

RESEARCH

Open Access



# Computational intractability law molds the topology of biological networks

Ali A. Atiia<sup>1,2\*</sup> , Corbin Hopper<sup>1,3</sup>, Katsumi Inoue<sup>4,5</sup>, Silvia Vidal<sup>2</sup> and Jérôme Waldispühl<sup>1</sup>

\*Correspondence: [atiia@cs.mcgill.ca](mailto:atiia@cs.mcgill.ca)

**Author Summary:** Our results indicate that the topology of molecular interaction networks is a selected-for adaptation that minimizes the evolutionary cost of re-wiring the network in response to an evolutionary pressure to conserve, delete or mutate existing genes and interactions.

<sup>1</sup>School of Computer Science, McGill University, Montreal, Canada  
<sup>2</sup>Research Centre on Complex Traits, McGill University, Montreal, Canada  
Full list of author information is available at the end of the article

## Abstract

Virtually all molecular interaction networks (MINs), irrespective of organism or physiological context, have a majority of loosely-connected 'leaf' genes interacting with at most 1-3 genes, and a minority of highly-connected 'hub' genes interacting with at least 10 or more other genes. Previous reports proposed adaptive and non-adaptive hypotheses describing sufficient but not necessary conditions for the origin of this majority-leaves minority-hubs (mLmH) topology. We modelled the evolution of MINs as a computational optimization problem which describes the cost of conserving, deleting or mutating existing genes so as to maximize (minimize) the overall number of beneficial (damaging) interactions network-wide. The model 1) provides sufficient and, assuming  $\mathcal{P} \neq \mathcal{NP}$ , necessary conditions for the emergence of mLmH as an adaptation to circumvent computational intractability, 2) predicts the percentage number of genes having  $d$  interacting partners, and 3) when employed as a fitness function in an evolutionary algorithm, produces mLmH-possessing synthetic networks whose degree distributions match those of equal-size MINs.

**Keywords:** Biological networks, Systems biology, Emergence, Computational complexity, Computational intractability, Combinatorial optimization, Evolutionary algorithms, Evolutionary adaptations

## Introduction

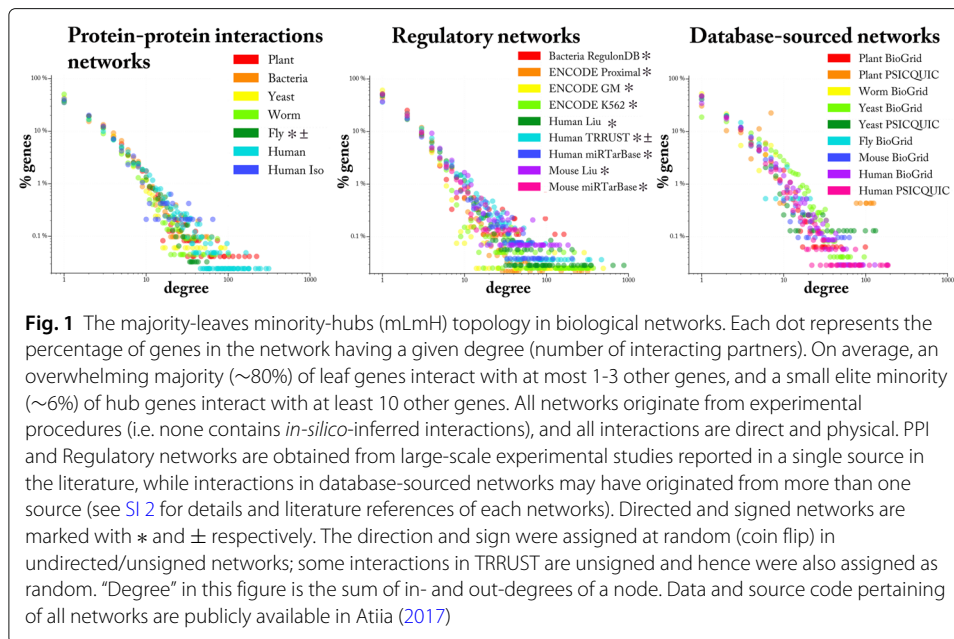
Molecular interaction networks (MINs) are typically represented as graphs where nodes represent proteins, nucleic acids, or metabolites and edges represent physical or functional interactions. By 'functional' we refer to the indirect effect that some gene  $g_i$  has on another  $g_j$ . None of the edges in the networks featured in this study represents a functional relationship. Instead, each edge represents a direct and physical interaction between  $g_i$  and  $g_j$ . The steady increase in scale (Rolland et al. 2014) and resolution (Yang et al. 2016) of experimentally validated interactions has not been matched with theoretical progress towards deciphering the underlying evolutionary forces shaping the structure of MINs. Earlier studies aimed to empirically show that MINs possess a certain property (say, small-world connectivity or power-law-fitted degree distribution), and subsequently advocate for the existence of a universal design principle behind it. However, the robustness of the observations and conclusions of these studies Barabási and Albert (1999); Fell and

Wagner (2000) has been questioned Arita (2004); Tanaka et al. (2005); Fox Keller (2005); Khanin and Wit (2006), and so too Stelling et al. (2004); Hahn et al. (2004) has the validity of the design principles (Albert et al. 2000; Barabási and Oltvai 2004) they inspired. Assuming those properties were indeed real, the universality of proposed design principles around them would still need to be justified. The statistical support of a property "is no evidence of universality without a concrete underlying theory to support it" (Stumpf and Porter 2012). Furthermore, a universal theory "must facilitate the inclusion of domain mechanisms and details" yet there is a sharp disconnect between current hypotheses' high abstractions and actual functional aspects of evolving biological systems (Alderson and Doyle 2010). A 'software' trait relates to *relationships* between genes (connectivity, community clustering, co-expression etc) rather than to their molecular ('hardware') properties (e.g. their amino acid sequence or 3D conformation). The assumption that natural selection can lead to the emergence of advantageous network-level software traits has itself been challenged by the non-adaptive hypothesis: the topology of MINs could be an indirect result of selection pressure on other traits (Papp et al. 2009) or a mere byproduct of non-adaptive evolutionary forces such as mutation and genetic drift (Lynch 2007; Sorrells and Johnson 2015). Both the adaptive and non-adaptive hypotheses present sufficient but not necessary conditions for the emergence of network traits and therefore one cannot objectively rule out the plausibility of the other.

### Perspective

A fundamental property in MINs is the majority-leaves minority-hubs (mLmH) topology whereby an overwhelming ~80% majority of leaf genes interact with at most 1-3 other genes, and an elite ~6% minority of hub genes interact with at least 10 other genes (these figures are the average over the 25 networks featured in this article which, to our knowledge, represent all large-scale experimentally validated networks of direct physical interactions available in the literature). mLmH-possessing networks are typically referred to as 'scale-free' (SF). Because of the association between the SF term and the controversial proposition (Arita 2004; Tanaka et al. 2005; Fox Keller 2005; Khanin and Wit 2006) that biological networks follow a power-law distribution, we use the loose term mLmH to emphasize the fact that our model does not assume, require nor advocate for or against the idea that the degree distribution of MINs follows a power-law in a strict technical sense.

mLmH is observed in virtually all MINs irrespective of organism or physiological context as shown in Fig. 1. Here we investigate the mLmH topology from a computational complexity perspective and derive necessary and sufficient conditions for its emergence. We assume that the organism is under some evolutionary pressure to change. Under this hypothetical scenario, we assume it critical for the system, in some regulatory state at some particular point in evolutionary time, that some genes be promoted and some be inhibited. An interaction is deemed beneficial if it promotes or inhibits a gene that should indeed be promoted or inhibited, respectively. Similarly, an interaction is deemed damaging if it is promotional or inhibitory towards a gene that should rather be inhibited or promoted, respectively. We assign a benefit (damage) score for each gene  $g_i$  as the sum of beneficial (damaging) interactions that  $g_i$  is *projecting onto* or *attracting from* other genes in the network. If gene  $g_i$  promotes or inhibits  $g_j$ , we say that  $g_i$  is projecting an interaction

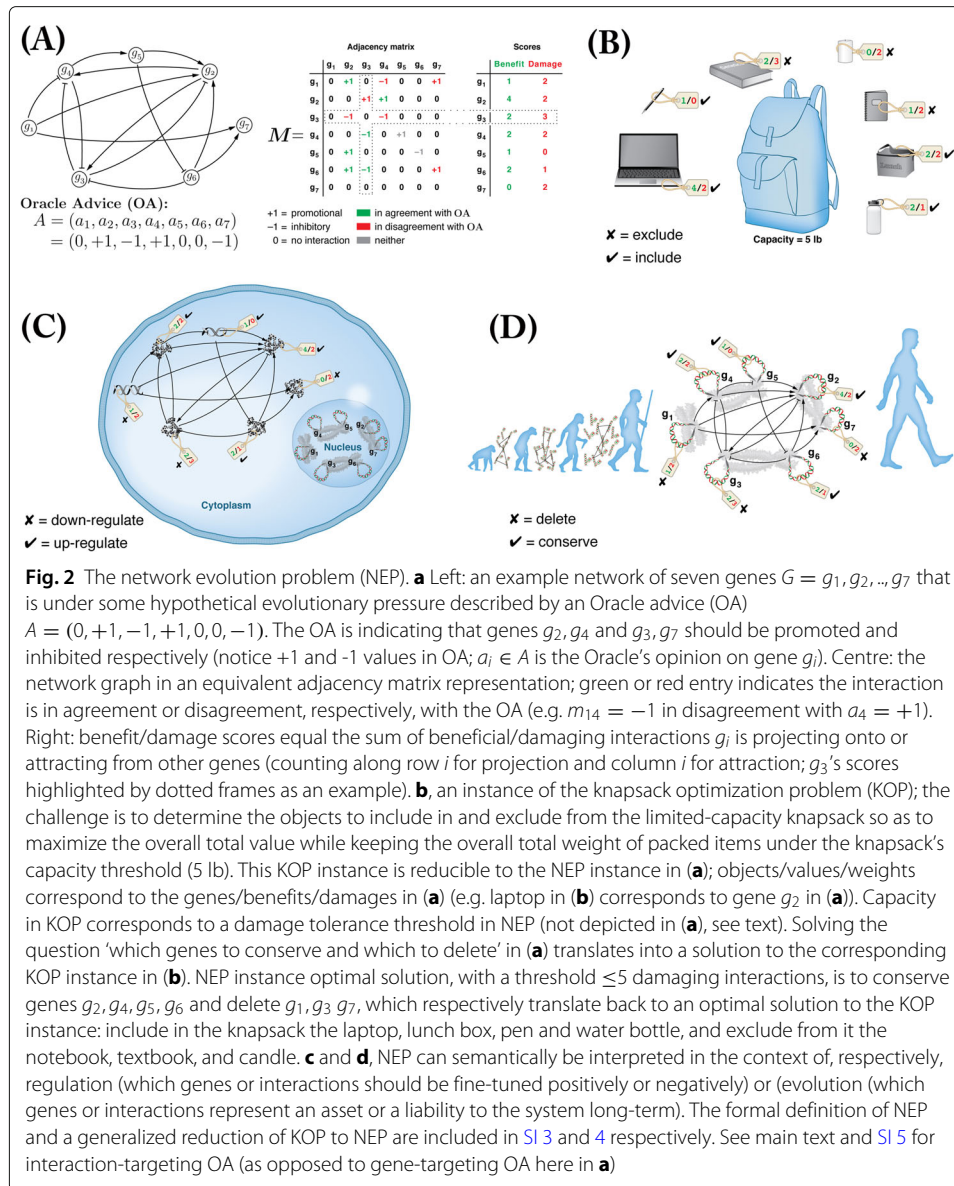


onto  $g_j$ , and  $g_j$  is attracting an interaction from  $g_i$ . The benefit or damage scores of both  $g_i$  and  $g_j$  is incremented by some  $\rho \in \mathbb{R}$ . If such an interaction is beneficial or damaging, respectively.  $\rho$  signifies the potency of such an interaction. A beneficial/damaging interaction therefore adds  $\rho$  to the total benefit or damage score of both the source and target genes. A gene may for example be projecting very few beneficial interactions while attracting many damaging interactions (making it more of a liability) and vice versa. Hence the benefit/damage scoring is not necessarily affected by how skewed the in- versus out-degree distribution of a gene is. Because we have no knowledge of what  $\rho$  is (i.e. a measure of interaction potency) at a global scale in MINs, we use  $\rho = 1$  for all interactions in this study.

Given the benefit and damage scores of genes under the current evolutionary pressure scenario, how hard of a computational problem would it be to determine the optimal immediate "next-move" for the system, i.e. which genes to conserve, mutate or delete, such that the overall total number of beneficial interactions is maximal *while* the total number of damaging interactions is minimal to a threshold? We refer to this optimization question as the network evolution problem (NEP).

### The network evolution problem

The nodes in a MIN correspond to a set of genes  $G = \{g_1, g_2, \dots, g_n\}$ , and the directed and signed edges represent interactions. The direction of an edge between  $g_j$  and  $g_k$  denotes which of the two genes is the source and which is the target. The nodes in the hypothetical MIN shown Fig. 2a, left, represent 7 genes, with promotional and inhibitory interactions denoted by arrow- and bar-terminated edges, respectively. A hypothetical 'Oracle advice' (OA) on all or some of the genes simulates the evolutionary pressure on the network. The OA is represented as a ternary sequence  $A = (a_1, a_2, \dots, a_n)$  where  $a_j = +1$  or  $a_j = -1$  indicates that the Oracle says  $g_j$  should be promoted or inhibited, respectively.  $a_j = 0$  implies the Oracle is neutral towards  $g_j$ . In Fig. 2a example, the OA



**Fig. 2** The network evolution problem (NEP). **a** Left: an example network of seven genes  $G = g_1, g_2, \dots, g_7$  that is under some hypothetical evolutionary pressure described by an Oracle advice (OA)  $A = (0, +1, -1, +1, 0, 0, -1)$ . The OA is indicating that genes  $g_2, g_4$  and  $g_3, g_7$  should be promoted and inhibited respectively (notice +1 and -1 values in OA;  $a_i \in A$  is the Oracle's opinion on gene  $g_i$ ). Centre: the network graph in an equivalent adjacency matrix representation; green or red entry indicates the interaction is in agreement or disagreement, respectively, with the OA (e.g.  $m_{14} = -1$  in disagreement with  $a_4 = +1$ ). Right: benefit/damage scores equal the sum of beneficial/damaging interactions  $g_i$  is projecting onto or attracting from other genes (counting along row  $i$  for projection and column  $i$  for attraction;  $g_3$ 's scores highlighted by dotted frames as an example). **b**, an instance of the knapsack optimization problem (KOP); the challenge is to determine the objects to include in and exclude from the limited-capacity knapsack so as to maximize the overall total value while keeping the overall total weight of packed items under the knapsack's capacity threshold (5 lb). This KOP instance is reducible to the NEP instance in **(a)**; objects/values/weights correspond to the genes/benefits/damages in **(a)** (e.g. laptop in **(b)** corresponds to gene  $g_2$  in **(a)**). Capacity in KOP corresponds to a damage tolerance threshold in NEP (not depicted in **(a)**, see text). Solving the question 'which genes to conserve and which to delete' in **(a)** translates into a solution to the corresponding KOP instance in **(b)**. NEP instance optimal solution, with a threshold  $\leq 5$  damaging interactions, is to conserve genes  $g_2, g_4, g_5, g_6$  and delete  $g_1, g_3, g_7$ , which respectively translate back to an optimal solution to the KOP instance: include in the knapsack the laptop, lunch box, pen and water bottle, and exclude from it the notebook, textbook, and candle. **c** and **d**, NEP can semantically be interpreted in the context of, respectively, regulation (which genes or interactions should be fine-tuned positively or negatively) or evolution (which genes or interactions represent an asset or a liability to the system long-term). The formal definition of NEP and a generalized reduction of KOP to NEP are included in SI 3 and 4 respectively. See main text and SI 5 for interaction-targeting OA (as opposed to gene-targeting OA here in **a**)

is  $A = (0, +1, -1, +1, 0, 0, -1)$ , meaning the Oracle says genes  $g_2, g_4$  should be promoted (notice  $a_2 = a_4 = +1$ ) while  $g_3, g_7$  should be inhibited ( $a_3 = a_7 = -1$ ). The Oracle has no opinion towards genes  $g_1, g_5$  and  $g_6$  ( $a_1 = a_5 = g_6 = 0$ ) in this example.

The adjacency matrix  $M$  in Fig. 2a, centre, is an equivalent representation of the network graph on the left. There is a non-zero entry  $m_{jk}$  for each edge between genes  $g_j$  and  $g_k$ . A promotional or inhibitory interaction is represented in  $M$  as a +1 or -1 entry, respectively. A green- or red-coloured matrix entry  $m_{jk}$  denotes whether the interaction between  $g_j$  and  $g_k$  is in agreement or disagreement with the OA on  $g_k$  (i.e. whether  $a_k = m_{jk}$ ). While  $m_{jk}$  describes what the effect of  $g_j$  on  $g_k$  actually is,  $a_k$  describes what that effect *should* ideally be.

For example,  $g_1$  promotes  $g_2$  ( $m_{12} = +1$ ) in agreement with the OA that  $g_2$  should be promoted ( $a_2 = +1$ ). On the other hand,  $g_1$  inhibits  $g_4$  ( $m_{14} = -1$ ) in disagreement with the OA on  $g_4$  ( $a_4 = +1$ ). The benefit/damage ( $b/d$ ) score of a gene  $g_i$  is calculated as the

sum of green/red interactions along row  $i$  (projection) and column  $i$  (attraction), as shown in the right tabular in Fig. 2a. The benefit and damage scores of  $g_3$ , and the corresponding row and column from which they have been summed up, are highlighted with dotted frame in Fig. 2a. Notice that although the Oracle is neutral towards  $g_1, g_5$  and  $g_6$ , they do have  $b/d$  values depending on their interactions with other genes and whether those interactions are in agreement with the OA or not. Given the set of genes with non-zero  $b/d$  scores, the optimization problem (formally defined in SI 3) is:

*What subset of genes should be conserved and which should be mutated/deleted so as to maximize the total number of beneficial interactions network-wide, while minimizing (to a threshold) the total number of damaging interactions?*

### The computational complexity of NEP

The computational complexity of NEP is equivalent to that of the well-known  $\mathcal{NP}$ -complete knapsack optimization problem (KOP). Figure 2b shows a KOP instance with 7 items, each tagged with certain values reflecting its utility and weight in pounds. Values and weights are indicated on items' tags with green and red numbers respectively. The optimization question in KOP is how to pack the knapsack with as many items as possible such that the total value of items included in the knapsack is maximal while their total weight does not exceed the knapsack maximum capacity, which is 5 pounds in this KOP instance. Some items have negligible weight and some are useless, such as pen and candle in this example, respectively. Clearly the pen (candle) should be included in (excluded from) the knapsack regardless, and as such need not be considered in the optimization search. Among the remaining objects, a search through the include/exclude combinations must be explored to determine the optimal solution.

We say that the KOP instance in Fig. 2b is reduced to the NEP instance in (A). Nodes  $g_1, g_2, \dots, g_7$  and their  $b/d$  scores in (A) respectively correspond to the items and value/weight scores of the KOP items (B). For example, KOP's laptop corresponds to NEP's gene  $g_2$ . The optimal solution to the NEP instance in Fig. 2a would be to conserve  $g_2, g_4, g_5$  and  $g_6$  and to mutate or delete  $g_1, g_3$  and  $g_7$ , assuming the maximum threshold of tolerable damaging interactions to be 5 (corresponding to the 5 pounds knapsack capacity in (B)). This solution translates back to an optimal solution to the KOP instance in (B): include in the knapsack the laptop, lunch box, pen and water bottle, and exclude the notebook, textbook, and candle. Generally, any instance of KOP can easily (i.e. with polynomially bounded computational cost) be reduced into a corresponding NEP instance. A generalized formal KOP-to-NEP reduction is included in SI 4.

With small number of KOP items it is trivial to find out the optimal solution, but as the number of items increases the combinatorial search space grows exponentially fast. Whether there exists a polynomially-bounded algorithm for solving any arbitrary instances of an  $\mathcal{NP}$ -complete ( $\mathcal{NPC}$ ) problems is the subject of the  $\mathcal{P}$  vs.  $\mathcal{NP}$  question, arguably the most important questions in computer science and mathematics today (Aaronson 2004; 2005; Fortnow 2009).  $\mathcal{NPC}$  problems have defied all attempts aimed at finding algorithms that consume polynomially-bounded amount of computational resources for solving arbitrary instances. Hence, the increasingly accepted conjecture is that problems in this class will always require super-polynomial computational resources, and that this should be accepted as a universal law by the same token that repeatedly experimentally verified laws in physics are accepted as such Wigderson (2014). In prac-

tice heuristics-based algorithms do exist and can find good approximation solutions (e.g. by exploiting structures in practically common instances) (Vazirani 2013; Lawler 1979). In evolutionary context, however, the only ‘algorithm’ at hand is random variation non-random selection (RVnRS), and the NEP instance difficulty is measured relative to it.

### Optimization in biological context

Biological systems do not employ sophisticated search algorithms to determine the optimal conserve or mutate/delete actions from one generation to the next, but rather proceed through iterations of RVnRS (Carvunis et al. 2012). This work shows the link between network topology and the number of needed RVnRS iterations before the composition (nodes) and connectivity (edges) of a network has sufficiently been transformed away from a deleterious state. We highlight the exponential increase in the required RVnRS iterations were the topology of biological network to deviate away from the mLmH topology. Particularly, this work shows that the number of RVnRS iterations is exponential in the number of ambiguous genes (those having non-zero score for both benefit and damage). The more ambiguous a set of genes is, the *less likely* it is for the iterative RVnRS process to alter just the right set of genes such that the total number of beneficial/damaging interactions is optimally maximized/minimized in *as few* evolutionary iterations as possible.

The  $\mathcal{NP}$ -hardness of NEP in and of itself is a rather weak measure of computational difficulty if one has the freedom to choose the algorithm. Indeed, in practice, heuristics-based approximation algorithms can produce fairly satisfactory solutions that may not be too suboptimal relative to *the* optimal solution. We assume however that the only algorithm available to an evolving organism is RVnRS. The efficacy of RVnRS relies on accumulating improvements with as little ‘backtracking’ through the search space as possible. The more unambiguously beneficial or damaging genes there are in a system that is under some evolutionary pressure to alter its MIN, the more effective RVnRS is at accumulating more beneficial, and cleansing more damaging, interactions.

A naive greedy strategy could be to impose the OA by conserving every gene  $g_i$  where  $a_i = +1$ , and mutating/deleting every  $g_j$  where  $a_j = -1$ . However, **conserving  $g_i$**  can inadvertently conserve OA-contradicting interactions if  $g_j$  happens to be a promoter or inhibitor of some  $g_k$  where  $a_k = -1$  or  $a_k = +1$  respectively. And **deleting/mutating  $g_j$**  can inadvertently disrupt OA-supporting interactions if  $g_j$  happens to be a promoter or inhibitor of some  $g_k$  where  $a_k = +1$  or  $a_k = -1$  respectively. In other words, a naive imposition of an OA is complicated by the reality of network connectivity. NEP optimization in biological evolutionary context is summarized in Table SI 4. Further details on the notions of gene ambiguity, search space size under RVnRS, and the role of mLmH topology at reducing it is discussed in “Effective instance size” section and “Search space size” section.

### The semantics of NEP

The OAs described thus far represent rather blunt opinions about a gene by stating it should be promoted or inhibited. In reality a gene may ideally have both a promoter and an inhibitor co-expressed whereby, for example, the latter serves to tone down the potency of the former. A more fine-tuned OA would therefore be on individual interactions rather than genes. The only syntactic modifications to NEP definition in this case



would be that the OA is a matrix equal in dimensions to the interaction matrix  $M$ . This is in contrast to a gene-targeting OA which is represented as a  $|G|$ -long sequence where  $G$  is the set of genes in the network (as was the case in the example of Fig. 2a). The complexity and instance difficulty of NEP remains the same whether the Oracle is stating its advice on genes or interactions, however (see SI 5 for further details).

An OA can further be interpreted semantically as describing short- or long-term evolutionary pressure. Short-term OA is specific to some regulatory state at a specific moment of the cell's life cycle for example (Fig. 2c). An OA on some gene or interaction indicates an evolutionary pressure to fine-tune its regulation positively or negatively. Long-term OA on the other hand indicates a more existential evolutionary pressure for or against a gene or interaction. In this case, the OA is interpreted as advising on whether a gene or interaction is rather advantageous or disadvantage to the organism's survival generally (Fig. 2d).

## Results

Here we summarize the results presented in subsequent subsections. We empirically study NEP instances obtained by generating hypothetical OAs on various MINs and compare that with instances on synthetic networks of various topologies. We assess instance difficulty assuming the working algorithm is RVnRS. The results show instances obtained from real MINs are invariably far easier to satisfy by virtue of the mLmH property, as real MINs have far less ambiguous genes per instance. The large number of low-degree leaf genes in MINs reduces instance size because such genes are certain (degree 1) or likely (degree 2, 3, .. with exponentially decreasing likelihood) to have all-beneficial (all-damaging) interactions and should therefore be conserved (mutated/deleted) regardless. Such unambiguous genes need not be considered in the computationally costly optimization search. For example, a leaf gene involved in only one interaction (i.e. it has degree 1) can either be totally beneficial or totally damaging under a given evolutionary pressure scenario, and as such there is no ambiguity as to whether it should be conserved or mutated/deleted. A gene of degree 2 has a 50% chance of being unambiguous: the two interactions it is engaged in are either both beneficial or both damaging. There are four possibilities of the state of the two interactions: (0,0), (0,1), (1,0) and (1,1) where 1 or 0 denote an interaction is beneficial or damaging respectively. In general, a gene of degree  $d$  has a  $\frac{2}{2^d} = 2^{1-d}$  probability of being unambiguous.

Based on the fact that the larger a gene's degree is the exponentially more likely it is to be ambiguous, the model predicts the expected number of genes of degree  $d$  in real MINs. The prediction formula is parameterized with a value proportional to the edge:node ratio of the MIN. All existing MINs are currently partial to different degrees of completion (i.e. not all interactions of all genes have been mapped out). Based on the predictability of the degree distribution of MINs from a computational intractability perspective, and the proportionality of fitted parameters in the prediction formula to the edge:node and node:edge ratios in MINs, we conjecture that there will be a universal edge:node ratio of  $\sim 2$  in all MINs regardless of organism or physiological context. The validity of this conjecture can be assessed once all genes and interactions have been accounted for in a large number of MINs.

We also simulated the evolution of synthetic networks using an evolutionary algorithm which, in each generation, selects the top 10% of networks that present the easiest

optimization task (mainly by having a large number of unambiguous nodes). These top networks breed the next generation of networks. The degree distribution of the fittest synthetic networks after ~hundreds to few thousands of generations of simulated evolution very closely match those of real MINs of equal size. The emergence of mLmH in synthetic networks is not sensitive to the starting conditions: whether networks are initially empty (and accumulate nodes/edges over the generations) or randomized (have the same number of nodes/edges as a corresponding real MIN, but edges are re-assigned over the generations as a form of mutation).

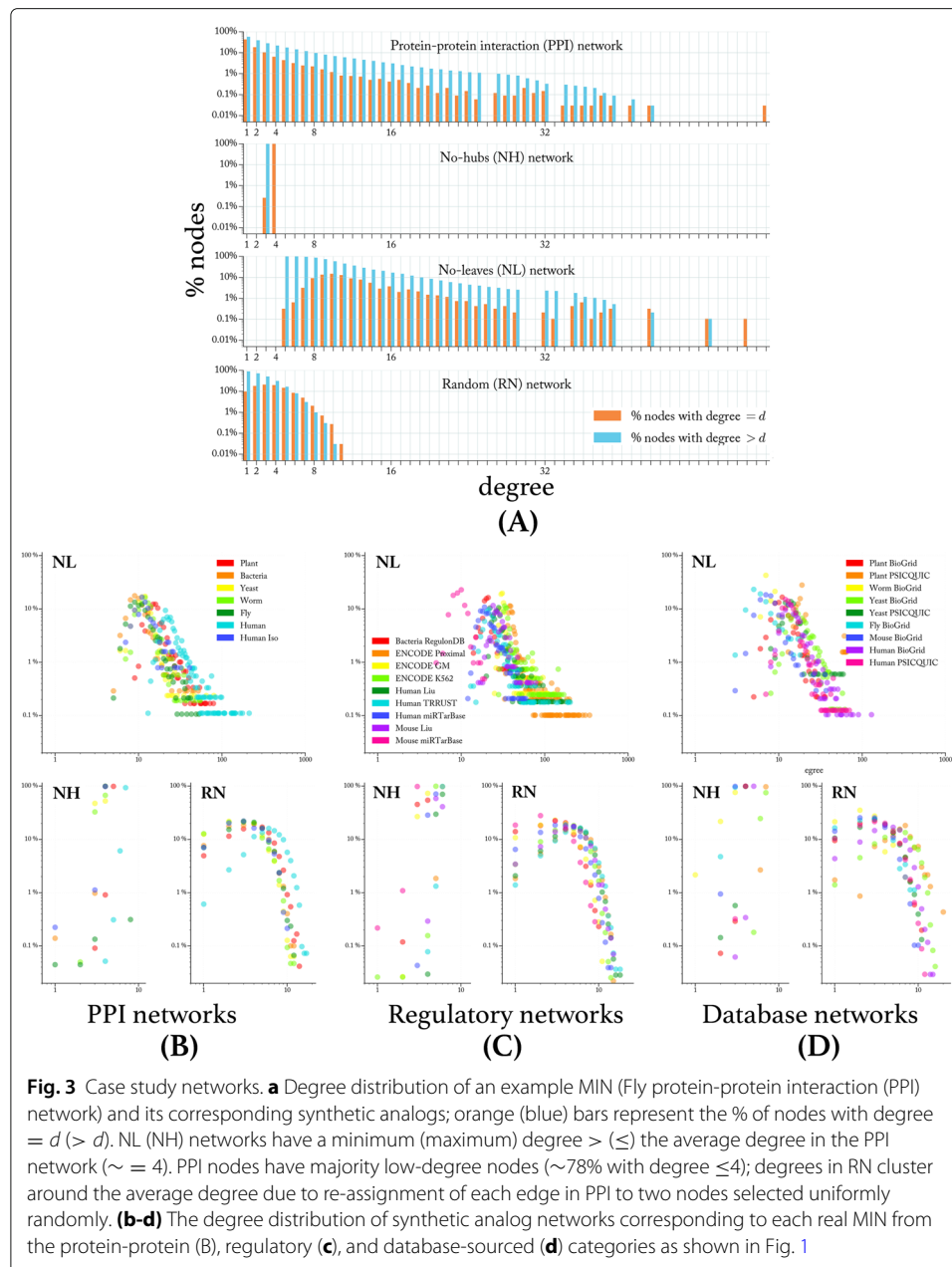
### Simulation of evolutionary pressure

NEP instances were generated on MINs (Fig. 1) as well as various synthetic networks. The direction and sign were assigned randomly in MINs that are undirected and/or unsigned (detailed in SI 2). For each MIN, we computer-generated synthetic analog networks having the same number of node and/or edges. Figure 3a shows the degree distribution of an example real network (*Drosophila melanogaster* (Fly) protein-protein interaction (PPI) network (Vinayagam et al. 2014)). No-leaves (NL) and no-hubs (NH) networks were generated by reassigning edges in PPI from leaves to nodes (for NL) or vice versa (for NH), so as to simulate the effect of depriving PPI of either property. Nodes in NL (NH) have a minimum (maximum) degree  $\geq$  ( $\leq$ ) the average node degree of PPI ( $\lceil \sim 3.6 \rceil = 4$ ). The redistribution of edges from leaves to hubs in NL results in some nodes having zero degree, which are eliminated, resulting in NL being a smaller (943 nodes), more dense network compared to PPI. NL and NH networks simulate two alternative topologies that biological networks could have evolved into if minimizing interactions per gene (NH) or total number of genes (NL) were the only driving forces in their evolution. We also applied the same simulation to a random (RN) analog of PPI network, whereby each edge in the latter is re-assigned to two randomly selected nodes (both edge direction and sign randomly assigned), and so nodes' degrees cluster around the average degree in PPI network. Figure 3b-d respectively show the degree distribution of analog (NL, NH, and RN) random networks for protein-protein, regulatory, and database-sourced networks (the degree distribution of corresponding real networks were shown in Fig. 1). 1-5K NEP instances are generated for each network by calculating benefit/damage values against a randomly generated OA on all nodes (for each node  $n_i, a_i \neq 0$ ). In Section 7.5 we show results where the OA is neutral to some nodes. Each instance is solved to optimality under a given tolerance threshold, expressed as the percentage of damaging edges to be tolerated. Details of the algorithmic workflow of the simulation, and the choice of sampling threshold, is provided in SI 6, 7, respectively.

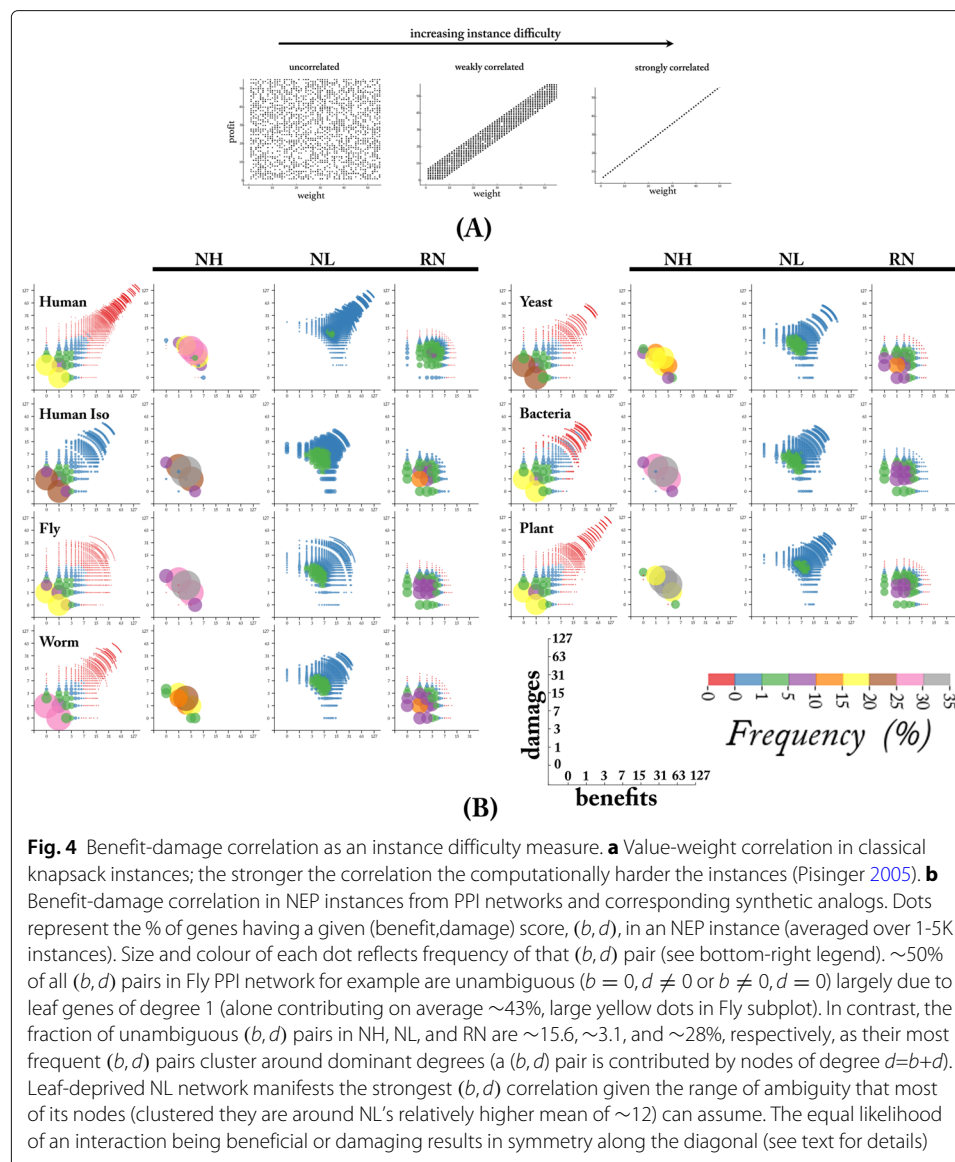
### Benefit-damage correlation

The correlation between benefit and damage scores is used to assess the difficulty of NEP instances in analogy to values-weights correlation in KOP (see NEP-to-KOP reverse-reduction in SI 4.3). Figure 4a shows the correlation plot of classical test instances (Pisinger 1999). It has previously been shown that the more correlated the values and weights in a knapsack instance are the more difficult the instance is (Pisinger 2005). Strong value-weight correlation increases the ambiguity as to which items to add/remove from the knapsack (or in NEP context, which genes to conserve/delete from the network). Figure 4b shows the average frequency of a (benefit, damage) pair ( $(b, d)$  hereafter)





over 1-5K NEP instances for PPI networks (for regulatory and DB-sourced networks see SI 8). The reduced ambiguity in instances from real PPI networks results by virtue of the large number of genes that are certain (degree 1) or likely (degree 2, 3, 4 .. with likelihoods 50, 12.5, 0.125 .. %, respectively) to be unambiguous: either totally advantageous ( $b \neq 0, d = 0$ ) or totally disadvantageous ( $b = 0, d \neq 0$ ). In Fly PPI network for example,  $\sim 50\%$  of all ( $b, d$ ) pairs are unambiguous, with 1:0 and 0:1 pairs (resulting from degree-1 leaf genes) alone representing  $\sim 43\%$  of those pairs (large brown dots). In contrast,  $\sim 15.6, \sim 3.1$ , and  $\sim 28\%$  of ( $b, d$ ) pairs in Fly network's NH, NL, and RN analogs are unambiguous. The role of leaves in decreasing ( $b, d$ ) correlation is especially

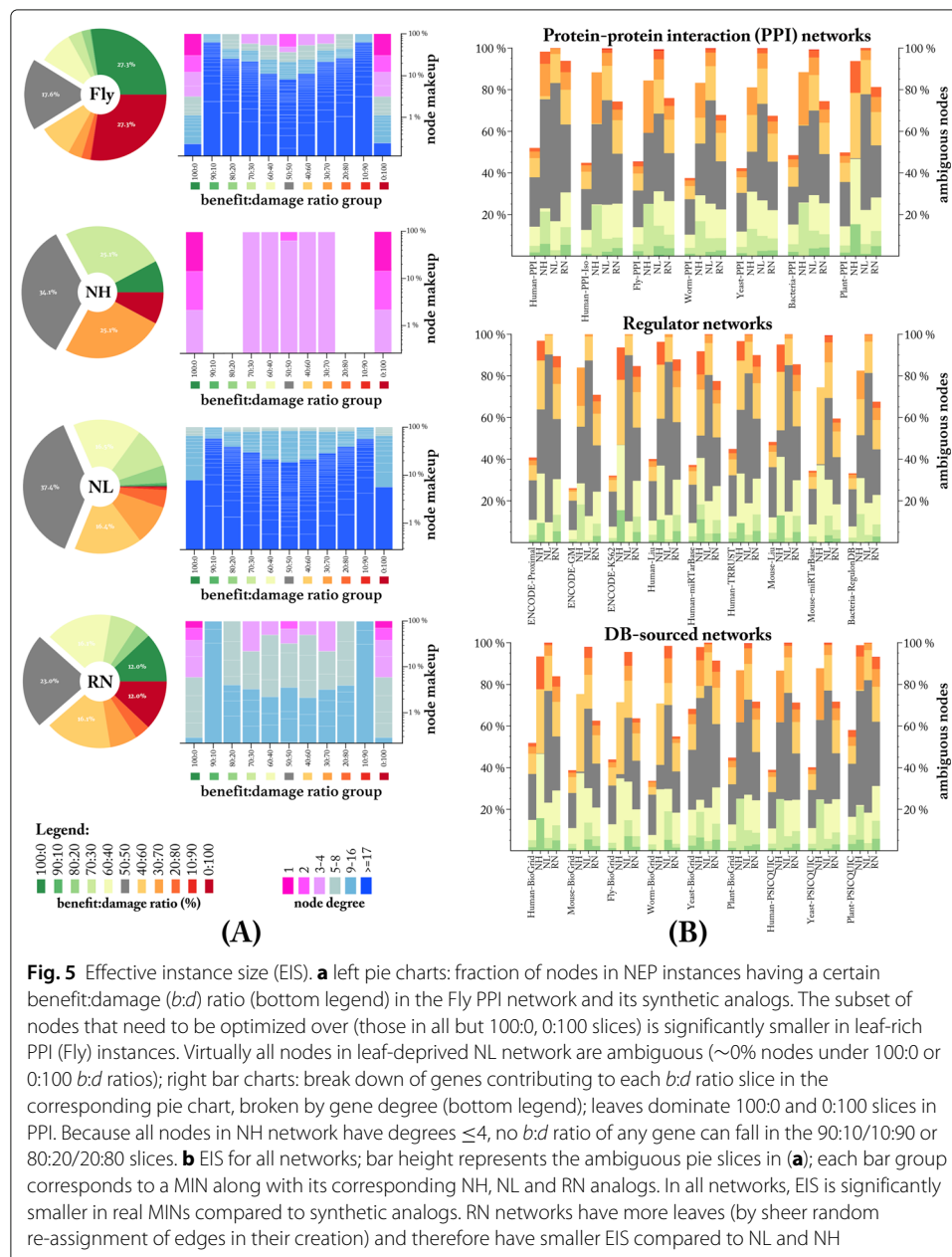


highlighted in contrast to leaf-deprived NL network which exhibits the strongest correlation around its higher mean degree (analysis of degree-to-ambiguity proportional relation is detailed in “[Prediction of degree distribution](#)” section). The symmetry along the diagonal in Fig. 4b is expected given that there is an equal probability of 50% that an interaction is beneficial or damaging under a random Oracle advice on the gene that interaction is targeting.

The signed Fly PPI network distinctly shows more sporadic benefit-damage correlation (red dots in Fly dot plot in Fig. 4) in higher-degree nodes, which results from the asymmetry of the number of promotional (67.5%) versus inhibitory (32.5%) interactions. Under random OA, such asymmetry results in higher likelihood of disparate benefit and damage values. The asymmetry of signs in Fly PPI network is consistent with a recent report that showed a similar promotional/inhibitory interaction distribution in yeast (Costanzo et al. 2016).

**Effective instance size**

Unambiguous genes can a priori be deemed advantageous or disadvantageous and therefore should be conserved or mutated/deleted, respectively. As such they need not be part of the optimization search because the optimal evolutionary outcome as to whether to conserve or alter them is independent of that of other genes in the network. Effective instance size (EIS) is the fraction of genes in an NEP instance that are ambiguous ( $b \neq 0$  and  $d \neq 0$ ). The smaller the  $|b - d|$  value the more ambiguous a gene is. Figure 5a, left (pie charts) shows the fraction of genes that on average (over 1-5K instances) falls under a certain  $b:d$  ratio slice where  $b:d = \frac{b}{b+d} : \frac{d}{b+d}$  for an example network (Fly PPI). NEP instances in PPI have ~42% EIS ( $b:d = 90:10, 80:20, \dots, 10:90\%$ ), compared to ~84, ~100, and ~72 % for NH, NL and RN networks respectively. Compared to NH and NL, EIS in RN is



**Fig. 5** Effective instance size (EIS). **a** left pie charts: fraction of nodes in NEP instances having a certain benefit:damage ( $b:d$ ) ratio (bottom legend) in the Fly PPI network and its synthetic analogs. The subset of nodes that need to be optimized over (those in all but 100:0, 0:100 slices) is significantly smaller in leaf-rich PPI (Fly) instances. Virtually all nodes in leaf-deprived NL network are ambiguous (~0% nodes under 100:0 or 0:100  $b:d$  ratios); right bar charts: break down of genes contributing to each  $b:d$  ratio slice in the corresponding pie chart, broken by gene degree (bottom legend); leaves dominate 100:0 and 0:100 slices in PPI. Because all nodes in NH networks have degrees  $\leq 4$ , no  $b:d$  ratio of any gene can fall in the 90:10/10:90 or 80:20/20:80 slices. **b** EIS for all networks; bar height represents the ambiguous pie slices in (a); each bar group corresponds to a MIN along with its corresponding NH, NL and RN analogs. In all networks, EIS is significantly smaller in real MINs compared to synthetic analogs. RN networks have more leaves (by sheer random re-assignment of edges in their creation) and therefore have smaller EIS compared to NL and NH

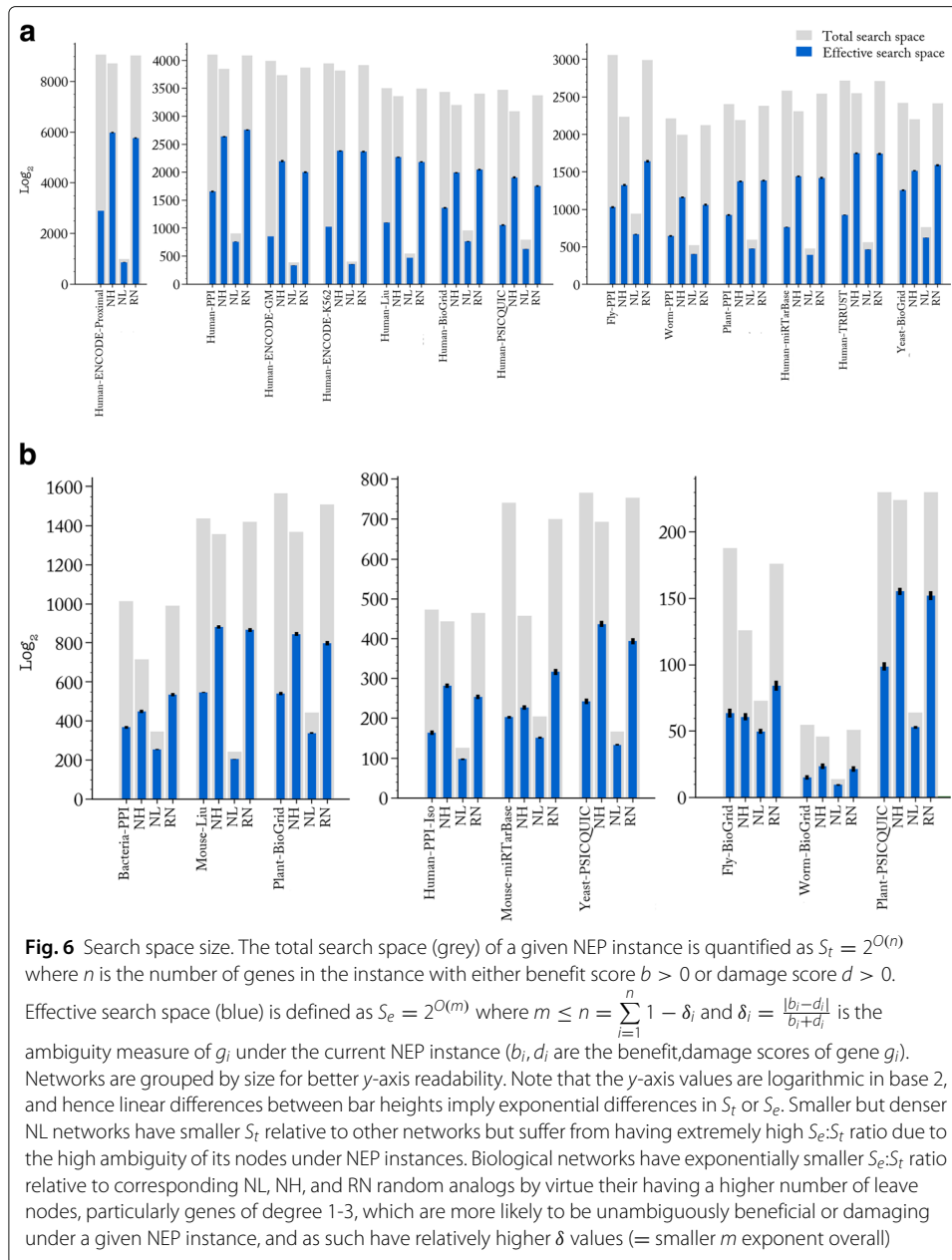
smaller to the extent that it has more leaf nodes (particularly degree 1-3) which relatively increases the size of its unambiguous slices ( $b:d$  100:0 or 0:100%).

The constituent genes in each pie slice are shown in Fig. 5a, right bar charts, for each network, broken down by degree range (bottom legend). Since the likelihood of a gene's ambiguity is inversely (and exponentially, see later discussion in the next section) proportional to its degree, leaf genes (degree  $\leq 4$ ) in PPI dominate the unambiguous 100:0 or 0:100%  $b:d$  ratio groups. With virtually all nodes in NH network having a degree 4 (NH bar chart in Fig. 3a), no  $b:d$  ratio of any node can fall in certain  $b:d$  ratio groups (more specifically, none of the possible  $(b, d)$  pairs 4:0, 3:1, 2:2, 1:3 ... 0:4 falls into any of the  $b:d$  ratios 90:10, 80:20, 10:90, or 20:80%). Figure 5b shows the EIS for all networks, where the bar height represents the size of ambiguous pie slices shown in (A) excluding the unambiguous 100:0 and 0:100 slices. Each bar group corresponds to a network, with the first bar being the real network and the 2nd, 3rd and 4th (left to right) corresponding to the network's NH, NL and RN analogs respectively. EIS is significantly smaller in real networks as compared to synthetic analogs regardless of physiological context or the source of the network. EIS in RN analogs is comparatively smaller in networks having smaller edge:node ( $e2n$ ) ratio since such networks are more likely to have leaf nodes after random re-assignment of edges (e.g. Human and Yeast PPI networks have 1.45 and 3.24  $e2n$  ratios respectively).

### Search space size

The total search space of a given NEP instance is defined as  $S_t = 2^{O(n)}$  where  $n$  is the number of ambiguous genes in the instance, i.e. those with  $b \neq 0$  and  $d \neq 0$ . The base 2 denotes the two general evolutionary outcome on an ambiguous gene  $g_i$ : its state will have been (1) altered or (2) unaltered after the next round of RVnRS. The alteration to a gene (for example through mutation) can be (dis-)advantageous relative to the current NEP instance. The state of an ambiguous gene can further be affected by the state changes of its interacting partners. For example, in the NEP instance of Fig. 2b, a mutation to gene  $g_2$  which causes it to lose its positive interactions with say,  $g_1$  and  $g_5$  would not only change its benefit/damage  $(b, d)$  scores from  $(4, 2)$  to  $(2, 2)$  but it would also result in the  $(b, d)$  scores of  $g_1$  and  $g_5$  to change even if they themselves did not undergo any alteration.  $g_1$  and  $g_5$  lose 1 benefit as a result of them no longer positively interacting with  $g_2$  due to the latter's aforementioned mutation.

Clearly the more connected a gene is the more its state alteration, and the alteration of its interacting partners, changes the optimal evolutionary target for such a gene from one generation to the next (i.e. whether it should ideally be conserved or mutated/deleted given the current evolutionary pressure described by the OA). Effective search space is defined as  $S_e = 2^{O(m)}$  where  $m \leq n = \sum_{i=1}^n 1 - \delta_i$  and  $\delta_i = \frac{|b_i - d_i|}{b_i + d_i}$  is the ambiguity measure of  $g_i$  under the current NEP instance. On the one extreme where  $\forall g_i, b_i = d_i \neq 0$ , then  $m = n$  and therefore  $S_e = S_t$ . On the other extreme where all genes are unambiguous,  $\forall g_i, b_i = 0 | d_i = 0$ , then  $S_e = 0$ . An unambiguously beneficial or damaging gene  $g_j$  ( $d_j = 0$  or  $b_j = 0$ , respectively) does not increase  $m$  regardless of the current evolutionary pressure scenario because  $\delta_j = 1$  (gene ambiguity vs. degree is discussed in details in "Prediction of degree distribution" section). Figure 6 shows  $S_t$  (grey) and  $S_e$  (blue) for all networks. Note that the  $y$ -axis values are logarithmic in base 2, and hence linear differences between bar heights imply exponential differences in  $S_t$  or  $S_e$ . The smaller

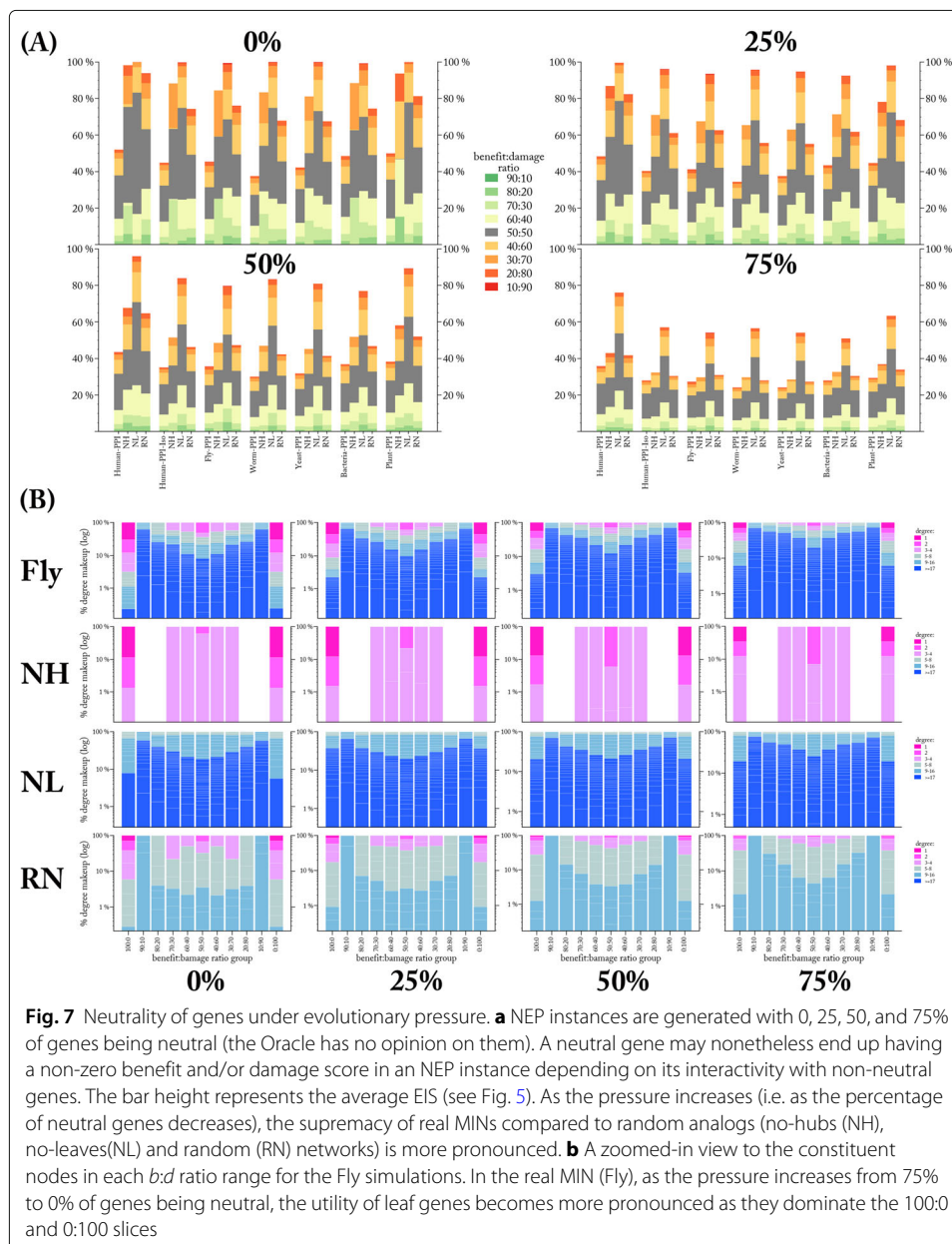


but denser NL networks show smaller  $S_t$  but suffer from exponentially higher  $S_e:S_t$  ratio compared to other networks. This implies that, while the search space is smaller, the optimal solution can change radically after each RVnRS given how deeply interconnected the genes are in such a network. Biological networks show exponentially smaller  $S_e:S_t$  ratio especially for larger more complete (e.g. ENCODE and Human PPI connectome) and/or curated (e.g. TRRUST) networks.

**Gene neutrality**

To test the effect of having a certain fraction of the genes as neutral under a given evolutionary pressure, NEP instances were generating assuming 0, 25, 50, and 75% of genes are

neutral (the Oracle has no opinion about them). A gene towards which the Oracle is indifferent may nonetheless end up having a non-zero benefit and/or damage value depending on its interaction profile with other genes on which the Oracle has an opinion. Subplots from top-left (0%) to bottom-right (75%) in Fig. 7a demonstrate the resulting effective instance size (EIS) as pressure decreases (i.e. as the number of neutral genes increases). The bar height represents the effective instance size (EIS), i.e. the fraction of genes that are ambiguous (in that they are engaged in both beneficial and damaging interactions in a given NEP instance). EIS shown here is the average over 1-5K instances. As the pressure increases (bottom to top), the supremacy of real MINs compared to synthetic analogs (no-hubs (NH), no-leaves(NL) and random (RN) networks) is more pronounced. Leaf-deprived NL synthetic analogs show the largest EIS across all settings, while NH and RN



**Fig. 7** Neutrality of genes under evolutionary pressure. **a** NEP instances are generated with 0, 25, 50, and 75% of genes being neutral (the Oracle has no opinion on them). A neutral gene may nonetheless end up having a non-zero benefit and/or damage score in an NEP instance depending on its interactivity with non-neutral genes. The bar height represents the average EIS (see Fig. 5). As the pressure increases (i.e. as the percentage of neutral genes decreases), the supremacy of real MINs compared to random analogs (no-hubs (NH), no-leaves(NL) and random (RN) networks) is more pronounced. **b** A zoomed-in view to the constituent nodes in each *b:d* ratio range for the Fly simulations. In the real MIN (Fly), as the pressure increases from 75% to 0% of genes being neutral, the utility of leaf genes becomes more pronounced as they dominate the 100:0 and 0:100 slices



networks show decreasing EIS to the extent that they both have more smaller-degree leaf nodes. The constituent nodes in each  $b:d$  bar segments is shown in Fig. 7b for the Fly network and its corresponding random analog. As the pressure decreases from 0% to 75%, the constituent nodes contributing to a given  $b:d$  ratio range remains largely unchanged as the nodes towards which the Oracle is not indifferent are selected randomly and as such the relative frequency of nodes of a certain degree remains of the same proportion under different neutrality assumptions.

**Prediction of degree distribution**

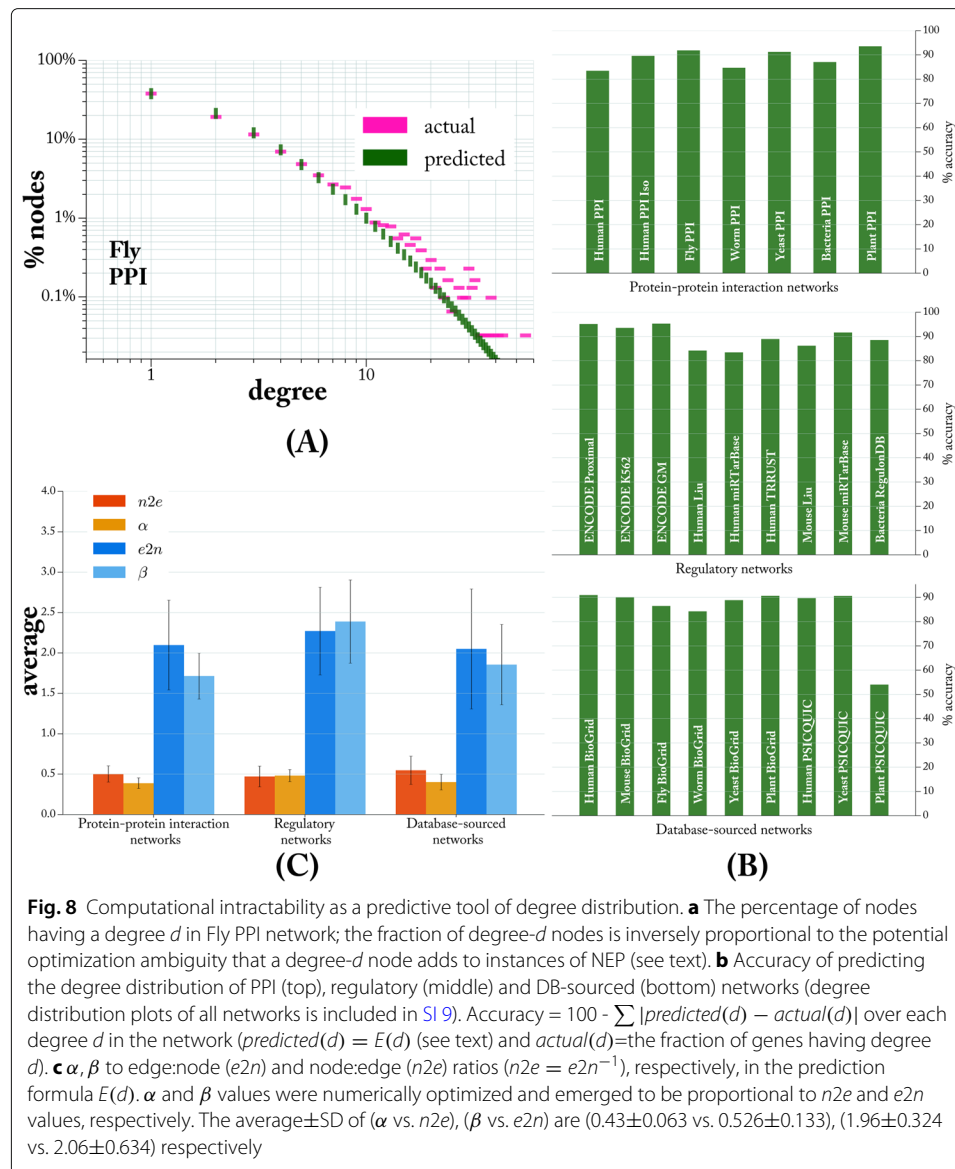
We considered whether computational intractability alone can predict the degree distribution of a biological network. More precisely, we considered whether the likelihood of a gene of degree  $d$  ("degree- $d$  gene" hereafter) to be totally advantageous or disadvantageous (belonging to green or red pie slices in Fig. 5a, respectively), which is exponentially inversely proportional to its degree, can predict the expected number of degree- $d$  genes in a biological network. For a degree-2 gene  $g_i$ , for example, there are  $2^2 = 4$  potential states of benefits/damages that  $g_i$  can assume under a given OA: 00, 01, 10, or 11 where 0 or 1 signify the edge (interaction) as being beneficial or damaging, respectively. States 01 or 10 are "ambiguous":  $g_i$  must be part of the overall optimization search to determine whether to conserve or delete it. In general, the number of ambiguous states for degree- $d$  gene is  $2^d - 2$ , albeit not all of equal ambiguity: while the 1000010 and 1111000 states of a degree-7 gene are both ambiguous, the former is significantly less so. Let  $k$  correspond to the number of 1's in a gene's given state (equivalently, its benefit score in a given NEP instance). We refer to a given state of a degree- $d$  gene as  $k$ -ambiguous ( $k$ -amb hereafter),  $0 \leq k \leq d$ , if it has  $k$  1's. For example, 0111 and 1000 are  $3 - amb$  and  $1 - amb$  states of a degree-4 gene. As  $k \rightarrow d$  (or  $k \rightarrow 0$ ) the ambiguity whether to conserve (or delete) decreases, while as  $k \rightarrow \frac{d}{2}$  both the ambiguity and (exponentially) number of states increases. For a degree-20 gene for example, there are  $\binom{20}{3} = 1140$   $3 - amb$  states compared to  $\binom{20}{10} = 184756$   $10 - amb$  states. Assuming an equal probability  $q = \frac{1}{2}$  for an edge to be beneficial or damaging, the likelihood of  $k - amb$  state for degree- $d$  gene is given by the expected number of  $k$  successes in  $d$  Bernoulli trials:

$$P(k - amb) = \binom{d}{k} q^k = \binom{d}{k} 2^{-d}$$

We define the expected frequency of degree- $d$  genes as:

$$E(d) = \frac{\alpha}{d + 1} \sum_{k=0}^d P(k - amb)^{\beta \log(d)}$$

where constants  $\alpha, \beta \in \mathbb{R}^+$  are proportional to node:edge ( $n2e$ ), edge:node ( $e2n$ ) ratios, respectively. Figure 8a shows the actual (pink) and predicted (green) degree frequency in Fly PPI network at  $\alpha, \beta = 0.43, 1.9$ , respectively. Prediction is further applied to all other networks with accuracy (defined as  $100 - \sum |predicted(d) - actual(d)|$ ) being  $\geq 84\%$  in 24 out of 25 networks as shown in Fig. 8b. Individual prediction plots for all networks is included in SI 9. An accurate account of all interactions currently is affected by the experimental bias against interactions involving lesser known genes (an inherent problem to small-scale studies (Rolland et al. 2014)). Figure 8c shows  $(\alpha, \beta)$  versus  $(n2e, e2n)$  ratios



of networks in (B) (all of which are currently partial, see SI 2). The optimal ( $\alpha$ ,  $\beta$ ) values were numerically determined by considering each  $\alpha$  in the interval [0.01, 1] in increments of 0.01 against each  $\beta$  in [0.1, 10] interval in increments of 0.1. The average of ( $\alpha$  vs.  $n2e$ ), ( $\beta$  vs.  $e2n$ ) are ( $0.43 \pm 0.063$  vs.  $0.526 \pm 0.133$ ), ( $1.96 \pm 0.324$  vs.  $2.06 \pm 0.634$ ) respectively. Note that the proportionality of optimal  $\alpha$  and  $\beta$  to  $n2e$  and  $e2n$  values is an emergent property.

An accurate account of all genes is not only limited by experimental coverage but also by alternatively-spliced isoforms of the same gene (which can have distinct interaction profiles (Yang et al. 2016)) being treated as a single gene. Hence, a gene's degree may be inflated in experiments where isoforms are not distinguished. For example, nodes in HumanIso network can be different isoforms of the same gene (Yang et al. 2016). The ( $\alpha$ ,  $\beta$ ) values versus ( $n2e$ ,  $e2n$ ) ratios in a partial MIN should indicate how well its coverage and resolution compares to other standard high-quality MINs. We conjecture that there

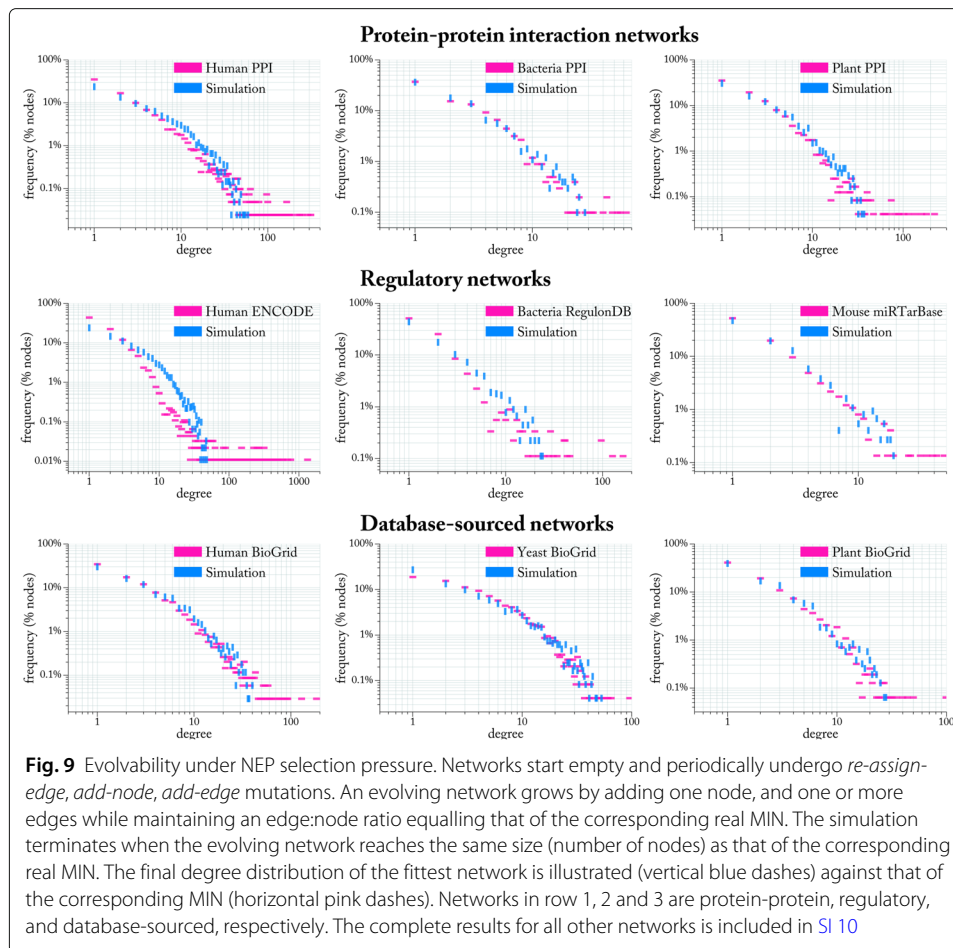
is ultimately (as experimental coverage and resolution of high-throughput interaction-detecting experiments increases) a universal  $e2n$  ratio of  $\sim 2$  in all MINs.

### Simulated evolution and adaptation

We conducted simulated network evolution and adaptation using an evolutionary algorithm that selects for networks that best minimize the difficulty of NEP instances generated against them. The simulation aims to assess the potency of NEP as a selection pressure that shapes network topology, rather than to mirror the details of recombination and mutation processes in biological systems. We conducted two sets of simulations depending on the starting conditions of synthetic networks. In the first, we begin with empty networks that grow over the generations by accumulating more nodes and edges. In the second, we begin with random synthetic networks that have the same size (number of nodes and edges) as corresponding real MINs. The size is kept unchanged from one generation to the next, but edges get re-assigned randomly in each generation as a form of mutation. We refer to the first and second sets of simulations as 'evolution' and 'adaptation' simulations, respectively. In both experiments, the idea is to examine the emerging degree distribution of synthetic networks after iterations of random mutation non-random selection, and compare that to the degree distribution of biological networks of equal size.

In simulated evolution, networks at the beginning of the simulation are near empty but are periodically assigned more new nodes and edges, in addition to being mutated by randomly re-assigning an existing edge to two randomly selected nodes. The simulation begins with a population of 60 near empty synthetic networks (8 nodes and  $8r$  randomly assigned edges, where  $r$  is the edge:node ( $e2n$ ) ratio to be maintained throughout the simulation). An NEP instance against each network in the population is obtained by generating a random interaction-targeting OA, resulting in some interactions in the network being beneficial and others damaging. Nodes' benefit/damage scores are calculated (as illustrated previously by example in Fig. 2, and further detailed in SI 5). Assuming a tolerance threshold of 5% of total damaging interactions network-wide, instances are solved to optimality (detailed further in SI 6).

The fitness of each synthetic network is based on two values averaged over the 60 instances: 1) effective instance size (EIS) as described in "Effective instance size" section and 2) the effective gained benefits (EGB) which measures the total number of beneficial interactions than can be obtained with the least number of nodes having to be conserved/deleted in an optimal NEP instances solution (see SI 10 for details). The top 6 fittest networks (10% of the population) that best minimize EIS and maximize EGB survive, while the remaining die. The surviving 6 networks are each mutated (*add-node*, *add-edge*, and *re-assign-edge* mutations) and 10 exact replicas are bred from each for the next round of mutate-and-select, resulting in a population of 60 networks once again. Another round starts by generating 100 NEP instances against each of the 60 networks in the population, and so on. The simulation is terminated when the number of nodes reaches that of a corresponding real MINs. Throughout the simulation, the *add-edge* mutation is conducted only if the edge-node ratio ( $e2n$ ) of the synthetic network does not exceed that of the corresponding real MIN. Figure 9 illustrates a sample of evolved synthetic networks and the corresponding MINs that have the same  $e2n$  ratio (results for all



other networks are shown in [SI 10](#)). The degree distributions of simulated networks (blue in [Fig. 9](#)) closely match their corresponding MINs (pink).

In simulated adaptations, the starting synthetic networks have the same size as the corresponding MIN, and no *add-node* or *add-edge* mutations are conducted (i.e. synthetic networks do not grow in size from one generation to the next). Only *re-assign-edge* mutation is conducted in each generation. All other aspects of the mutation/selection process is the same as the aforementioned simulated evolution experiments. mLmH property still emerges (see plots in [SI 11](#)). These results show the sufficiency of NEP-based evolutionary pressure to mold the network to mLmH topology within a number of generations that is extremely small relative to real evolutionary time. The number of generations is proportional to the number of genes in the target real MIN, i.e.  $\sim$ thousands (see [SI 2](#) for network sizes). These results also confirm findings from previous simulation and adaptation experiments on a smaller set of networks reported in [Atiia et al. \(2017\)](#).

The fitness function does not change throughout the simulation, while in reality there can be evolutionary periods where natural selection is acting more aggressively on genes of certain connectivity more than others. This could explain the discrepancy in hub concentration in simulated Human PPI, Plant PPI and Human ENCODE networks for example. It should be noted that hub concentration in real networks may, at least in part, be due to the low-resolution of network data which does not represent isoforms of the

same gene as separate nodes. This is supported by the fact that in Human PPI-Iso, which has high resolution (Yang et al. 2016), the discrepancy is very minimal (Figures SI 13).

### Concluding remarks

**contribution** The power of random variation and non-random selection to produce fine-tuned ‘hardware’ has been extensively documented at various biomechanical levels. Classic examples include allometric scaling and anatomical adaptations (organism), width and material of blood vessels (organ), compartmentalization of sub-cellular components (cellular) and fast-folding and aggressively-functioning enzymes (molecular). Recent reports (Livnat et al. 2011; Chastain et al. 2014) have highlighted evidence for ‘software’ optimization in biological systems from a computational complexity perspective, proposing justification for the evolutionary advantage for the role of sex for example (Livnat and Papadimitriou 2016). Such investigations into evolutionary biology through the lens of computational complexity have hitherto been too high-abstracted. Nonetheless, they do demonstrate that computational-complexity perspective has the potential to be the much-needed theoretical framework for turning massive volumes of biological datasets about biological systems into actionable knowledge (Brenner 2012).

There is currently a gap between the ever increasing scale and quality of molecular interaction networks (MINs) and the theoretical understanding of the origin of their architectural properties. It has long been debated whether properties such as the majority-leaves minority-hubs (mLmH) are adaptive. Here we showed how computational intractability can provide sufficient and necessary conditions for the emergence of mLmH as an adaptation to circumvent computational intractability. The model predicts and evolves the mLmH based on the fact as a gene’s degree increases linearly, its optimization “ambiguity” potential increases exponentially and hence the more computationally expensive is the task of adaptively re-wiring the network away from a deleterious and into an advantageous state. We conjectured based on predictability results that there is ultimately a universal edge:node ratio of  $\sim 2$  in all MINs. Given the increasing pace at which the coverage (Gerstein et al. 2012; Rolland et al. 2014) and resolution (Han et al. 2015; Yang et al. 2016) of high-throughput interaction-detecting experimental procedures is being reported, we expect the validity of such conjecture to be tested within a few years.

**sufficiency and necessity** We described the need for biological networks to change their node (gene) and edge (interaction) composition in response to some evolutionary pressure as a computational optimization problem which we showed to be  $\mathcal{NP}$ -hard. By itself, the  $\mathcal{NP}$ -hardness of a problem is a rather weak measure of computational difficulty. Instances dealt with in practice may possess exploitable properties and as such their optimal solutions could potentially be found without incurring exponential computational resource expenditure (Vazirani 2013). There could also exist good approximation algorithms that can solve to various degrees of optimality while consuming sub-exponential computational resources (Lawler 1979). However, in the context of biological evolution, the running algorithm is random-variation and non-random selection (RVnRS) and the instance difficulty of the network re-wiring problem must be assessed relative to it. Our results show that the search space size for the described network re-wiring problem would be exponentially orders of magnitude larger were the topology of biological networks to deviate significantly from mLmH.

While the model sufficiently predicts and generates the mLMH topology, it also provides a necessary condition for the emergence of mLMH as an adaptation around inescapable computational intractability considering that (1) the computational cost of the network rewiring problem is, in the general case, universally insurmountable (assuming  $\mathcal{P} \neq \mathcal{NP}$ ) and (2) RVnRS does not endow the organism with any heuristic-based approximation shortcuts that offer better than brute-force exploration of the search space. A sparse MIN where a gene has at most one interacting partner produces the easiest possible instances of the network-rewiring problem: regardless of the nature of the current evolutionary pressure, any gene can either be beneficial or damaging depending on whether the interaction it is engaged in is beneficial or damaging to the network as a whole. There is therefore no ambiguity as to which genes to conserve and which to delete/mutate in such a fully sparse network. However, such networks would necessarily contain more genes, since functions that could have otherwise been accomplished by a single hub gene (e.g. a phosphatase targeting multiple proteins) must now be handled by a large number of specialty genes. This clearly leads to an explosion of genome size. On the other extreme, a dense network where functions are concentrated in as few multi-tasking hub genes as possible would lead to an exponential search space. Particularly, the number of iterations of random-variation non-random selection needed before the network has been re-wired away from a deleterious and into an advantageous state would be exponential in the number of ambiguous genes given some evolutionary pressure. That all genes in such a network are engaged in more than one interaction exponentially increases their likelihood of being ambiguous under a given evolutionary pressure scenario. An exponential number of iterations of random-variation and non-random selections are needed before the right set of genes have been conserved, deleted or mutated such that the total number of damaging interactions network-wide is under a tolerable threshold.

mLMH topology is the middle ground between the two aforementioned extremes (the totally sparse but large, and the highly dense but unadaptable networks): essential functions are concentrated (Gerstein et al. 2012) in 'hub' genes that are unlikely to be damaging in and of themselves (Khurana et al. 2013) while continuous (and cheap) experimentation (e.g. fine-tuning micro-RNA regulation (Gerstein et al. 2012)) is conducted at the periphery of the network (Kim et al. 2007) with loosely connected leaf genes.

**applications and extensions** An immediate application of the model is to complement statistical tests used to infer the quality and coverage of large-scale interactome-mapping wet experiments (Rolland et al. 2014) or *in-silico* network inference (Mitra et al. 2013), by testing whether the resulting networks over- or under-represents real interactions relative to the prediction. There are aspects of the model that should be extended. In this work we treated all interactions as equal, but in reality some interactions are more potent than others. Future work could extend the model by considering the potency of each interaction, a not so trivial task since no large-scale data exist yet to facilitate its inference. The model is also static, in that assigning benefit/damage to a gene is based on immediate neighbours only. The implication is that all genes are equal, but in reality a central gene (many shortest paths pass through it) has much more effect network-wide than a gene residing at the periphery of the network. Trivially, implementing a dynamic variant (where the cascading effect of a gene's beneficial (damaging) effects are accounted for) does not change the complexity class of NEP, although its simulations will be more computationally demanding.



## Supplementary information

Supplementary information accompanies this paper at <https://doi.org/10.1007/s41109-020-00268-0>.

**Additional file 1:** Supplementary material Figures S1 to S17 and Tables S1 - S4.

### Abbreviations

NPC:  $\mathcal{NP}$ -complete; MIN: Molecular interaction network; mLmH: majority-leaves minority-Hubs network topology; OA: Oracle Advice; RVnRS: Random variation non-random selection; NEP: Network evolution problem; KOP: Knapsack optimization problem; PPI: Protein-protein interaction; NL: No-Leaf network; NH: No-Hub network; *amb*: ambiguous; EIS: Effective instance size; EGB: Effective gained benefits; PSICQUIC: Proteomics standard initiative common QUery InterfaCe; MIQL: Molecular interaction query language;  $n2e$ : node:edge ratio of a network;  $e2n$ : edge:node ratio of a network;

### Acknowledgements

Computations were made on the supercomputing cluster Guillimin from McGill University, managed by Calcul Québec and Compute Canada. The operation of these supercomputers is funded by CFI, MESI, and FRQ-NT.

### Authors' contributions

AAA, KI, SV and JW designed research; AAA performed research, implemented computer simulations, analyzed data, and wrote the paper; CH implemented simulated evolution experiments; JW, KI, SV, CH revised and corrected the paper. The author(s) read and approved the final manuscript.

### Funding

The authors declare no funding for this submission.

### Availability of data and materials

**Data availability** The data/reanalysis that support the findings of this study are publicly available online at: <https://github.com/aliatiia/evolution.by.computatioanl.selection/tree/master/data>.

**Code availability** Simulation and data processing source code are available at: <https://github.com/aliatiia/evolution.by.computatioanl.selection/tree/master/src>.

### Competing interests

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

### Author details

<sup>1</sup>School of Computer Science, McGill University, Montreal, Canada. <sup>2</sup>Research Centre on Complex Traits, McGill University, Montreal, Canada. <sup>3</sup>École normale supérieure Paris-Saclay, Cachan, France. <sup>4</sup>National Institute of Informatics, Tokyo, Japan. <sup>5</sup>Tokyo Institute of Technology, Tokyo, Japan.

Received: 20 November 2019 Accepted: 27 May 2020

Published online: 10 July 2020

### References

- Aaronson S (2004) Limits on efficient computation in the physical world. arXiv preprint quant-ph/0412143. <https://arxiv.org/abs/quant-ph/0412143>
- Aaronson S (2005) Guest column: NP-complete problems and physical reality. *ACM Sigact News* 36(1):30–52. <http://dl.acm.org/citation.cfm?id=1052804>. Accessed 1 Sept 2019
- Albert R, Jeong H, Barabasi AL (2000) Error and attack tolerance of complex networks : Article : Nature. *Nature* 406(6794):378–382. <http://www.nature.com/nature/journal/v406/n6794/full/406378a0.html>. Accessed 1 Sept 2019
- Alderson DL, Doyle JC (2010) Contrasting views of complexity and their implications for network-centric infrastructures. *IEEE Trans Syst Man Cybern Syst Hum* 40(4):839–852. [http://ieeexplore.ieee.org/xpls/abs\\_all.jsp?arnumber=5477188](http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5477188). Accessed 1 Sept 2019
- Arita M (2004) The metabolic world of Escherichia coli is not small. *Proc Nat Acad Sci U S A* 101(6):1543–1547. <http://www.pnas.org/content/101/6/1543.short>. Accessed 1 Sept 2019
- Atiia A (2017) Case-Study Biological Networks. <https://github.com/aliatiia/KenanDB>. Accessed 1 Sept 2019
- Atiia A, Hopper C, Waldispühl J (2017) Computational Intractability Generates the Topology of Biological Networks. In: Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics. ACM, pp 500–509. <http://dl.acm.org/citation.cfm?id=3107453>. Accessed 1 Sept 2019
- Barabási AL, Albert R (1999) Emergence of Scaling in Random Networks. *Science* 286(5439):509–512. <http://www.sciencemag.org/content/286/5439/509>. Accessed 1 Sept 2019
- Barabási AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. *Nat Rev Genet* 5(2):101–113. <http://www.nature.com/nrg/journal/v5/n2/full/nrg1272.html>. Accessed 1 Sept 2019
- Brenner S (2012) Turing centenary: Life's code script. *Nature* 482(7386):461–461. <http://www.nature.com/nature/journal/v482/n7386/full/482461a.html>. Accessed 1 Sept 2019
- Carvunis AR, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, et al. (2012) Proto-genes and de novo gene birth. *Nature* 487(7407):370–374. <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3401362/>. Accessed 1 Sept 2019
- Chastain E, Livnat A, Papadimitriou C, Vazirani U (2014) Algorithms, games, and evolution. *Proc Nat Acad Sci* 111(29):10620–10623. <http://www.pnas.org/content/111/29/10620>. Accessed 1 Sept 2019

- Costanzo M, VanderSluis B, Koch EN, Baryshnikova A, Pons C, Tan G, et al. (2016) A global genetic interaction network maps a wiring diagram of cellular function. *Science* 353(6306):aaf1420. <http://science.sciencemag.org/content/353/6306/aaf1420>. Accessed 1 Sept 2019
- Fell DA, Wagner A (2000) The small world of metabolism. *Nature Biotechnology* 18(11):1121–1122. [http://www.nature.com/nbt/journal/v18/n11/full/nbt1100\\_1121.html](http://www.nature.com/nbt/journal/v18/n11/full/nbt1100_1121.html). Accessed 1 Sept 2019
- Fortnow L (2009) The status of the P versus NP problem. *Commun ACM* 52(9):78–86. <http://dl.acm.org/citation.cfm?id=1562186>. Accessed 1 Sept 2019
- Fox Keller E (2005) Revisiting "scale-free" networks. *BioEssays* 27(10):1060–1068. <http://onlinelibrary.wiley.com/doi/10.1002/bies.20294/full>. Accessed 1 Sept 2019
- Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan KK, Cheng C, et al. (2012) Architecture of the human regulatory network derived from ENCODE data. *Nature* 489(7414):91–100. <https://www.nature.com/nature/journal/v489/n7414/full/nature11245.html>. Accessed 1 Sept 2019
- Hahn MW, Conant GC, Wagner A (2004) Molecular Evolution in Large Genetic Networks: Does Connectivity Equal Constraint?. *J Mol Evol* 58(2):203–211. <https://link.springer.com/article/10.1007/s00239-003-2544-0>. Accessed 1 Sept 2019
- Han H, Shim H, Shin D, Shim JE, Ko Y, Shin J, et al. (2015) TRRUST: a reference database of human transcriptional regulatory interactions. *Sci Rep* 5. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4464350/>. Accessed 1 Sept 2019
- Khanin R, Wit E (2006) How scale-free are biological networks. *J Comput Biol* 13(3):810–818. <http://online.liebertpub.com/doi/abs/10.1089/cmb.2006.13.810>. Accessed 1 Sept 2019
- Khurana E, Fu Y, Chen J, Gerstein M (2013) Interpretation of Genomic Variants Using a Unified Biological Network Approach. *PLoS Comput Biol* 9(3):e1002886. <http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002886>. Accessed 1 Sept 2019
- Kim PM, Korbel JO, Gerstein MB (2007) Positive selection at the protein network periphery: Evaluation in terms of structural constraints and cellular context. *Proc Natl Acad Sci* 104(51):20274–20279. <http://www.pnas.org/content/104/51/20274>. Accessed 1 Sept 2019
- Lawler EL (1979) Fast approximation algorithms for knapsack problems. *Math Oper Res* 4(4):339–356. Bibtex: lawler\_fast\_1979-1. <http://pubsonline.informs.org/doi/abs/10.1287/moor.4.4.339>. Accessed 1 Sept 2019
- Livnat A, Papadimitriou C, Feldman MW (2011) An analytical contrast between fitness maximization and selection for mixability. *J Theor Biol* 273(1):232–234. <http://www.sciencedirect.com/science/article/pii/S002251931000634X>. Accessed 1 Sept 2019
- Livnat A, Papadimitriou C (2016) Sex as an algorithm: the theory of evolution under the lens of computation. *Commun ACM* 59(11):84–93. <http://dl.acm.org/citation.cfm?id=2934662>. Accessed 1 Sept 2019
- Lynch M (2007) The evolution of genetic networks by non-adaptive processes. *Nat Rev Genet* 8(10):803–813. <http://www.nature.com/nrg/journal/v8/n10/full/nrg2192.html>. Accessed 1 Sept 2019
- Mitra K, Carvunis AR, Ramesh SK, Ideker T (2013) Integrative approaches for finding modular structure in biological networks. *Nat Rev Genet* 14(10):719–732. <https://www.nature.com/articles/nrg3552>. Accessed 1 Sept 2019
- Papp B, Teusink B, Notebaart RA (2009) A critical view of metabolic network adaptations. *HFSP J* 3(1):24–35. <http://www.tandfonline.com/doi/abs/10.2976/1.3020599>. Accessed 1 Sept 2019
- Pisinger D (1999) Core problems in knapsack algorithms. *Oper Res* 47(4):570–575. <http://pubsonline.informs.org/doi/abs/10.1287/opre.47.4.570>. Accessed 1 Sept 2019
- Pisinger D (2005) Where are the hard knapsack problems?. *Comput Oper Res* 32(9):2271–2284. <http://www.sciencedirect.com/science/article/pii/S030505480400036X>. Accessed 1 Sept 2019
- Rolland T, Tazan M, Charlotiaux B, Pevzner SJ, Zhong Q, Sahni N, et al. (2014) A proteome-scale map of the human interactome network. *Cell* 159(5):1212–1226. <http://www.sciencedirect.com/science/article/pii/S0092867414014226>. Accessed 1 Sept 2019
- Sorrells T, Johnson A (2015) Making Sense of Transcription Networks. *Cell* 161(4):714–723. <http://www.sciencedirect.com/science/article/pii/S0092867415004316>. Accessed 1 Sept 2019
- Stelling J, Sauer U, Szallasi Z, Doyle FJ, Doyle J (2004) Robustness of cellular functions. *Cell* 118(6):675–685. <http://www.sciencedirect.com/science/article/pii/S0092867404008402>. Accessed 1 Sept 2019
- Stumpf MPH, Porter MA (2012) Critical Truths About Power Laws. *Science* 335(6069):665–666. <http://science.sciencemag.org/content/335/6069/665>. Accessed 1 Sept 2019
- Tanaka R, Yi TM, Doyle J (2005) Some protein interaction data do not exhibit power law statistics. *FEBS Lett* 579(23):5140–5144. <http://onlinelibrary.wiley.com/doi/10.1016/j.febslet.2005.08.024/full>. Accessed 1 Sept 2019
- Vazirani VV (2013) Approximation algorithms. Springer Science & Business Media. [https://scholar.google.ca/scholar?hl=en&as\\_sdt=0%2C5&q=Vazirani%20V%20\(2013\)%20Approximation%20algorithms](https://scholar.google.ca/scholar?hl=en&as_sdt=0%2C5&q=Vazirani%20V%20(2013)%20Approximation%20algorithms)
- Vinayagam A, Zirin J, Roesel C, Hu Y, Yilmazel B, Samsonova AA, et al. (2014) Integrating protein-protein interaction networks with phenotypes reveals signs of interactions. *Nat Methods* 11(1):94–99. <http://www.nature.com/nmeth/journal/v11/n1/abs/nmeth.2733.html>. Accessed 1 Sept 2019
- Wigderson A (2014) Opening Remarks and Introduction of Biology Session. IAS Princeton. <https://video.ias.edu/computationconference/2014/1122-LeslieValiant>. Accessed 1 Sept 2019
- Yang X, Coulombe-Huntington J, Kang S, Sheynkman G, Hao T, Richardson A, et al. (2016) Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing. *Cell* 164(4):805–817. <http://www.sciencedirect.com/science/article/pii/S0092867416300435>. Accessed 1 Sept 2019

## Publisher's Note

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