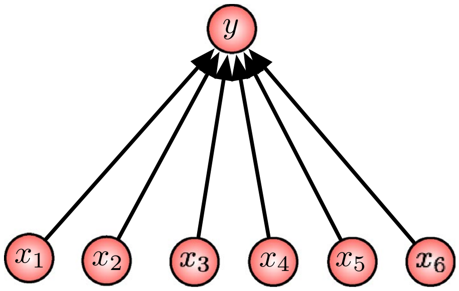

Fig. 1 An input–output map for the study of robustness

behaviour of the remaining system with input nodes $J = I \backslash K$ will be different. Without further assumptions, the post-knockout function is not determined by κ and can be anything. We therefore consider another stochastic map, a Markov kernel, $\kappa_J : \mathcal{X}_J \times \mathcal{Y} \to [0,1]$ as model of the post-knockout function (see Fig. 2).

A complete specification of the function is given by the family $(\kappa_J)_{J \subseteq I}$ of all possible post-knockout functions, which we refer to as *functional modalities*. The Markov kernel κ itself, which describes the behaviour of the system prior to any knockouts, can be identified with κ_I. Let us now come back to robustness. Intuitively, a system is considered to be robust if its post-knockout function coincides with, or does not deviate too much from, its pre-knockout function. Clearly, this cannot be non-trivially achieved if it has to hold for all knockout perturbations. However, if we upper bound the number of deleted input nodes, typically by a much smaller number than n, then the problem of characterising robust systems turns out to be quite interesting. We will review such characterisations in this article.

In “Exclusion dependence” section, we provide a quantitative approach to the outlined invariance of function with respect to knockouts, referred to as *exclusion dependence*. Basically, it measures the deviation of the post-knockout function from the pre-knockout function. This already allows us to present our first characterisation of robust systems in “Neutrality” section, which is related to *neutrality*. In “Mechanistic modelling of knockout interventions” section, we introduce a mechanistic description of knockout perturbations, thereby relying on notions from statistical physics. This allows us to solve the problem of system identification based on experimental knockouts in terms of the Möbius inversion. The solution is closely related to the Gibbs–Markov equivalence known from statistical physics. We present the subject of system identification in “Knockout interventions for system identification” section. Intuitively, the difficulty of a system’s identification in terms of knockout experiments is directly correlated with the robustness of that system. This follows formally from our second characterisation of robust systems, presented in “Robustness and interaction order” section. In the “Conclusions” section, we compare the two presented descriptions of robust systems and outline directions of research.

Exclusion dependence

By exclusion dependence, we mean the deviation or distance of the post-knockout function from the original function. With an appropriate distance measure, knockout interventions can be studied in a quantitative manner, which allows us to derive a general robustness measure. Conceptually, any measure of robustness should incorporate the invariance of the system’s function against knockout of input nodes, say the nodes of a subset K. After the knockout, the system will have access only to the remaining input nodes, those in the complement $J := I \backslash K$. In general, the function will be affected by such a knockout intervention. We measure the deviation of the post-knockout function κ_J from the original function κ in $\boldsymbol{x} = (\boldsymbol{x}_J, \boldsymbol{x}_K)$ by using the relative entropy or Kullback-Leibler divergence (KL-divergence) as a distance measure. For two probability vectors p and q on a finite set $\mathcal{Z}$, it is defined as

$$D(p \,\|\, q) := \sum_{z \in \mathcal{Z}} p(z) \ln \frac{p(z)}{q(z)},$$

with the convention $p(z) \ln \frac{p(z)}{q(z)} = 0$ if $p(z) = 0$, and $p(z) \ln \frac{p(z)}{q(z)} = \infty$ if $p(z) > q(z) = 0$. The KL-divergence satisfies the following fundamental property:

$$D(p \| q) \geq 0, \quad \text{and} \quad D(p \| q) = 0 \quad \text{if and only if} \quad p = q. \tag{1}$$

The KL-divergence is well-known as a canonical divergence in information geometry (Amari and Nagaoka 2000; Amari 2016; Ay and Amari 2015; Ay et al. 2017). We now use the KL-divergence in order to quantify the deviation of the post-knockout function κ_J from the original function κ, given that they both have input $\boldsymbol{x} = (\boldsymbol{x}_J, \boldsymbol{x}_K)$:

$$D\big(\kappa(\boldsymbol{x}_J, \boldsymbol{x}_K; \cdot) \,\|\, \kappa_J(\boldsymbol{x}_J; \cdot)\big) := \sum_y \kappa(\boldsymbol{x}_J, \boldsymbol{x}_K; y) \ln \frac{\kappa(\boldsymbol{x}_J, \boldsymbol{x}_K; y)}{\kappa_J(\boldsymbol{x}_J; y)}. \tag{2}$$

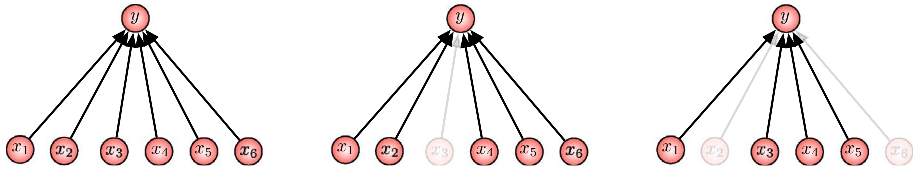

Fig. 2 Left: unperturbed function. Middle: knockout of node 3. Right: knockout of nodes 2 and 6

Given the property (1) of the KL-divergence, this deviation vanishes if and only if the post-knockout function coincides with the original function in $\boldsymbol{x} = (\boldsymbol{x}_J, \boldsymbol{x}_K)$, that is

$$\kappa(\boldsymbol{x}_J, \boldsymbol{x}_K; y) = \kappa_J(\boldsymbol{x}_J; y) \qquad \text{for all } y. \tag{3}$$

Note that this invariance of the function refers to the input $\boldsymbol{x}$. This does not necessarily mean that the function will also be invariant in the context of another input $\boldsymbol{x}'$. Let us consider an input distribution μ with support $\mathcal{S}$. Taking the mean of the individual KL-divergences (2) with respect to μ yields

$$D_\mu(\kappa \parallel \kappa_J) \;:=\; \sum_{\boldsymbol{x}} \mu(\boldsymbol{x}) \sum_y \kappa(\boldsymbol{x}; y) \ln \frac{\kappa(\boldsymbol{x}; y)}{\kappa_J(\boldsymbol{x}_J; y)},$$

which we refer to as *exclusion dependence* (Ay and Krakauer 2007). Obviously, the exclusion dependence vanishes if and only if the invariance (3) holds for all $\boldsymbol{x} \in \mathcal{S}$. We now relate the invariance of function to stochastic independence. To be more precise, we set

$$\begin{aligned} p(\boldsymbol{x}_J, \boldsymbol{x}_K, y) &:= \mu(\boldsymbol{x}_J, \boldsymbol{x}_K)\, \kappa(\boldsymbol{x}_J, \boldsymbol{x}_K; y) \\ &= \mathrm{Prob}\left\{X_J = \boldsymbol{x}_J, X_K = \boldsymbol{x}_K, Y = y\right\} \end{aligned}$$

and define the *conditional mutual information* of Y and X_K, given X_J:

$$\begin{aligned} & I(Y; X_K \mid X_J) \\ &\quad := \sum_{\boldsymbol{x}_J} p(\boldsymbol{x}_J) \sum_{y, \boldsymbol{x}_K} p(y, \boldsymbol{x}_K | \boldsymbol{x}_J) \ln \frac{p(y, \boldsymbol{x}_K | \boldsymbol{x}_J)}{p(y | \boldsymbol{x}_J)\, p(\boldsymbol{x}_K | \boldsymbol{x}_J)} \\ &\quad = \sum_{\boldsymbol{x}_J, \boldsymbol{x}_K} p(\boldsymbol{x}_J, \boldsymbol{x}_K) \sum_y p(y | \boldsymbol{x}_J, \boldsymbol{x}_K) \ln \frac{p(y | \boldsymbol{x}_J, \boldsymbol{x}_K)}{p(y | \boldsymbol{x}_J)}. \end{aligned}$$

Instead of considering the actual post-knockout function κ_J, we now ask how close we can be in principle to the original function, if we allow any post-knockout function κ'_J from the set $\mathcal{K}_J$ of Markov kernels that only use the remaining nodes J as input units. Clearly, the actual post-knockout function κ_J is contained in that set, which gives us the following inequality:

$$\begin{aligned} D_\mu(\kappa \parallel \kappa_J) \;&\geq\; \inf_{\kappa'_J \in \mathcal{K}_J} D_\mu(\kappa \parallel \kappa'_J) \\ &=\; D_\mu(\kappa \parallel \kappa^*_J) \;=\; I(Y; X_K \mid X_J) \;\geq\; 0. \end{aligned} \tag{4}$$

Here, κ^*_J is chosen such that, among all post-knockout functions, it has the smallest achievable deviation from the original function. The geometric interpretation of this minimisation is shown in Fig. 3.

One can show that any such minimiser has to satisfy the following condition.

$$\kappa^*_J(\boldsymbol{x}_J; y) \;=\; p(y | \boldsymbol{x}_J), \qquad \text{whenever } p(\boldsymbol{x}_J) > 0. \tag{5}$$

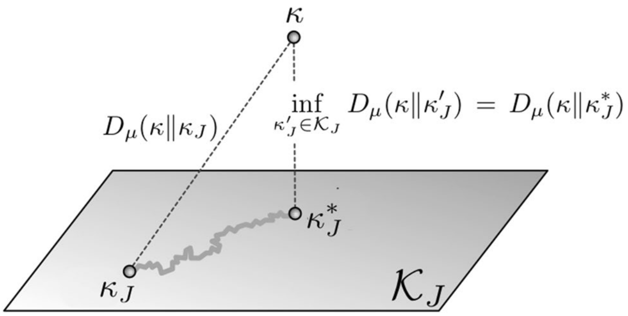
κ
$D_\mu(\kappa \| \kappa_J)$
$\inf_{\kappa'_J \in \mathcal{K}_J} D_\mu(\kappa \| \kappa'_J) \;=\; D_\mu(\kappa \| \kappa^*_J)$
κ^*_J
κ_J
$\mathcal{K}_J$

Fig. 3 Optimal function after knockout

Thus, the post-knockout function is almost uniquely determined. If the exclusion dependence vanishes then, with the estimate (4), we have:

1. The conditional mutual information $I(Y; X_K \mid X_J)$ vanishes. This is equivalent to the conditional independence $Y \perp\!\!\!\perp X_K \mid X_J$. This is a property of μ and κ together.
2. The post-knockout function κ_J is a minimiser of the exclusion dependence, so that it is (almost) uniquely determined by the RHS of (5).

In what follows, we want to require exclusion independence for a number of knockout scenarios, not just restricted to one subset K. Assume that we want the function to be invariant with respect to the deletion of at most k input units. In that case, we have exclusion independence for all K with $|K| \leq k$. Thus,

$$Y \perp\!\!\!\perp X_K \mid X_{I \setminus K} \quad \text{for all } K \subseteq I, |K| \leq k. \tag{6}$$

These conditional independence statements encode in an implicit way those probability distributions that ensure exclusion independence. Again, these conditions refer to the input distribution μ and the Markov kernel κ. The family of post-knockout functions $\kappa_{I \setminus K}, |K| \leq k$, is then (almost) uniquely determined by μ and κ. Using algebraic geometric methods, one can determine those pairs μ and κ that satisfy Eq. (6) (Rauh 2013). An outcome of this analysis is Theorem 1 presented below. It highlights the role of neutrality for robust systems. To be more precise, assume that we have exclusion independence against the knockout of any node set K with at most k elements. Then, for two configurations $(\boldsymbol{x}_J, \boldsymbol{x}_K)$ and $(\boldsymbol{x}_J, \boldsymbol{x}'_K)$ in $\mathcal{S}$, we have

$$\begin{aligned} \kappa(\boldsymbol{x}_J, \boldsymbol{x}_K; y) &= p(y | \boldsymbol{x}_J, \boldsymbol{x}_K) \;=\; p(y | \boldsymbol{x}_J) \\ &= p(y | \boldsymbol{x}_J, \boldsymbol{x}'_K) \;=\; \kappa(\boldsymbol{x}_J, \boldsymbol{x}'_K; y). \end{aligned} \tag{7}$$

Note that this property refers to κ and does not involve the post-knockout function κ_J. It basically says that any function κ that is invariant with respect to knockouts has also to be invariant with respect to mutations (change of the states of

K from $\boldsymbol{x}_K$ to $\boldsymbol{x}'_K$). Mutational neutrality has been the subject of genetic robustness (Arjan et al. 2003; Schuster et al. 1994; Fontana 2006). We are now going to highlight neutrality from a geometric perspective.

Neutrality

Both, the input distribution μ as well as the Markov kernel κ, are involved in the condition (6). We now present a way how to construct κ so that the condition (6) is satisfied, based on a given input distribution μ. Let us assume that we want to delete at most k input nodes while keeping the function (stochastic map) invariant. The Hamming distance of two input configurations $\boldsymbol{x} = (x_1, \dots, x_n)$ and $\boldsymbol{x}' = (x'_1, \dots, x'_n)$ is given by the number of indices i for which $x_i \neq x'_i$. We now connect two configurations $\boldsymbol{x}$ and $\boldsymbol{x}'$ in the support $\mathcal{S}$ of μ if their Hamming distance is smaller than or equal to k. This way, we obtain a graph with node set $\mathcal{S}$. The connected components of this graph play a particularly important role for exclusion independence. It turns out that exclusion independence is achieved by the stochastic map κ, which takes an input configuration from $\mathcal{X}_I$ and generates a stochastic output $y \in \mathcal{Y}$, if and only if it is neutral on these connected components. More precisely, if $\kappa(\boldsymbol{x};y) = \kappa(\boldsymbol{x}';y)$ for all $y \in \mathcal{Y}$ whenever $\boldsymbol{x}$ and $\boldsymbol{x}'$ are from the same connected component, then the condition (6) is satisfied. The values of the Markov kernel outside of $\mathcal{S}$ do not play a role here.

In order to illustrate the construction, let us consider an example. Say that we have only two input units with state sets $\mathcal{X}_1 = \mathcal{X}_2 = \{1, \dots, 16\}$, and assume that out of the 256 joint configurations, some of them are in the support $\mathcal{S}$ of the input distribution μ, as shown in Fig. 4a.

For the illustration of the construction, it is easier to consider the graph metric given by the network shown in Fig. 4 a, instead of the Hamming distance. This yields the induced network where two elements in $\mathcal{S}$ are connected if their distance equals one. The resulting connected components are shown in Fig. 4b. We denote the set of connected components N by $\mathfrak{N}(\mathcal{S})$. Now we are free to define κ in such a way that it outputs the same distribution $\lambda_N(y)$ on $\mathcal{Y}$ for two elements $\boldsymbol{x}$, $\boldsymbol{x}'$ contained in the same connected component N. This neutrality is illustrated by colouring each of the connected components by a single colour in Fig. 4c.

Theorem 1 (see Proposition 3 in Rauh and Ay 2014) *Let $\mathcal{S}$ be a non-empty set of input configurations. We connect two points in $\mathcal{S}$ if their Hamming distance is at most k. The resulting network divides $\mathcal{S}$ into connected components, and we denote their set by $\mathfrak{N}(\mathcal{S})$. Furthermore, consider a probability measure α on $\mathfrak{N}(\mathcal{S})$, and for each $N \in \mathfrak{N}(\mathcal{S})$ a probability measure μ_N on $\mathcal{X}_I$ with support N and a probability measure λ_N on $\mathcal{Y}$. Then the following joint distribution satisfies the conditional independence statements* (6):

$$p(\boldsymbol{x}, y) \; = \sum_{N \in \mathfrak{N}(\mathcal{S})} \alpha(N)\, \mu_N(\boldsymbol{x})\, \lambda_N(y). \tag{8}$$

On the other hand, any joint distribution that satisfies the conditional independence statements (6) *has the structure* (8).

The situation is illustrated in Fig. 5. The result of Theorem 1 highlights the concept of neutrality which is an important concept in robustness studies (Wagner 2007).

Given an input distribution μ, this theorem allows us to construct stochastic maps κ that ensure k-exclusion independence. This is done in several steps.

1. Consider the connected components $\mathfrak{N}(\mathcal{S})$ of the support set $\mathcal{S}$ of μ.
2. For each $N \in \mathfrak{N}(\mathcal{S})$, choose a probability distribution λ_N on $\mathcal{Y}$.
3. For $\boldsymbol{x} \in \mathcal{S}$, we denote by $N_{\boldsymbol{x}}$ the connected component $N \in \mathfrak{N}(\mathcal{S})$ that contains $\boldsymbol{x}$ and define

$$\kappa(\boldsymbol{x};y) \; := \; \lambda_{N_{\boldsymbol{x}}}(y).$$

Outside of $\mathcal{S}$, κ can be chosen arbitrarily.

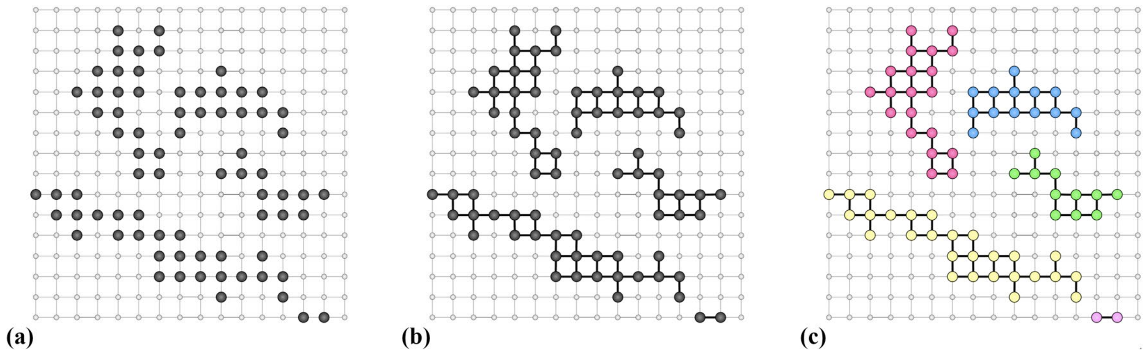

Fig. 4 **a** A set $\mathcal{S}$ in $\mathcal{X}_I$; **b** the connected components of $\mathcal{S}$; **c** neutrality of the map on the connected components, indicated by colour

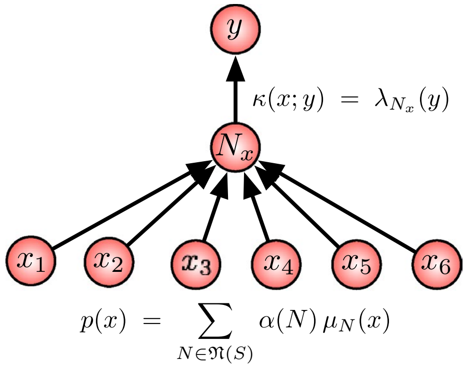
y
$\kappa(x;y) = \lambda_{N_x}(y)$
N_x
x_1 x_2 x_3 x_4 x_5 x_6
$p(x) = \sum_{N \in \mathfrak{N}(S)} \alpha(N)\, \mu_N(x)$

Fig. 5 Structure of a robust map

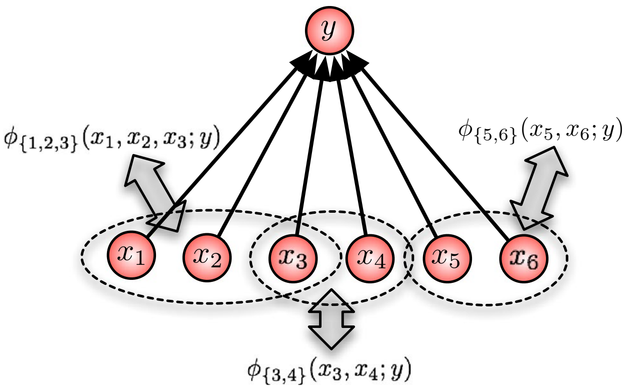
y
$\phi_{\{1,2,3\}}(x_1, x_2, x_3; y)$
$\phi_{\{5,6\}}(x_5, x_6; y)$
x_1 x_2 x_3 x_4 x_5 x_6
$\phi_{\{3,4\}}(x_3, x_4; y)$

Fig. 6 Decomposition of the affinity

Mechanistic modelling of knockout interventions

In what follows, we introduce a particular way of specifying a family (κ_J) of functional modalities. We consider a function $Q(\boldsymbol{x};y)$ that describes the affinity of the node to assume the state y given the input $\boldsymbol{x} = (x_1, \ldots, x_n)$. With such a function, the corresponding Markov kernel can be defined as

$$\kappa(\boldsymbol{x};y) = \frac{e^{Q(\boldsymbol{x};y)}}{\sum_{y'} e^{Q(\boldsymbol{x};y')}}. \tag{9}$$

Here, we adopt the terminology from statistical physics, where the structure (9) of a so-called *Boltzmann–Gibbs distribution* originates from, and refer to the function Q also as an *energy function*. Without any further specification of the energy function, the structure (9) of a stochastic map is quite general. The only restriction comes from the fact that all transition probabilities $\kappa(\boldsymbol{x};y)$ are positive. This implies, for instance, that deterministic functions, such as Boolean functions, cannot be modelled in this way directly. However, these functions can be approximated in terms of stochastic maps of the form (9) arbitrarily well.

One can interpret the structure (9) as a "black-box" description of a stochastic input–output map, which does not incorporate any information about the mechanisms that actually give rise to that map. In fact, one can imagine many mechanisms that generate the same input–output mapping. In general, each of these mechanisms will lead to a different function as result of a deletion of input units. Therefore, the prediction of a post-knockout function requires some level of mechanistic description. We now introduce such a description and demonstrate how it can be used to predict any post-knockout function of the system. The main idea is to decompose the energy function Q into interaction terms $\phi_A(\boldsymbol{x}_A;y)$, thereby employing another concept from statistical physics. Each of these terms quantifies to what extent the input nodes in A work together and thereby contribute to the overall function. Their contribution can be positive or negative, that is excitatory or inhibitory.

As an illustration, consider the example shown in Fig. 6. In this example, there are three interaction terms that contribute to the function Q and thereby to the output y via the stochastic map (9):

$$\begin{aligned} Q(x_1, \ldots, x_6;y) &= \phi_{\{1,2,3\}}(x_1, x_2, x_3;y) \\ &\quad + \phi_{\{3,4\}}(x_3, x_4;y) + \phi_{\{5,6\}}(x_5, x_6;y). \end{aligned} \tag{10}$$

This additional information on how the function Q is composed in individual interaction terms already allows us to model and study knockouts. To illustrate this, let us assume that, for instance, node 3 is knocked out. As it is not present anymore, it cannot interact with any other node. Therefore, it is natural to assume that, as result of the knock-out intervention, all interactions $\phi_A(x_A;y)$ that involve node 3 are now discarded. In our example, this corresponds to removing $\phi_{\{1,2,3\}}(x_1, x_2, x_3;y)$ and $\phi_{\{3,4\}}(x_3, x_4;y)$ from the right-hand side of (10), as illustrated in Fig. 7.With $J := \{1, 2, 4, 5, 6\}$— node 3 is removed—the resulting post-knockout function is then given as

$$\kappa_J(\boldsymbol{x}_J;y) = \frac{e^{\phi_{\{5,6\}}(x_5,x_6;y)}}{\sum_{y'} e^{\phi_{\{5,6\}}(x_5,x_6;y')}}. \tag{11}$$

Note that this is a modification of the stochastic map (9), where all contributions of node three are removed from the energy function Q. Let us discuss two situations here. Let us first assume that the post-knockout function is different from the original function. This implies that node 3 is involved in at least one interaction term, so that we would be able to infer some mechanistic aspects of the role that node 3 plays in generating the output y. However, we cannot conclude that node 3 does not play any role if we do not see a difference between the pre- and post-knockout functions. This invariance of function is an essential requirement for robustness. A robust system, therefore, does not easily reveal

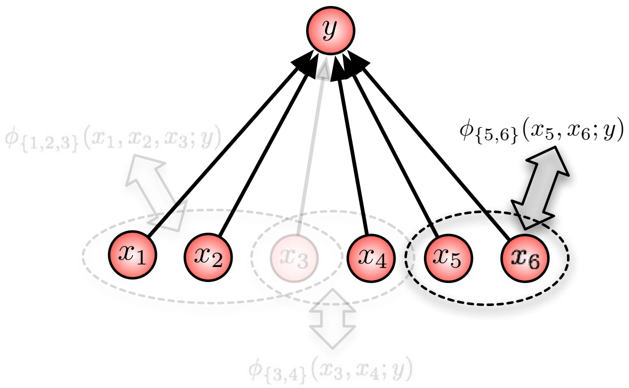
y
$\phi_{\{1,2,3\}}(x_1, x_2, x_3; y)$
$\phi_{\{5,6\}}(x_5, x_6; y)$
x_1
x_2
x_3
x_4
x_5
x_6
$\phi_{\{3,4\}}(x_3, x_4; y)$

Fig. 7 Removal of the interaction terms $\phi_{\{1,2,3\}}(x_1, x_2, x_3;y)$ and $\phi_{\{3,4\}}(x_3, x_4;y)$ as result of the knockout of node 3

its internal mechanisms as a result of knockout interventions. Our simple example already highlights two roles of knockout interventions or perturbations: We can distinguish between experimental knockouts, with which one aims at understanding the inner working of a system, and knockouts as exogenous perturbations that the system has to compensate in order to perform its function. The subjects of system identification and robustness of systems will be discussed in further detail in the following sections. However, let us first extend our instructive example to the general setting of an arbitrary number of input nodes and a general interaction structure.

Assume that the energy function Q can be decomposed in terms of interactions ϕ_A, $A \subseteq I$, that is

$$Q(\boldsymbol{x};y) \;=\; \sum_{A \subseteq I} \phi_A(\boldsymbol{x}_A;y), \tag{12}$$

which corresponds to (10). Consider the situation where a number of input nodes is knocked out and denote the set of remaining nodes by J. Then each interaction term that involves a node from the complement of J will be removed from the decomposition (12), leading to the direct extension of (11)

$$\kappa_J(\boldsymbol{x}_J;y) = \frac{1}{Z(\boldsymbol{x}_J)} \,\exp\left\{ \sum_{A \subseteq J} \phi_A(\boldsymbol{x}_A;y) \right\}, \qquad J \subseteq I, \tag{13}$$

$$\text{with} \qquad Z(\boldsymbol{x}_J) = \sum_{y'} \exp\left\{ \sum_{A \subseteq J} \phi_A(\boldsymbol{x}_A;y') \right\}. \tag{14}$$

This family of stochastic maps, indexed by the subsets J of remaining nodes, describes all post-knockout functions, which represents one instance of the *functional modalities* we discussed above. As we will see, this is a quite general mechanistic perspective in the sense that any family of strictly positive functional modalities can be represented this way.

Let us further illustrate this approach to modelling post-knockout functions in terms of the following example from the field of neural networks.

Example 1 We consider a neuron which receives an input $\boldsymbol{x} = (x_1, \ldots, x_n) \in \{-1, +1\}^n$ and generates the output $+1$ with probability

$$\kappa(\boldsymbol{x}; +1) \;:=\; \frac{1}{1 + e^{-\sum_{i=1}^{n} w_i x_i}}. \tag{15}$$

For a general output $y \in \{-1, +1\}$, this implies

$$\kappa(\boldsymbol{x};y) \;:=\; \frac{e^{\frac{1}{2}\sum_{i=1}^{n} w_i x_i y}}{e^{\frac{1}{2}\sum_{i=1}^{n} w_i x_i \cdot (-1)} + e^{\frac{1}{2}\sum_{i=1}^{n} w_i x_i \cdot (+1)}}. \tag{16}$$

This representation of the stochastic map κ has a structure that allows us to infer the function after a knockout of a set K of input nodes, by simply removing the contribution of all the nodes in K. This leads to

$$\kappa_J(\boldsymbol{x}_J; +1) \;=\; \frac{1}{1 + e^{-\sum_{i \in J} w_i x_i}},$$

where $J = I \backslash K$. This inference of the post-knockout function is based on the decomposition of the sum that appears in the enumerator on the RHS of (16), $\frac{1}{2}\sum_{i=1}^{n} w_i x_i y$. Mapping this to the representation (12), we clearly have $\phi_A = 0$ whenever $|A| \neq 1$, and $\phi_{\{i\}}(x_i;y) = \frac{1}{2} w_i x_i y$, $i = 1, \ldots, n$.

Knockout interventions for system identification

If the removal of some part of the system causes the loss or disturbance of one of its functions, then it is reasonable to attribute that function to the removed part. This intuition underlies, for instance, some of our understanding of brain function which has been deduced from brain lesions. Such lesions are typically caused by an injury or disease. However, the same fundamental intuition also underlies knockout experiments that aim at the identification of the system's mechanisms. One well-known example is given by gene knockouts.

Let us now study this intuition based on our formal model of knockout interventions which we developed in "Mechanistic modelling of knockout interventions" section. We begin with the example shown in Fig. 6. We have already discussed the knockout of node 3, which gave us the post-knockout function (11). This implies

$$\ln \kappa_{\{1,2,4,5,6\}}(x_1, x_2, x_4, x_5, x_6;y) \\ = Z(x_1, x_2, x_4, x_5, x_6) + \phi_{\{5,6\}}(x_5, x_6;y). \tag{17}$$

If we knock out nodes 4 and 5, the remaining contribution is from the nodes 1, 2, and 3:

$$\ln \kappa_{\{1,2,3\}}(x_1, x_2, x_3;y) \\ = Z(x_1, x_2, x_3) + \phi_{\{1,2,3\}}(x_1, x_2, x_3;y). \tag{18}$$

Finally, for the knockout of nodes 1 and 5, the remaining contribution comes from nodes 3 and 4:

$$\ln \kappa_{\{3,4\}}(x_3, x_4;y) \\ = Z(x_3, x_4) + \phi_{\{3,4\}}(x_3, x_4;y). \tag{19}$$

This shows that we can recover the interactions, up to some function of $\boldsymbol{x}$, given by the normalisations Z, if we know the post-knockout function of some appropriately chosen knockouts. This backs up the strategy of system identification in terms of knockout experiments. The non-uniqueness of the recovered interaction terms, due to the functions that only depend on the input $\boldsymbol{x}$, does not harm here. Any family of interaction terms will represent the functional modalities κ_J, $J \subseteq I$, equally well.

Note that the knock-outs that reveal the individual interaction terms are also not unique. There are several appropriate knockout protocols. For instance, in the above example, we could also consider the three knockouts of node 1, node 4, and node 5, leading to a linear system

$$\begin{aligned} &\ln \kappa_{\{2,3,4,5,6\}}(x_2, x_3, x_4, x_5, x_6;y) \\ &\quad = Z(x_2, x_3, x_4, x_5, x_6) \\ &\qquad + \phi_{\{3,4\}}(x_3, x_4;y) + \phi_{\{5,6\}}(x_5, x_6;y), \\ &\ln \kappa_{\{1,2,3,5,6\}}(x_1, x_2, x_3, x_5, x_6;y) \\ &\quad = Z(x_1, x_2, x_3, x_5, x_6) \\ &\qquad + \phi_{\{1,2,3\}}(x_1, x_2, x_3;y) + \phi_{\{5,6\}}(x_5, x_6;y), \\ &\ln \kappa_{\{1,2,3,4,6\}}(x_1, x_2, x_3, x_4, x_6;y) \\ &\quad = Z(x_1, x_2, x_3, x_4, x_6) \\ &\qquad + \phi_{\{1,2,3\}}(x_1, x_2, x_3;y) + \phi_{\{3,4\}}(x_3, x_4;y). \end{aligned}$$

This can be easily solved, and we obtain again the interactions from the post-knockout functions. Altogether, we have shown that two different knockout protocols, each of them involving only three knockout interventions, allow us to identify the three interaction terms of the system. In general, the situation is much more complex. On the one hand, there can be many more interaction terms involved so that we will require correspondingly many knockout interventions. We cannot assume that performing these interventions will always be feasible. On the other hand, even if only a few interactions are actually involved, without prior knowledge on these interactions it is not possible to determine a correspondingly small set of knockouts that would allow us to reveal the interactions within the system. Let us assume, for the moment, that we can perform all possible knockouts and measure the corresponding post-knockout functions. In theory, the resulting set of equations can be solved in terms of the Möbius inversion. Let us be more precise: Clearly, each κ_J in (13) is strictly positive. Using the Möbius inversion, it is easy to see that each strictly positive family (κ_J) has such a representation. In order to see this, we simply set

$$\phi_A(\boldsymbol{x}_A, y) \;:=\; \sum_{J \subseteq A} (-1)^{|A \setminus J|} \ln \kappa_J(\boldsymbol{x}_J;y)\,. \tag{20}$$

Note that this representation is not unique: If an arbitrary function of $\boldsymbol{x}_A$ is added to the function ϕ_A, then the functional modalities remain unchanged. We have a one-to-one correspondence, modulo this ambiguity, between functional modalities and interactions:

$$\kappa_J,\ J \subseteq I \;\longleftrightarrow\; \phi_A,\ A \subseteq I.$$

In this section, we do not interpret a knockout as a natural perturbation but as an experimental test. The intention is to reveal the inner working, the mechanisms, of the system based on its response to these unnatural perturbations. Indeed, we can identify important aspects of the interactions. This is remarkable, because these interactions provide the basis of the unperturbed function. Thus, we perturb the system in order to understand the constituents of its unperturbed function. This provides a formal basis for knockout experiments.

It is important to note that the choice of the ϕ_A, $A \subseteq I$, for describing knockouts is essential. If we choose to represent a post-knockout function in terms of different interactions, say ψ_B, $B \subseteq I$, we have to translate the effect of knockouts accordingly. To be more precise, assume that we can represent the initial functions as

$$\phi_A(\boldsymbol{x}_A;y) \;=\; \sum_{B \subseteq A} \alpha_{A,B}\, \psi_B(\boldsymbol{x}_B;y), \qquad A \subseteq I. \tag{21}$$

This implies

$$\ln \kappa_J(\boldsymbol{x}_J;y) \sim \sum_{A \subseteq J} \phi_A(\boldsymbol{x}_A;y) \qquad \text{(by (13))} \tag{22}$$

$$\begin{aligned} &= \sum_{A \subseteq J} \sum_{B \subseteq A} \alpha_{A,B}\, \psi_B(\boldsymbol{x}_B;y) \qquad \text{(by (21))} \\ &= \sum_{B \subseteq J} \left(\sum_{C \subseteq J \setminus B} \alpha_{B \cup C, B} \right) \psi_B(\boldsymbol{x}_B;y) \\ &= \sum_{B \subseteq J} \beta_{J,B}\, \psi_B(\boldsymbol{x}_B;y), \end{aligned} \tag{23}$$

where we define the coefficients $\beta_{J,B}$ accordingly. This calculation highlights the following fact: When describing the post-knockout function κ_J in terms of the ϕ_A we simply remove all terms ϕ_A for which A is not contained in J, as described by (22). This is the basic requirement that defines the ϕ_A, $A \subseteq I$. When trying to represent the post-knockout function in terms of different interactions, ψ_B, $B \subseteq I$, this rule changes. In order to see this, let us analyse the representation (23). Also in this case, the terms ψ_B do not appear whenever B is not contained in J. However, the coefficients $\beta_{J,B}$ change as a consequence of a knockout. That means, if we have an expansion of $\ln \kappa(\boldsymbol{x};y)$, where κ is the unperturbed function, into terms ψ_B, with weights $\beta_{I,A}$, then simply removing weighted terms as result of a knockout is not sufficient.

Robustness and interaction order

Let (κ_J) be functional modalities, which we assume to be strictly positive. We know that they can be represented in terms of interactions, by the choice (20). How is exclusion independence mapped to this representation? In order to answer this question, let us consider the situation where at most k input nodes are knocked out. In this case, k-exclusion independence implies that for $|J| \geq r := n - k$, $B \subset J$ with $|B| = r$, and $\boldsymbol{x} \in \mathcal{S}$, we have the equality $\kappa_J(\boldsymbol{x}_J;\cdot) = \kappa_B(\boldsymbol{x}_B;\cdot)$. Below we use the Möbius inversion (20) for the derivation of the interaction terms $\phi_A(\boldsymbol{x}_A;y)$ under the assumption of k-exclusion independence. In order to do so, we need to evaluate the following sum:

$$
\begin{aligned}
&\sum_{\substack{J \subseteq A \\ |J| \geq r}} (-1)^{|A \setminus J|} \ln \kappa_J(\boldsymbol{x}_J;y) \\
&= \sum_{\substack{J \subseteq A \\ |J| \geq r}} (-1)^{|A \setminus J|} \frac{1}{\binom{|J|}{r}} \sum_{\substack{B \subseteq J \\ |B| = r}} \ln \kappa_J(\boldsymbol{x}_J;y) \\
&= \sum_{\substack{J \subseteq A \\ |J| \geq r}} (-1)^{|A \setminus J|} \frac{1}{\binom{|J|}{r}} \sum_{\substack{B \subseteq J \\ |B| = r}} \ln \kappa_B(\boldsymbol{x}_B;y) \\
&= \sum_{\substack{B \subseteq A \\ |B| = r}} \left\{ \sum_{R \subseteq A \setminus B} (-1)^{|A|-|R|-r} \frac{1}{\binom{|R|+r}{r}} \right\} \ln \kappa_B(\boldsymbol{x}_B;y).
\end{aligned}
$$

Together with (20) this gives us

$$
\begin{aligned}
\phi_A(\boldsymbol{x}_A;y) = & \sum_{\substack{J \subseteq A \\ |J| < r}} (-1)^{|A \setminus J|} \ln \kappa_J(\boldsymbol{x}_J;y) \\
& + \sum_{\substack{J \subseteq A \\ |J| \geq r}} (-1)^{|A \setminus J|} \ln \kappa_J(\boldsymbol{x}_J;y) \\
= & \sum_{\substack{J \subseteq A \\ |J| \leq r}} \alpha_{A,J} \ln \kappa_J(\boldsymbol{x}_J;y),
\end{aligned}
\tag{24}
$$

where

$$
\alpha_{A,J} = \begin{cases} (-1)^{|A|-|J|}, & \text{if } |J| < r \\ \sum_{R \subseteq A \setminus J} (-1)^{|A|-|R|-r} \frac{1}{\binom{|R|+r}{r}}, & \text{if } |J| = r \end{cases}
$$

depends only on the cardinalities of A and J. The representation (24) shows that k-exclusion independence implies a special structure of the underlying interactions. It suggests to consider the family of functional modalities (κ_J) given in terms of interactions ϕ_A with the structure

$$
\phi_A(\boldsymbol{x}_A, y) := \sum_{\substack{J \subseteq A \\ |J| \leq r}} \alpha_{A,J} \, \psi_J(\boldsymbol{x}_J;y),
\tag{25}
$$

where ψ_J are arbitrary functions that do not depend on x_i, $i \notin J$. Note that these functions are not indexed by A. Given this structure, we have

$$
\begin{aligned}
\ln \kappa_J(\boldsymbol{x}_J;y) \sim & \sum_{A \subseteq J} \phi_A(\boldsymbol{x}_A, y) \\
= & \sum_{A \subseteq J} \sum_{\substack{J' \subseteq A \\ |J'| \leq r}} \alpha_{A,J'} \, \psi_{J'}(\boldsymbol{x}_{J'};y) \\
= & \sum_{\substack{B \subseteq J \\ |B| \leq r}} \beta_{J,B} \, \psi_B(\boldsymbol{x}_B;y),
\end{aligned}
\tag{26}
$$

which represents a special case of (23). This proves that all functional modalities of a robust system involve interactions of order at most r. We denote by $\mathcal{M}_r$ the set of functional modalities (κ_J) that have the representation (26). Note that in this representation, each κ_J can have different coefficients $\beta_{J,B}$. More precisely, if ψ_B appears in the representation of κ_J and $\kappa_{J'}$, the corresponding coefficients $\beta_{J,B}$ and $\beta_{J',B}$ will be typically different. Summarising, we obtain the following result.

Theorem 2 (see Theorem 15 in Rauh and Ay 2014) *Let* (κ_J) *be a family of strictly positive functional modalities that is*

k-exclusion independent in all $\boldsymbol{x} \in \mathcal{S}$. Then there exists a family (κ'_J) of functional modalities in $\mathcal{M}_r$, with $r = n - k$, such that $\kappa_J(\boldsymbol{x};\cdot) = \kappa'_J(\boldsymbol{x};\cdot)$ for all $\boldsymbol{x} \in \mathcal{S}$ and all $J \subseteq I$.

This result implies that, in order to achieve k-exclusion independence, the order of interaction should not be too large. Such an interaction will break down as result of a knockout so that the remaining units will not be able to compensate it. On the other hand, if the interactions are small enough, their removal can be compensated. This compensation is achieved by an appropriate change of the coefficients β in the representation (26). To be more precise, say that the interaction ψ_B has a coefficient $\beta_{I,B}$ in the unperturbed function $\kappa = \kappa_I$ and assume that the node set K is knocked out, where B is still a subset of the remaining set $J = I \backslash K$. Thus, in the post-knockout function κ_J, the term ψ_B will still be involved. However, the coefficient $\beta_{J,B}$ will typically change in order to ensure the invariance of the function. Therefore, in order to design a robust system, we need the whole family of coefficients $\beta_{J,B}$. This would allow the system to react to a knockout by immediately switching to the right post-knockout function κ_J. Alternatively, the knockout might initiate a re-adaptation process, which takes time to recover the original function. In this case, it is not required to hard-wire all coefficients $\beta_{J,B}$ into the system. It is then sufficient to store the coefficients involved in the original function, that is $\beta_{I,B}$. The other coefficients are obtained in terms of the recovery process. Such a process is particularly relevant within the study of biological robustness mechanisms.

Conclusions

We have reviewed results related to the robustness of function against knockout perturbations. In comparison with previous works, this article tries to draw a conceptually more complete picture, relying on existing theoretical results. On the one hand, it provides a solid justification of knockout experiments for system identification. On the other hand, it highlights the role of neutrality and low interaction order in the context of robust systems. These properties are associated with two corresponding characterisations of k-exclusion independent systems, Theorems 1 and 2. Thereby, neutrality and low interaction order, two at first sight seemingly unrelated aspects, are revealed to be equivalent. Even though we have this equivalence, we highlight an important difference in the corresponding representations. In the "neutrality representation" (8) the map κ is only specified on the support $\mathcal{S}$ of the input distribution μ. The output of the map κ for configurations in the space between the connected components $N \in \mathfrak{N}(\mathcal{S})$, on which κ is constant, can be arbitrarily chosen. This is fundamentally different in the "interaction representation" (26) of κ, that is for $J = I$. It is indeed possible that the sum $\sum_{\substack{B \subseteq I \\ |B| \leq r}} \beta_{I,B}\, \psi_B(\boldsymbol{x}_B;y)$ is well-defined even for $\boldsymbol{x} \notin \mathcal{S}$. The only requirement for $\boldsymbol{x}$ is that for each restriction $\boldsymbol{x}_B$ with $|B| \leq r$, there is a configuration $\boldsymbol{x}'_{I \backslash B}$ such that the concatenation $(\boldsymbol{x}_B, \boldsymbol{x}'_{I \backslash B})$ is in the support $\mathcal{S}$ of the input distribution. Thus, the "interaction representation" comes with an extrapolation of the values on the connected components $N \in \mathfrak{N}(\mathcal{S})$ to some part of the space in between. The role of such an extrapolation for robustness is currently unclear. It might allow the system to generalise to and deal with unforeseen situations in a reasonable way. This is absolutely not possible with the "neutrality representation", which is nothing but a lookup table for all $\boldsymbol{x} \in \mathcal{S}$, with no restriction at all for $\boldsymbol{x} \notin \mathcal{S}$. Coping with unforeseen situations in itself is a robustness property, which, however, is different from the way we have defined robustness in this article. Thus, the presented formalism might allow us to compare different notions of robustness. We already discussed one instance of this at the end of "Exclusion dependence" section: Invariance of function with respect to knockout perturbations implies invariance of function with respect to mutations.

Acknowledgements The author is grateful for stimulating discussions with David Krakauer, Jessica Flack, and Johannes Rauh.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Amari S (2016) Information geometry and its applications, vol 194. Springer, Berlin

Amari S, Nagaoka H (2000) Methods of information geometry. Oxford University Press, Oxford

Ay N, Amari S (2015) A novel approach to canonical divergences within information geometry. Entropy 17(12):8111–8129

Ay N, Krakauer DC (2007) Geometric robustness theory and biological networks. Theory Biosci 125(2):93–121

Ay N, Flack J, Krakauer DC (2007) Robustness and complexity coconstructed in multimodal signalling networks. Philos Trans R Soc B Biol Sci 362(1479):441–447

Ay N, Jost J, Vân Lê H, Schwachhöfer L (2017) Information geometry. Springer, Berlin

Boldhaus G, Bertschinger N, Rauh J, Olbrich E, Klemm K (2010) Robustness of boolean dynamics under knockouts. Phys Rev E 82(2):021916

de Visser JA, Hermisson J, Wagner GP, Meyers LA, Bagheri-Chaichian H, Blanchard JL, Chao L, Cheverud JM, Elena SF, Fontana W, Gibson G (2003) Perspective: evolution and detection of genetic robustness. Evolution 57(9):1959–1972

Flack JC, Girvan M, De Waal FB, Krakauer DC (2006) Policing stabilizes construction of social niches in primates. Nature 39(26):426–429

Fontana W (2006) The topology of the possible. Palgrave Macmillan, London, pp 67–84

Krakauer DC, Page KM, Erwin DH (2009) Diversity, dilemmas, and monopolies of niche construction. Am Nat 173(1):26–40

Krakauer DC, Flack J, Ay N (2010) Probabilistic design principles for robust multi-modal communication networks. Cambridge University Press, Cambridge, pp 255–268

Rauh J (2013) Generalized binomial edge ideals. Adv Appl Math 50(3):409–414

Rauh J, Ay N (2014) Robustness, canalyzing functions and systems design. Theory Biosci 133(2):63–78

Schuster P, Fontana W, Stadler PF, Hofacker IL (1994) From sequences to shapes and back: a case study in RNA secondary structures. Proc R Soc Lond Ser B Biol Sci 255(1344):279–284

Wagner A (2007) Robustness and evolvability in living systems. Princeton University Press, Princeton