Supplementary Material

Adding levels of complexity enhances robustness and evolvability in a multi-level genotype-phenotype map

Pablo Catalán, Andreas Wagner, Susanna Manrubia and José A. Cuesta

Journal of the Royal Society Interface

Contents

1 toyLIFE

toyLIFE was originally presented in [1]. We give here its main details, with slight modifications in the definition of the model.

1.1 Building blocks: genes, proteins, metabolites

The basic building blocks of toyLIFE are toyNucleotides (toyN), toyAminoacids (toyA), and toySugars (toyS). Each block comes in two flavors: hydrophobic (H) or polar (P). Random polymers of basic blocks constitute toyGenes (formed by 20 toyN units), toyProteins (chains of 16 toyA units), and toyMetabolites (sequences of toyS units of arbitrary length). These elements of toyLIFE are defined on two-dimensional space (Supplementary Figure 1).

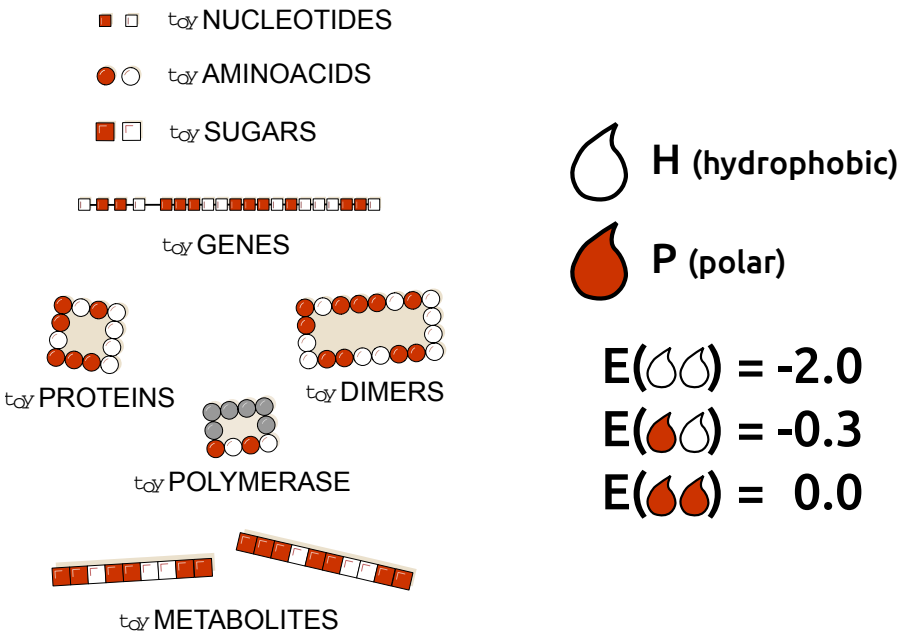
toyGenes

toyGenes are composed of a 4-toyN promoter region followed by a 16-toyN coding region. There are 2^4 different promoters and 2^{16} coding regions, leading to $2^{20} \approx 10^6$ toyGenes. An ensemble of toyGenes forms a genotype. If the toyGene is expressed, it will produce a chain of 16 toyA that represents a toyProtein. Translation follows a straightforward rule: H (P) toyN translate into H (P) toyA. Point mutations in toyLIFE are easy to implement: they are changes in one of the nucleotides in one of the genes in the genotype. If the sequence has a H toyN in that position, then a mutation will change it to a P toyN, and vice versa.

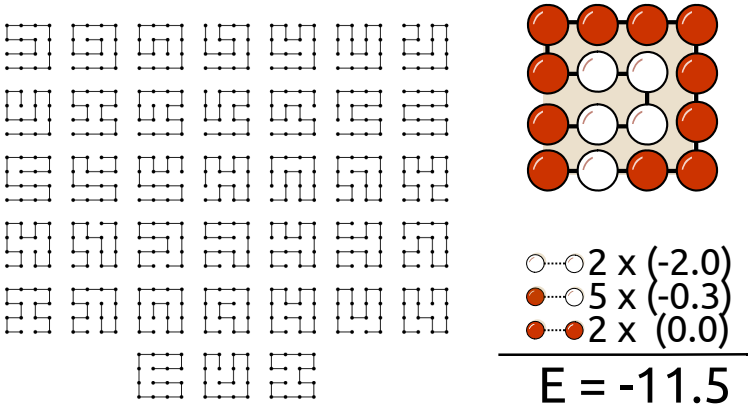
toyProteins

toyProteins correspond to the minimum energy, maximally compact folded structure of the 16 toyA chain arising from a translated toyGene. Their folded configuration is calculated through the hydrophobic-polar (HP) protein lattice model [2, 3].

We only consider maximally compact structures. That is, every toyProtein must fold on a 4×4 lattice, following a self-avoiding walk (SAW) on it. After accounting for symmetries —rotations and reflections—, there are only 38



Supplementary Figure 1: Building blocks and interactions defining toyLIFE. The three basic building blocks of toyLIFE are toyNucleotides, toyAminoacids, and toySugars. They can be hydrophobic (H, white) or polar (P, red), and their random polymers constitute toyGenes, toyProteins, and toyMetabolites. The toyPolymerase is a special polymer that will have specific regulatory functions. These polymers will interact between each other following an extension of the HP model (see text), for which we have chosen the interaction energies $E_{\mathrm{HH}} = -2$, $E_{\mathrm{HP}} = -0.3$ and $E_{\mathrm{PP}} = 0$ [2].



Supplementary Figure 2: Protein folding in ToyLIFE. toyProteins fold on a 4×4 lattice, following a self-avoiding walk (SAW). Discarding for symmetries, there are 38 SAWs (left). For each binary sequence of length 16, we fold it into every SAW and compute its folding energy, following the HP model. For instance, we fold the sequence PHPPPPPPPPPHHHHP into one of the SAWs and compute its folding energy (right). There are two HH contacts, five HP contacts and two PP contacts —we only take into account contacts between non-adjacent toyAminoacids. Summing all this contacts with their corresponding energies, we obtain a folding energy of -11.5. Repeating this process for every SAW, we obtain the minimum free structure.

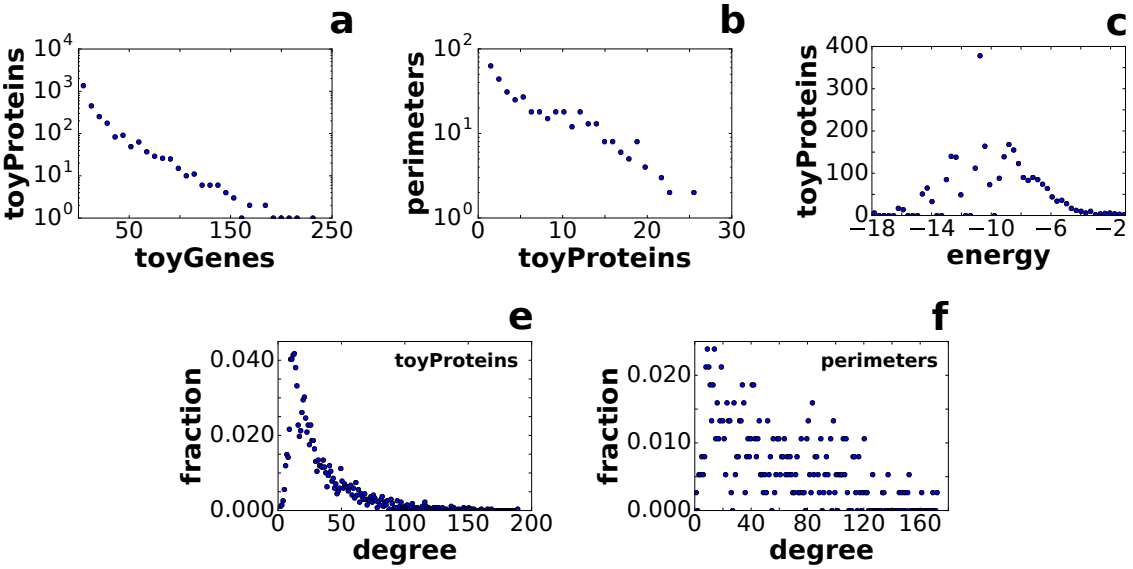
SAWs on that lattice (Supplementary Figure 2).

The energy of a fold is the sum of all pairwise interaction energies between toyA that are not contiguous along the sequence. Pairwise interaction energies are $E_{\mathrm{HH}} = -2$, $E_{\mathrm{HP}} = -0.3$ and $E_{\mathrm{PP}} = 0$, following the conditions set in [2] that $E_{\mathrm{PP}} > E_{\mathrm{HP}} > E_{\mathrm{HH}}$ (Supplementary Figure 2). toyProteins are identified by their folding energy and their perimeter. If there is more than one fold with the same minimum energy, we select the one with fewer H toyAminoacids in the perimeter. If still there is more than one fold fulfilling both conditions, we discard that protein by assuming that it is intrinsically disordered and thus non-functional. Note, however, that sometimes different folds yield the same folding energy and the same perimeter. In those cases, we do not discard the resulting toyProtein [1]. Out of $2^{16} = 65,536$ possible toyProteins, $12,987$ do not yield unique folds. We find $2,710$ different toyProteins with 379 different perimeters. Not all toyProteins are equally abundant: although every toyProtein is coded by 19.4 toyGenes on average, most of them are coded by only a few toyGenes. For instance, $1,364$ toyProteins —roughly half of them!— are coded by less than 10 toyGenes each. On the other hand, only 4 toyProteins are coded by more than 200 toyGenes each, the maximum being 235 toyGenes coding for the same toyProtein. The distribution is close to an exponential decay (Supplementary Figure 3**a**). The same happens with the perimeters, although with less skewness: each perimeter is mapped by 7.15 toyProteins on average, but the most abundant perimeters correspond to 26 toyProteins, and 100 are mapped by 1 or 2 toyProteins each (Supplementary Figure 3**b**). As we will see later, this already induces a certain degree of neutrality in ToyLIFE phenotypes.

Folding energies range from -18.0 to -0.6, with an average in -9.63. The distribution is unimodal, although very rugged (Supplementary Figure 3**c**). Note that folding energies are discrete, and that separations between them are not equal. For instance, there are 6 toyProteins that have a folding energy of -18.0, but the next energy level is -16.3, realised by 17 toyProteins, and yet the next level is -16.0, realised by 14 toyProteins. The mode of the distribution is -10.6, realised by 202 toyProteins.

We can also study the structure of the toyProtein network (Supplementary Figure 3**e**, **f**). The nodes of this network will be the $2,710$ toyProteins. toyProtein 1 and toyProtein 2 will be neighbors if there is a pair of toyGenes that express each toyProtein and whose sequence is equal but for one toyN. The weight of the edge between toyProtein1 and 2 will be the sum of such pairs of toyGenes. It is surprising that there are no self-loops in this network —there are no mutations connecting one toyProtein to itself. In other words, although there is a strong degeneracy in the mapping from toyGenes to toyProteins, there are no connected neutral networks. If we consider just the perimeters, however, the neutrality is somewhat recovered: out of the 379 perimeters, 224 of them have neutral neighbors. So there are many mutations that alter the folding energy of a toyProtein without changing the perimeter. In this sense,

[1] In [1], where we first presented ToyLIFE, we did not use this rule: whenever a sequence folded into two folds with the same folding energy and same number of Hs in the perimeter, we would discard them. This version of ToyLIFE, therefore, is slightly different. However, the results are qualitatively similar.



Supplementary Figure 3: Distributions of toyProteins in ToyLIFE. (a) Distribution of toyProtein abundances —that is, the number of toyGenes that code for them. Most toyProteins are coded by few toyGenes, but some of them are very abundant: the most abundant toyProtein is coded by 235 toyGenes. **(b)** Distribution of the perimeters associated with each toyProtein. Again, not all perimeters are equally abundant, and some of them correspond to as many as 25 toyProteins, while 100 correspond to 1 or 2 toyProteins. **(c)** Distribution of folding energies. The range of folding energies goes from -18.0 to -0.6, with a unimodal, rugged distribution. The mode is -10.6, a folding energy achieved by 202 toyProteins. **(d)** Degree distribution in the toyProtein network. Two toyProteins are connected if there are two toyGenes coding for them that have the same sequence, except for one toyN. The average degree is 32.2. **(e)** Degree distribution in the perimeter network. Two perimeters are neighbors if the toyProteins associated to them are neighbors. The average degree is 53.3.

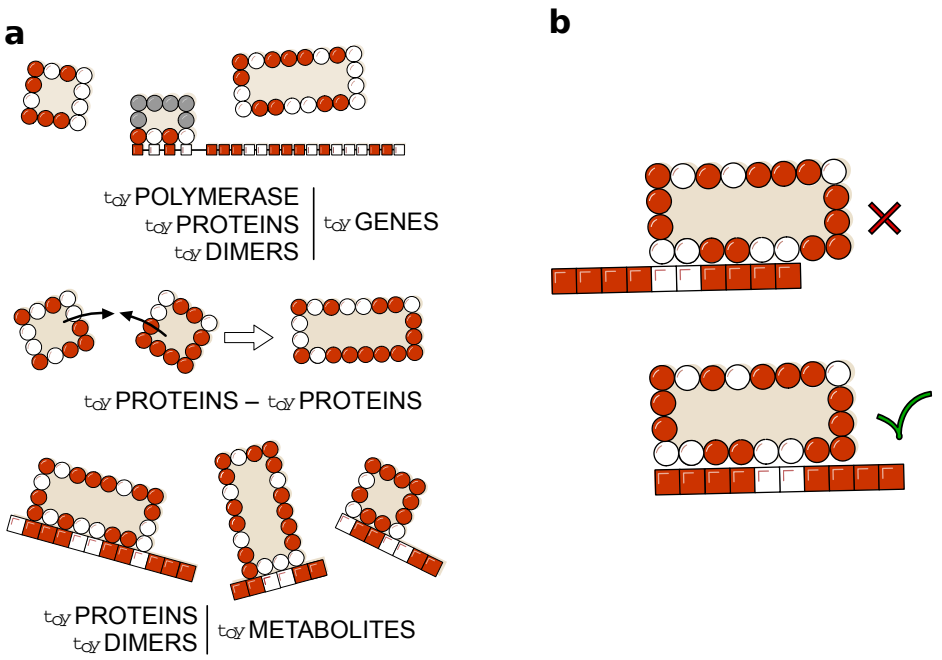
ToyLIFE is capturing a complex detail of molecular biology: mutations appear to be neutral from one point of view —in this case, perimeter— but are rarely entirely neutral. In other words, the value of a mutation is context and environment-dependent. There are always some small changes in the molecule —in this case, folding energy— that may affect their function later down the line. Real world examples of this *cryptic* effects of mutations on molecules are everywhere [4–7]. Connections between toyProteins are scarce too: the average degree in the toyProtein network is 32.2 (with a standard deviation of 25.7), a very small number — on average, each toyProtein is connected to hardly 1% of the rest of toyProteins! (Supplementary Figure 3**e**). The maximum degree is 190. This means that mutating from one toyProtein to another is not easy in general. In terms of perimeters this is more relaxed, as the average degree in the perimeter network is 53.3 (standard deviation is 38.1), with a maximum degree of 173. On average, every perimeter is connected to 14% of the rest of perimeters: it is a small number, but it is still higher than in the toyProtein case (Supplementary Figure 3**f**).

In the ToyLIFE universe, only the folding energy and perimeter of a toyProtein matter to characterise its interactions, so folded chains sharing these two features are indistinguishable. This is a difference with respect to the original HP model, where different inner cores defined different proteins and the composition of the perimeter was not considered as a phenotypic feature. However, subsequent versions of HP had already included additional traits [8].

The toyPolymerase (Supplementary Figure 1) is a special toyA polymer, similar to a toyProtein in many aspects, but that is not coded for by any toyGene. It has only one side, with sequence PHPH, and its folding energy is taken to be -11.0. We will discuss its function and place later on.

1.2 Extending the HP model: interactions

toyProteins interact through any of their sides with other toyProteins, with promoters of toyGenes, and with toyMetabolites (see Supplementary Figure 4**a**). When toyProteins bind to each other, they form a toyDimer, which is the only protein aggregate considered in ToyLIFE. The two toyProteins disappear, leaving only the toyDimer. Once formed, toyDimers can also bind to promoters or toyMetabolites through any of their sides —binding to other toyProteins or toyDimers, however, is not permitted. In all cases, the interaction energy (E_{int}) is the sum of pairwise interactions for



Supplementary Figure 4: Interactions in toyLIFE. (a) Possible interactions between pairs of toyLIFE elements. toyGenes interact through their promoter region with toyProteins (including the toyPolymerase and toyDimers); toyProteins can bind to form toyDimers, and interact with the toyPolymerase when bound to a promoter; both toyProteins and toyDimers can bind a toyMetabolite at arbitrary regions along its sequence. **(b)** When a toyDimer or toyProtein binds to a toyMetabolite with the same energy in many places, we choose the most centered binding position. If two or more binding positions have the same energy and are equally centered, then no binding occurs.

all HH, HP and PP pairs formed in the contact —these interactions follow the rules of the HP model as well. Bonds can be created only if the interaction energy between the two molecules E_{int} is lower than a threshold energy $E_{\text{thr}} = -2.6$. Note that a minimum binding energy threshold is necessary to avoid the systematic interaction of any two molecules. Low values of the threshold would lead to many possible interactions, which would increase computation times. High values would lead to very few interactions, and we would obtain a very dull model. Our choice of $E_{\text{thr}} = -2.6$ achieves a balance: the number of interactions is large enough to generate complex behaviours, as we will see later on, while at the same time keeping the universe of interactions small enough to handle computationally. If below threshold, the total energy of the resulting complex is the sum of E_{int} plus the folding energy of all toyProteins involved. The lower the total energy, the more stable the complex. When several toyProteins or toyDimers can bind to the same molecule, only the most stable complex is formed. Consistently with the assumptions for protein folding, when this rule does not determine univocally the result, no binding is produced.

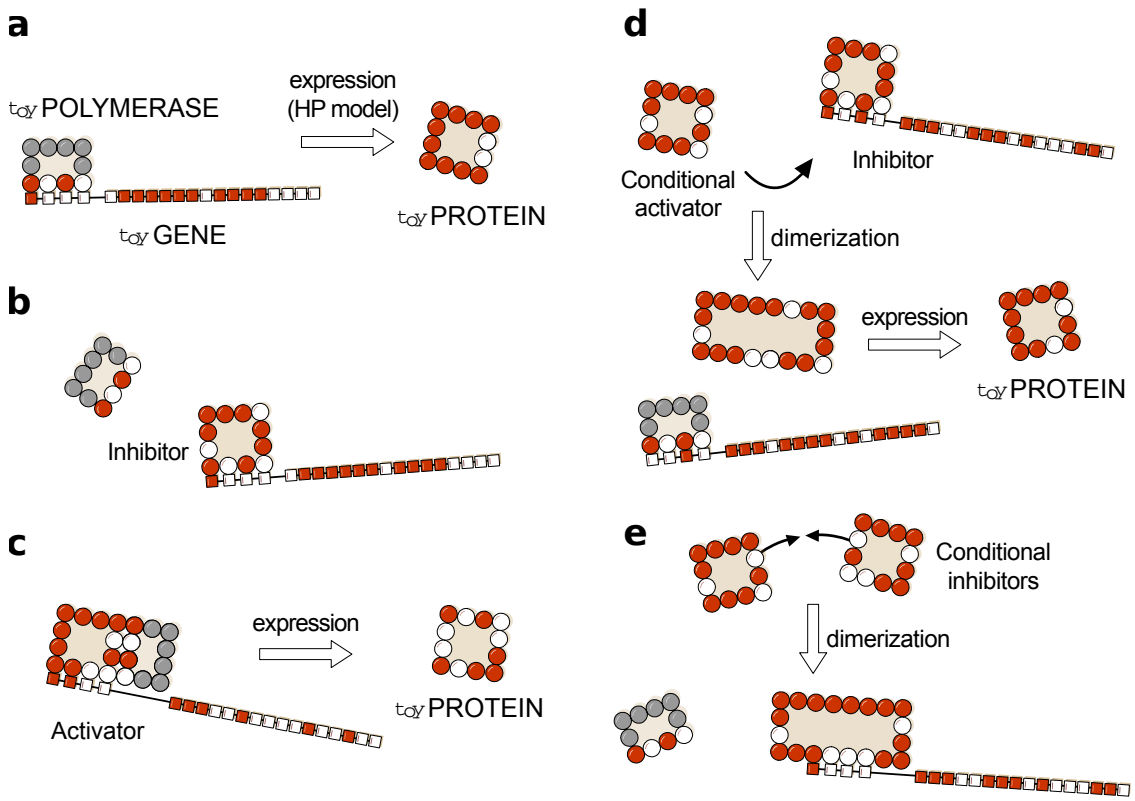
As the length of toyMetabolites is usually longer than 4 toyS (the length of interacting toyProtein sites), several binding positions between a toyMetabolite and a toyProtein might share the same energy. In those cases we select the sites that yield the most centered interaction (Supplementary Figure 4**b**). If ambiguity persists, no bond is formed. Also, no more than one toyProtein / toyDimer is allowed to bind to the same toyMetabolite, even if its length would permit it. toyProteins / toyDimers bound to toyMetabolites cannot bind to promoters.

Interaction rules in toyLIFE have been devised to remove any ambiguity. When more than one rule could be chosen, we opted for computational simplicity, having made sure that the general properties of the model remained unchanged. A detailed list of the specific disambiguation rules implemented in the model follows:

1. **Folding rule:** if a sequence of toyAminoacids can fold into two (or more) different configurations with the same energy and two different perimeters with the same number of H, it is considered degenerate and does not fold.

2. **One-side rule:** any interaction in which a toyProtein can bind any ligand with two (or more) different sides and the same energy is discarded.

3. **Annihilation rule:** if two (or more) toyProteins can bind a ligand with the same energy, the binding does not

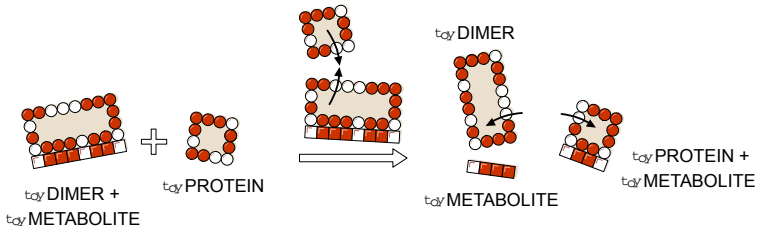


Supplementary Figure 5: Regulatory functions in toyLIFE. (a) A toyGene is expressed (translated) when the toyPolymerase binds to its promoter region. The sequence of Ps and Hs of the toyProtein will be exactly the same as that of the toyGene coding region. **(b)** If a toyProtein binds to the promoter region of a toyGene with a lower energy than the toyPolymerase does, it will displace the latter, and the toyGene will not be expressed. This toyProtein acts as an *inhibitor*. **(c)** The toyPolymerase does not bind to every promoter region. Thus, not all toyGenes are expressed constitutively. However, some toyProteins will be able to bind to these promoter regions. If, once bound to the promoter, they bind to the toyPolymerase with their rightmost side, the toyGene will be expressed, and these toyProteins act as *activators*. **(d)** More complex interactions —involving more elements— appear. For example, a toyProtein that forms a toyDimer with an inhibitor —preventing it from binding to the promoter— will effectively activate the expression of the toyGene. However, it does neither interact with the promoter region nor with the toyPolymerase, and its function is carried out only when the inhibitor is present. We call this kind of toyProteins *conditional activators*. **(e)** Two toyProteins can bind together to form a toyDimer that inhibits the expression of a certain toyGene. As they need each other to perform this function, we call them *conditional inhibitors*. As the number of genes increases, this kind of complex relationships can become very intricate.

occur. However, if a third toyProtein can bind the ligand with greater (less stable) energy than the other two, and does so uniquely, it will bind it.

4. **Identity rule:** an exception to the Annihilation rule occurs if the competing toyProteins are the same. In this case, one of them binds the ligand and the other(s) remains free.

5. **Stoichiometric rule:** an extension of the Identity rule. If two (or more) copies of the same toyProtein / toyDimer / toyMetabolite are competing for two (or more) different ligands, there will be binding if the number of copies of the toyProtein / toyDimer / toyMetabolite equals the number of ligands. For example, say that P1 binds to P2, P3 and P4 with the same energy. Then, (a) if P1, P2 and P3 are present, no complex will form; (b) if there are two copies of P1, dimers P1-P2 and P1-P3 will both form; but (c) if P4 is added, no complex will form. Conversely, if all ligands are copies as well, the Stoichiometry rule does not apply. For example, three copies of P1 and two copies of P2 will form two copies of dimer P1-P2, and one copy of P1 will remain free.



Supplementary Figure 6: Metabolism in toyLIFE. A toyDimer is bound to a toyMetabolite when a new toyProtein comes in. If the new toyProtein binds to one of the two units of the toyDimer, forming a new toyDimer energetically more stable than the old one, the two toyProteins will unbind and break the toyMetabolite up into two pieces. We say that the toyMetabolite has been catabolised.

1.3 Regulation

Expression of toyGenes occurs through the interaction with the toyPolymerase, which is a special kind of toyProtein (see Supplementary Figure 1). The toyPolymerase only has one interacting side (with sequence PHPH) and its folding energy is fixed to value -11.0: it is more stable than more than half the toyProteins. It is always present in the system. The toyPolymerase binds to promoters or to the right side of a toyProtein / toyDimer already bound to a promoter. When the toyPolymerase binds to a promoter, translation is directly activated and the corresponding toyGene is expressed (Supplementary Figure 5**a**). However, a more stable (lower energy) binding of a toyProtein or toyDimer to a promoter precludes the binding of the toyPolymerase. This inhibits the expression of the toyGene, except if the toyPolymerase binds to the right side of the toyProtein / toyDimer, in which case the toyGene can be expressed.

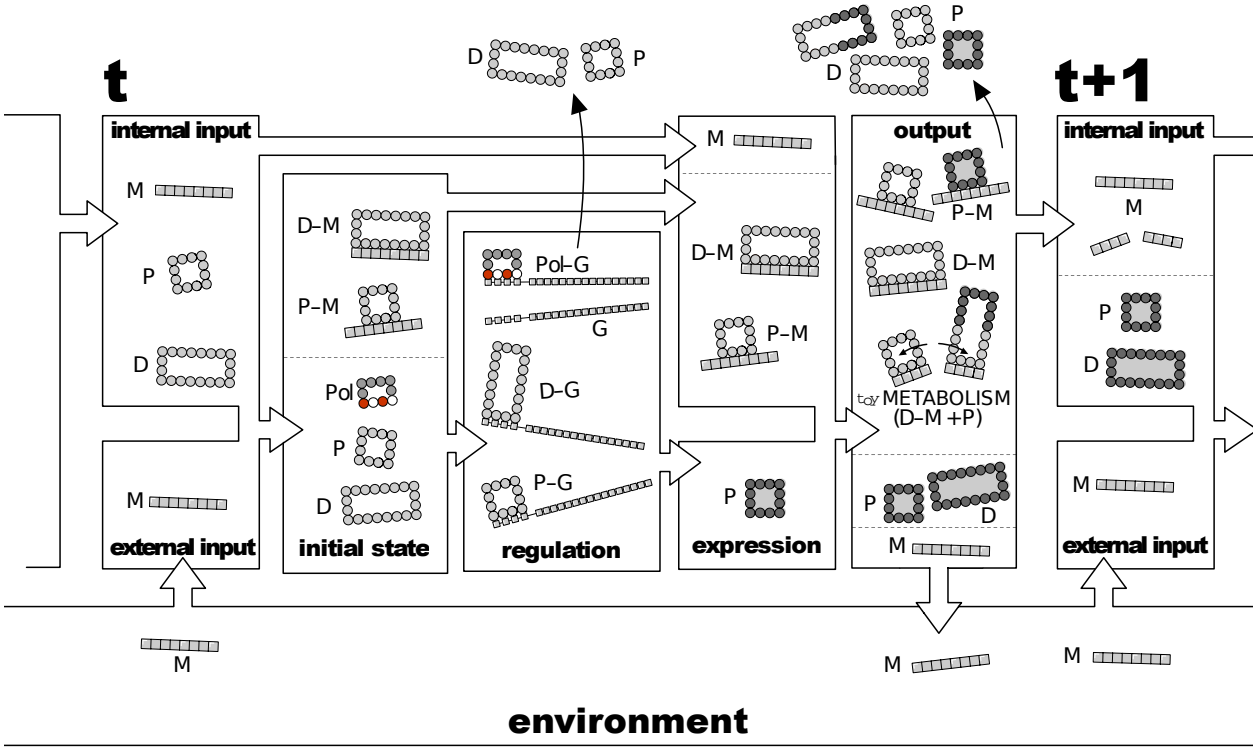
The minimal interaction rules that define toyLIFE dynamics endow toyProteins with a set of possible activities not included *a priori* in the rules of the model (see Supplementary Figure 5). For example, since the 4-toyN interacting site of the toyPolymerase cannot bind to all promoter regions —because some of these interactions have $E_{\text{int}} > E_{\text{thr}}$—, translation mediated by a toyProtein or toyDimer binding might allow the expression of genes that would otherwise never be translated. These toyProteins thus act as activators (Supplementary Figure 5**c**). This process finds a counterpart in toyProteins that bind to promoter regions more stably than the toyPolymerase does, and therefore prevent gene expression —this happens if $E_{\text{int(PROT)}} + E_{\text{PROT}} < E_{\text{int(POLY)}} + E_{\text{POLY}}$. They are acting as inhibitors (Supplementary Figure 5**b**). There are two additional functions that could not be foreseen and involve a larger number of molecules. A toyProtein that forms a toyDimer with an inhibitor —preventing its binding to the promoter— effectively behaves as an activator for the expression of the toyGene. However, it interacts neither with the promoter region nor with the toyPolymerase, and its activating function only shows up when the inhibitor is present. This toyProtein thus acts as a conditional activator (Supplementary Figure 5**d**). On the other hand, two toyProteins can bind together to form a toyDimer that inhibits the expression of a particular toyGene. As the presence of both toyProteins is needed to perform this function, they behave as conditional inhibitors (Supplementary Figure 5**e**). This flexible, context-dependent behavior of toyProteins is reminiscent of phenomena observed in real cells [9], and permits the construction of complex toyGene Regulatory Networks (toyGRNs).

1.4 Metabolism

When a toyDimer is bound to a toyMetabolite, another toyProtein can interact with this complex and break it. This reaction will take place if the toyProtein can bind to one of the subunits of the toyDimer and the resulting complex has less total energy than the toyDimer. As with the rest of interactions, the catabolic reaction will only take place if this binding is unambiguous. As a result of this reaction, the toyDimer will be broken in two: one of the pieces will be bound to the toyProtein (forming a new toyDimer), and the other one will remain free. The toyMetabolite will break accordingly: the part of it that was bound to the first subunit will stay with it, and the other part will stay with the second subunit. Note that the toyMetabolite need not be broken symmetrically: this will depend on how the toyDimer binds to it (Supplementary Figure 6).

1.5 Dynamics in toyLIFE

The dynamics of the model proceeds in discrete time steps and variable molecular concentrations are not taken into account. A step-by-step description of toyLIFE dynamics is summarised in Supplementary Figure 7. There is an initial



Supplementary Figure 7: Dynamics of toyLIFE. Input molecules at time step t are toyProteins (Ps) (including toyDimers (Ds)) and toyMetabolites, either produced as output at time step $t-1$ or environmentally supplied (all toyMetabolites denoted Ms). Ps and Ds interact with Ms to produce complexes P-M and D-M. Next, the remaining Ps and Ds and the toyPolymerase (Pol) interact with toyGenes (G) at the regulation phase. The most stable complexes with promoters are formed (Pol-G, P-G and D-G), activating or inhibiting toyGenes. P-Ms and D-Ms do not participate in regulation. Ps and Ds not in complexes are eliminated and new Ps (dark grey) are formed. These Ps interact with all molecules present and form Ds, new P-M and D-M complexes, and catabolise old D-M complexes. At the end of this phase, all Ms not bound to Ps or Ds are returned to the environment, and all Ps and Ds in P-M and D-M complexes unbind and are degraded. The remaining molecules (Ms just released from complexes, as well as all free Ps and Ds) go to the input set of time step $t+1$.

set of molecules which results from the previous time step: toyProteins (including toyDimers and the toyPolymerase) and toyMetabolites, either endogenous or provided by the environment. These molecules first interact between them to form possible complexes (see Section 1.2) and are then presented to a collection of toyGenes that is kept constant along subsequent iterations. Regulation takes place, mediated by a competition for binding the promoters of toyGenes, possibly causing their activation and leading to the formation of new toyProteins. Binding to promoters is decided in sequence. Starting with any of them (the order is irrelevant), it is checked whether any of the toyProteins / toyDimers (including the toyPolymerase) available bind to the promoter —remember that complexes bound to toyMetabolites are not available for regulation—, and then whether the toyPolymerase can subsequently bind to the complex and express the accompanying coding region. If it does, the toyGene is marked as active and the toyProtein / toyDimer is released. Then a second promoter is chosen and the process repeated, until all promoters have been evaluated. toyGenes are only expressed after all of them have been marked as either active or inactive. Each expressed toyGene produces one single toyProtein molecule. There can be more units of the same toyProtein, but only if multiple copies of the same toyGene are present.

toyProteins / toyDimers not bound to any toyMetabolite are eliminated in this phase. Thus, only the newly expressed toyProteins and the complexes involving toyMetabolites in the input set remain. All these molecules interact yet again, and here is where catabolism can occur. Catabolism happens when, once a toyMetabolite-toyDimer complex is formed, an additional toyProtein binds to one of the units of the toyDimer with an energy that is lower than that of the initial toyDimer. In this case, the latter disassembles in favor of the new toyDimer, and in the process the toyMetabolite is broken, as already mentioned in Section 1.4 and Supplementary Figure 6. The two pieces of

the broken toyMetabolites will contribute to the input set at the next time step, as will free toyProteins / toyDimers. However, toyProteins / toyDimers bound to toyMetabolites disappear in this phase —they are degraded—, and only the toyMetabolites are kept as input to the next time step. Unbound toyMetabolites are returned to the environment. This way, the interaction with the environment happens twice in each time step: at the beginning and at the end of the cycle.

2 A note on toyMetabolites

There are 2^m binary strings —toyMetabolites— of length m. From lengths 4 to 8, therefore, there are

$$\sum_{m=4}^{8} 2^m = 496$$

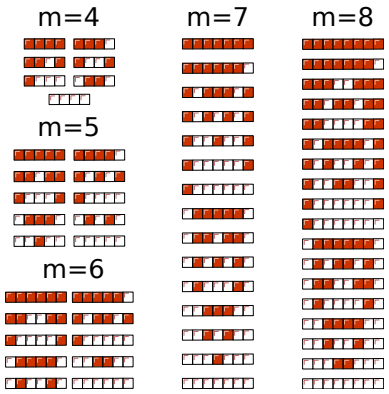
toyMetabolites. However, due to the interaction rules of tOYLIFE, a particular string and its reverse —i.e. HPPHPPPP and PPPPHPPH— will be treated the same way by tOYLIFE organisms. Therefore, for all practical purposes, we will consider each string and its reverse as the same toyMetabolite, thus staying with 274 of them. Additionally, there are 60 toyMetabolites that cannot be catabolised in tOYLIFE (Supplementary Figure 8). For all lengths, toyMetabolites formed by all Ps and one H at one extreme, or all Hs and one P at one extreme, are unbreakable. This is because there is no unambiguous way in which a toyDimer can bind to these toyMetabolites. There are two of these toyMetabolites for each length, making a total of 10. Additionally, the toyMetabolite PPHP cannot be broken due to the same reason. Symmetrical toyMetabolites, in general, cannot be catabolised either. Because of the interaction rules described in Section 1, only symmetrical toyDimers can bind to these toyMetabolites. But symmetrical toyDimers cannot be broken: any toyProtein that can bind to one subunit will be able to bind the other one. Because of the disambiguation rules, no binding is produced, and catabolism does not occur. There are 52 symmetric toyMetabolites —because they repeat half the sequence, there are

$$\sum_{m=4}^{8} 2^{\left[\frac{m+1}{2}\right]} = 52$$

of them, $[x]$ being the integer part of x —odd-length symmetrical toyMetabolites repeat $m+1$ toySugars, hence the $[(m+1)/2]$ exponent. However, three symmetrical toyMetabolites of length 7 —namely, PPPHPPP, PPHPHPP and PPHHHPP— can actually be broken. So there are 49 unbreakable symmetrical toyMetabolites. Added to the previous 11 unbreakable toyMetabolites, we get the total of 60. As a result, the total number of toyMetabolites up to length 8 is 214.

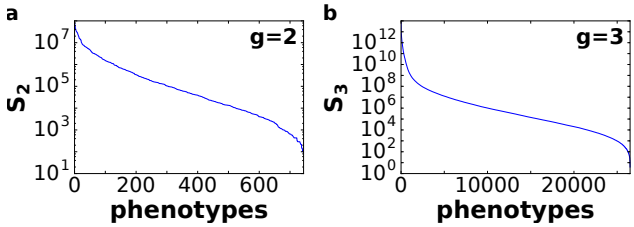
It is somewhat interesting that, as an emergent property of the model, some toyMetabolites are not able to be catabolised. Moreover, it is not that these toyMetabolites are irrelevant to the model: if they are present, they will interact with symmetric toyDimers, affecting the regulatory output of cells. So these toyMetabolites could function as signalling molecules.

What happens with longer toyMetabolites? Because of the way interactions have been defined in tOYLIFE, longer toyMetabolites can be considered as unions of shorter ones. For instance, a toyMetabolite of length 9 is (in terms of interactive potential) equal to two toyMetabolites of length 8. If a genotype is able to catabolise one of these, it will be able to catabolise the longer one, so the metabolic phenotype for toyMetabolites of arbitrary length is uniquely determined by considering lengths up to 8 toySugars.



Supplementary Figure 8: Unbreakable toyMetabolites. There are 60 unbreakable toyMetabolites: 49 of them are symmetrical, other 10 are chains of all Hs or all Ps in a row, and the last one is PPHP. Because of the interaction rules in tOYLIFE, only symmetrical toyDimers would be able to bind these toyMetabolites, and therefore they cannot be broken.

3 Rank plots for phenotypes in $g = 2$ and $g = 3$



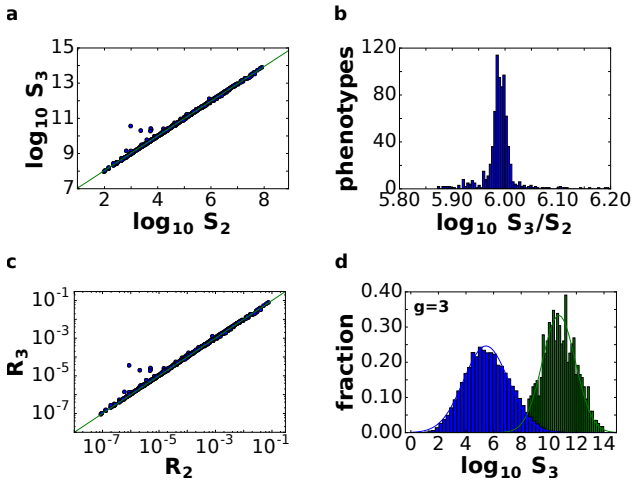
Supplementary Figure 9: Phenotype frequencies vary enormously in toyLIFE. Rank plots for all phenotypes in $g = 2$ **(a)** and $g = 3$ **(b)**. Both plots show a long tail of small phenotypes. In particular, for $g = 3$, only 300 phenotypes in $\mathcal{P}_3$ represent almost 99% of all genotypes. The remaining $26,000$ phenotypes are extremely rare by comparison.

4 Comparison between the $g = 2$ and $g = 3$ case

In Supplementary Figure 10**a**, we represent the abundance of a phenotype for $g = 2$ (S_2) versus its corresponding abundance for $g = 3$ (S_3), for each phenotype in $\mathcal{P}_2$. The Figure also shows a power-law fit, $\log_{10} S_3 = 6.064 + 0.986 \log_{10} S_2$, corresponding to $S_3 = 10^{6.064} S_2^{0.986} \approx 10^6 S_2$, a linear fit. This means that the abundance ordering between these phenotypes does not change when exploring genotypes with one more gene. The goodness of the fit is further shown in Supplementary Figure 10**b**, which represents the histogram of values of $\log_{10}(S_3/S_2)$. The distribution is concentrated around its mean, 5.996, very close to the value 6.064 obtained in Supplementary Figure 10**a**. This second result confirms that the abundance of $\mathcal{P}_2$ phenotypes in $g = 3$ space is equal to their corresponding abundance in $g = 2$ space times 10^6. Where does this factor come from? Recall that there are $2^{20} \sim 10^6$ toyGenes in toyLIFE. A factor of almost 10^6 between S_3 and S_2 means that we can add almost any toyGene to a given two-gene genotype, and the resulting phenotype will be the same: it will not interfere with the original function. This is a remarkable fact.

Moreover, if we look at the distribution of relative abundances of $\mathcal{P}_2$ phenotypes —computed as phenotype abundance divided by the total number of viable genotypes— for $g = 2$ and $g = 3$ (Supplementary Figure 10**c**), we obtain a linear relationship again: $R_3 = R_2$. Which means that the relative abundance of the phenotypes for $g = 2$ is very similar to the relative abundance they represent for $g = 3$. But the sum of the relative abundances for $g = 2$ is equal to 1 —there are only 775 phenotypes in $\mathcal{P}_2$. Accordingly, the sum of relative abundances for $g = 3$ is close to 1 —actually, it is 0.9964.

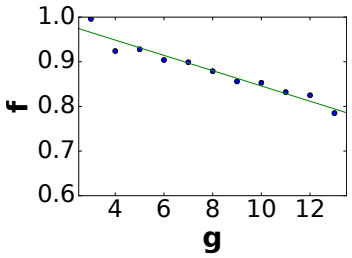
Finally, let us look again at the histogram of phenotype abundance distributions in $g = 3$ that we obtained in Figure 1**c** (main text). We can re-compute the histogram taking the 775 phenotypes from $\mathcal{P}_2$ as a separate set from the remaining $25,717$ phenotypes in $\mathcal{P}_3 - \mathcal{P}_2$. If we compute the respective histograms for both sets, we obtain Supplementary Figure 10**d**. In green we have represented the 775 phenotypes in $\mathcal{P}_2$. It is not surprising that their distribution follows a log-normal law again: it follows immediately from Figure 1**a** (main text) and from the linear relationship shown in Supplementary Figure 10**a**. What is relevant, however, is that the *bump* we observed in Figure 1**c** (main text) is gone in the histogram of the remaining $25,717$ phenotypes (in blue).



Supplementary Figure 10: Two-gene phenotypes dominate phenotype space in the three-gene case. (a) The 775 phenotypes belonging to $\mathcal{P}_2$ also appear in $\mathcal{P}_3$. This figure represents the corresponding abundance of each phenotype in both genotype spaces: S_2 and S_3 are, respectively, phenotype abundance for $g = 2$ and $g = 3$. Green line represents the linear fit $\log_{10} S_3 = 6.064 + 0.986 \log_{10} S_2$, which is close to the linear fit $S_3 \sim 10^6 S_2$. **(b)** Histogram of $\log_{10}(S_3/S_2)$ for each of the 775 phenotypes in $\mathcal{P}_2$. The mean of the distribution is 5.996. **(c)** Relative abundance of the 775 phenotypes in $\mathcal{P}_2$ (R_2) versus their relative abundance for $g = 3$ (R_3) —computed as phenotype abundance divided by number of viable genotypes. Green line is $R_3 = R_2$. The close fit means that the phenotypes from $\mathcal{P}_2$ dominate phenotype space for $g = 3$. **(d)** Abundance distribution of phenotypes in $\mathcal{P}_3$, taking the 775 phenotypes in $\mathcal{P}_2$ and rescaling them — we have obtained the two histograms as if they came from independent distributions for clarity. The green histogram represents the phenotypes in $\mathcal{P}_2$, and the blue histogram the remaining $25,717$ phenotypes in $\mathcal{P}_3$. New log-normal fits are drawn: $\mu_3 = 5.449$, $\sigma_3 = 1.619$ (blue line), $\mu_2 = 10.730$, $\sigma_2 = 1.196$ (green line). Note that the log-normal fit for three-gene phenotypes is much better once we take into account the 775 phenotypes in $\mathcal{P}_2$. All fits in this and subsequent Supplementary Figures have been done using the least squares method.

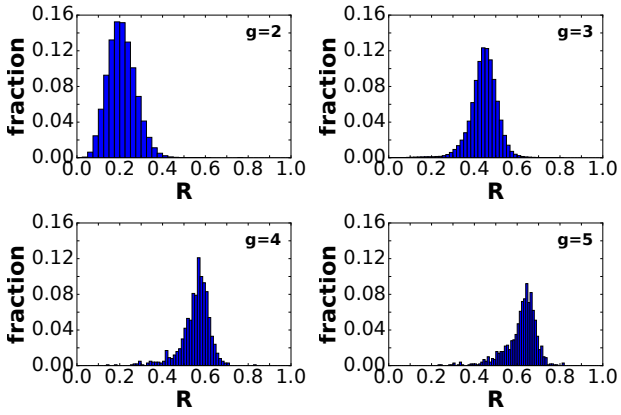
5 Relevance of $\mathcal{P}_2$ phenotypes

A relevant question now is how important the 775 phenotypes in $\mathcal{P}_2$ are for larger genotypes. Exhaustive sampling of genotype spaces larger than $g = 3$ is out of our possibilities, but we can perform random samples of genotypes for different values of g and observe the fraction f of observed phenotypes that belong to $\mathcal{P}_2$. This is represented in Supplementary Figure 11. Observe that, although this fraction decays linearly with gene size as $f = 1.02 - 0.02g$, the slope of the decay is very small, and therefore the fraction is always high —higher than 80% for $g \leq 13$. In other words, phenotypes in $\mathcal{P}_2$ continue to dominate phenotype space in tOYLIFE for a moderate number of genotype sizes.



Supplementary Figure 11: The dominance of two-gene phenotypes decays linearly with genotype size. For each g, we sample $10,000$ viable genotypes and compute their phenotypes, counting how many phenotypes belong to $\mathcal{P}_2$. We then represent the fraction f versus g. The data can be fitted to a linear function: $f = 1.02 - 0.02g$ (green line). The fraction of phenotypes belonging to $\mathcal{P}_2$ decays with g, albeit very slowly.

6 Robustness histograms in toyLIFE



Supplementary Figure 12: Genotypes in toyLIFE typically have a large number of neutral neighbors. Distribution of robustness for genotypes for different values of g (gene number) for $g = 2$ to $g = 5$. Robustness is defined as the normalized degree of a node in the networks: $R = k/k_{\max}$, where k is the degree of a node in the neutral network, and $k_{\max} = 20g$ is the maximum degree in the network. Normalisation allows us to compare values for different genotype sizes. For $g = 2$ and $g = 3$, we sampled 10^7 genotypes, whereas for $g = 4$ and $g = 5$ we sampled $1,000$ genotypes. All distributions are unimodal, and more or less concentrated around the mean.

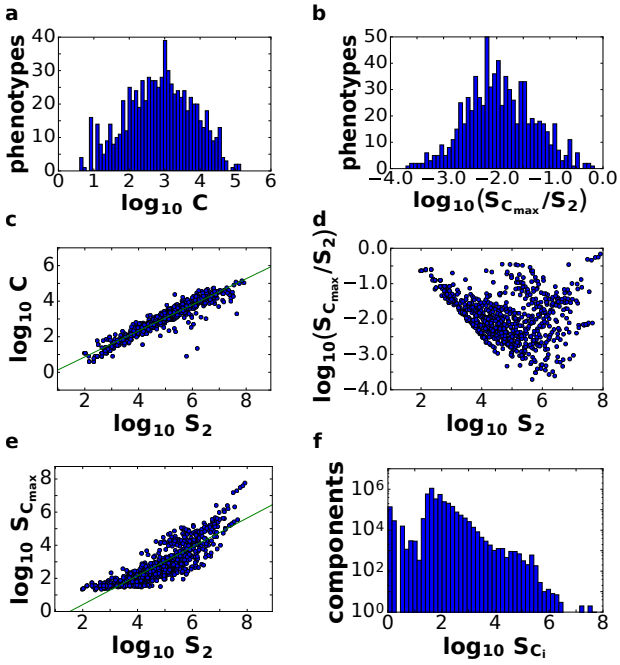
7 Connected components for $g = 2$

For $g = 2$ we can perform network analyses on all 775 phenotypes exhaustively, and compute their connected components (Supplementary Figure 13). We observe that most phenotypes are distributed in highly fragmented neutral networks: the genotypes corresponding to a given phenotype cluster in many disjoint connected components (Supplementary Figure 13**a**): the number of connected components C is never smaller than 4 and is usually much larger. Moreover, these connected components tend to be small: if we consider Cmax, the maximal component associated to each neutral network, its average relative abundance $S_{C\text{max}}/S_2$ is 0.033 (Supplementary Figure 13**b**). Only 63 phenotypes have connected components that are larger than 10% the phenotype abundance —among these are the largest connected components for $g = 2$, including one giant network that contains $56,889,472$ nodes!

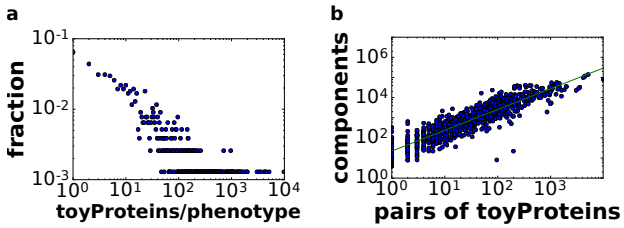
Large phenotypes tend to have a larger number of connected components, and we can find a relatively good power-law fit between the abundance of the phenotype S_2 and the number of components C: $C = 0.25S_2^{0.7}$ (Supplementary Figure 13**c**). The relationship between S_2 and the relative size of Cmax is noisy (Supplementary Figure 13**d**): smaller phenotypes have less connected components and therefore the relative size of the maximal component is high. As the number of components increases, most of them tend to have equal, small sizes. However, the largest phenotypes with the greatest number of connected components also have the largest connected components, as we pointed out before, so there is a positive correlation between S_2 and the absolute size of its maximal component, $S_{C\text{max}}$. This last fact is represented in Supplementary Figure 13**e**.

In short, there is a huge variation in the size of connected components in $g = 2$. We can plot the distribution of sizes of all connected components C_i —irrespective of the phenotype they belong to (Supplementary Figure 13**f**). The average component size, S_{C_i}, is 301.4, but we can see from the histogram that the distribution has a long tail. Therefore, although most connected components are smaller than $1,000$ nodes — roughly 98.5%!— some of them reach up to $\sim 10^7$ nodes.

The high disconnection in connected components is due to the HP model that underlies toyProtein folding. Any



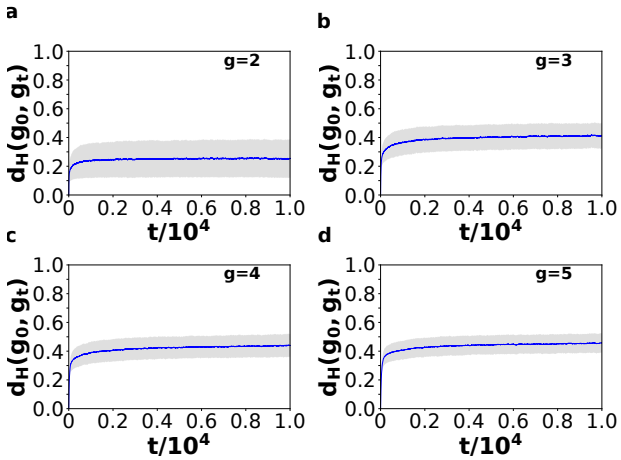
Supplementary Figure 13: Neutral networks in toyLIFE are highly fragmented for $g = 2$. (a) For all 775 phenotypes in $\mathcal{P}_2$, we computed the number of connected components (C) of the associated neutral network. This figure represents the distribution of the decimal logarithm of C per neutral network. No single phenotype has less than 4 connected components. **(b)** For each neutral network, we take the maximal component Cmax and plot the distribution of the logarithm of its relative size —that is, the logarithm of $S_{C\text{max}}$ divided by S_2. **(c)** The abundance of the phenotype and the number of components are related via a power law: $C = 0.25S_2^{0.7}$. **(d)** The relationship between the relative abundance of Cmax and the abundance of the phenotype is very noisy, but **(e)** there is a positive correlation between the absolute abundance of Cmax and the abundance of the phenotype. The green line represents the power law fit $S_{C\text{max}} = 0.05_2^{0.9}$. **(f)** Distribution of the logarithm of abundance of all connected components C_i for $g = 2$.



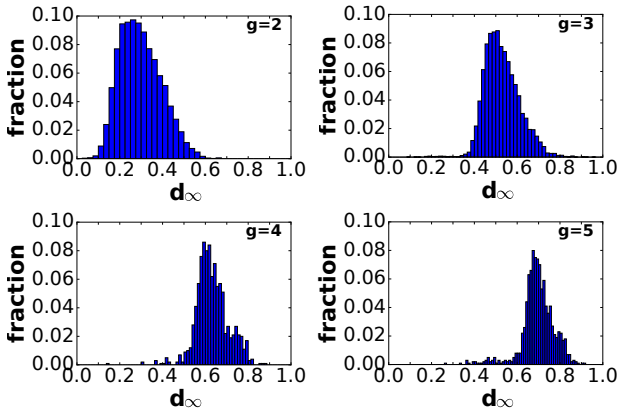
Supplementary Figure 14: Most phenotypes in $\mathcal{P}_2$ are obtained by a small number of pairs of toyProteins. (a) Distribution of the number of pairs of toyProteins that generate a given phenotype. For example, if both $\{1,1\}$, $\{1,2\}$ and $\{3,4\}$ generate a given phenotype, there are 3 pairs of toyProteins that generate it. **(b)** Due to the HP model that underlies toyProtein folding, the more pairs of toyProteins are able to generate a given phenotype, the larger the phenotype and, because of the power-law relationship obtained in Supplementary Figure 13**c**, the more connected components that will belong to the phenotype. The green line represents the power-law fit $C = 22.093P^{1.032}$.

given phenotype in $\mathcal{P}_2$ will be obtained by some set of pairs of toyProteins. Supplementary Figure 14**a** shows that this distribution is highly skewed, with a long tail: 28.64% of phenotypes in $\mathcal{P}_2$ are obtained by less than 10 pairs of toyProteins, while one phenotype is obtained by $9,808$ pairs of toyProteins. The problem, therefore, is not due to a small set of toyProteins associated to each phenotype. Rather, the cause of the disconnection between connected components is due to the lack of neutral mutations among proteins and the difficulty to reach different proteins in toyLIFE (see main text).

8 Random walks in toyLIFE



Supplementary Figure 15: Neutral networks in toyLIFE span a large fraction of genotype space (1). For each genotype size, from $g = 2$ to $g = 5$, we performed $1,000$ neutral random walks starting at randomly chosen genotypes. The length of the random walks was $10,000$ time steps. The figure represents the average Hamming distance $\langle d_H \rangle$ (blue line) between the genotype visited at time t, g_t, and the original genotype g_0, plus minus one standard deviation (grey area), empirically obtained from the data.

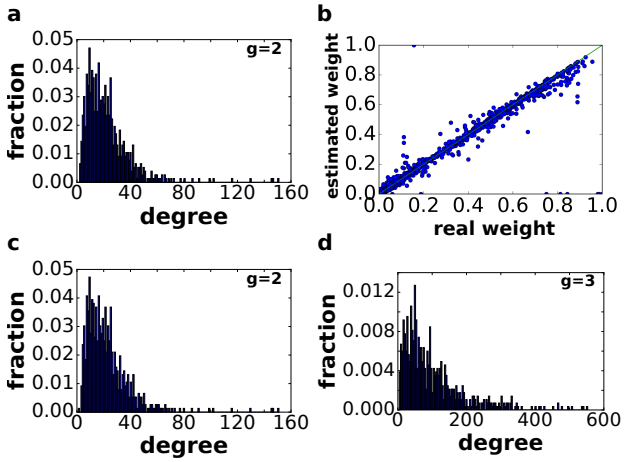


Supplementary Figure 16: Neutral networks in toyLIFE span a large fraction of genotype space (2). Distribution of d_∞ for genotypes with different values of g (gene number) for $g = 2$ to $g = 5$. We performed $10,000$ (for $g \leq 3$) or $1,000$ (for $g > 3$) neutral random walks, forcing them to increase the Hamming distance to the original genotype. We stopped when the random walk could not get farther.

9 Connections between phenotypes

For $g = 2$ we can build the phenotype network exhaustively. The network is not entirely connected: there is a giant component that includes 767 nodes out of the 775, and six additional tiny components, five of them with just one node and the remaining one with three nodes. Additionally, the results show that the average degree is low, just 22.1, with a standard deviation of 17.3 (Supplementary Figure 17**a**). The maximum degree is 151 and the minimum is 2. The largest weights are always those of the self-loops —that is, the majority of connections in the genotype network do not change phenotype, consistently with our previous discussion on robustness. In fact, because not all phenotypes are equally large, we can compute the weighted average degree of the network —giving more weight to larger phenotypes. The result is an average degree of 54.0, illustrating that larger phenotypes are more connected than the average.

For $g = 3$, we cannot build the phenotype network exhaustively. We will resort to a numerical approximation, in order to estimate the degrees of the nodes and their relative weights. Suppose we perform a random walk over all viable genotypes —jumping among them without any additional rule. If all genotypes are connected to each other —given our results for $g = 2$, this does not seem a terrible assumption— then we expect that, as the length of the random walk tends to infinity, every phenotype is visited proportionally to its abundance, and that the visits from one phenotype to another are proportional to the actual number of connections between them. The average number of visits (per time step) from phenotype i to j as time tends to infinity will be the same as the number of connections between phenotypes i and j, divided by the total number of connections leaving i. We can check if this approach is accurate by performing the random walk for $g = 2$, for which we have the actual connection data. We performed a random walk starting at a randomly chosen genotype for 10^9 time steps. The relative weights computed by this method are close to the actual weights, as shown in Supplementary Figure 17**b**. The correlation between both variables is 0.978: the outliers correspond to small phenotypes, which are hardly visited in the random walk. Supplementary Figure 17**c** shows that the estimated degree distribution is very similar to the actual one (Supplementary Figure 17**a**). Having made sure that this approach works, we repeated it for $g = 3$, again with a random walk of length 10^9 time steps. We restricted the random walk to the 775 phenotypes in $\mathcal{P}_2$: we wanted to study how the addition of one gene altered the connections between these phenotypes. When one mutation left this set of phenotypes, we considered it as lethal. The results obtained show that all phenotypes in $\mathcal{P}_2$ now belong to one giant component —there is one phenotype that does not appear in the sample, but did belong to the giant component in $g = 2$, so it must belong to it in $g = 3$. The average degree is higher, 101.1, with a standard deviation of 90.3 (Supplementary Figure 17**d**). The maximum degree is 553, and the minimum is 4. The degree distribution is much wider, and the connectivity between phenotypes has been greatly enhanced. The weighted average degree is 333.3, again showing that larger phenotypes are much more connected than smaller ones.



Supplementary Figure 17: Connections between phenotypes in toyLIFE. (a) Degree distribution of the phenotype network in $g = 2$. Two phenotypes are connected if there is at least one genotype belonging to the first that can mutate into another genotype belonging to the second phenotype. The average degree is 22.134. **(b)** Estimated relative weight between phenotypes versus actual relative weight. Estimation performed by a random walk among all viable genotypes in $g = 2$. Length of the random walk is 10^9. The correlation between both variables is 0.978. **(c)** Estimated degree distribution from the previous random walk, for $g = 2$. **(d)** Estimated degree distribution for $g = 3$, using a random walk among genotypes belonging to phenotypes in $\mathcal{P}_2$.

References

[1] Arias CF, Catalán P, Manrubia S, Cuesta JA. toyLIFE: a computational framework to study the multi-level organisation of the genotype-phenotype map. Sci Rep. 2014;4:7549.

[2] Li H, Helling R, Tang C, Wingreen N. Emergence of preferred structures in a simple model of protein folding. Science. 1996;273:666–669.

[3] Dill KA. Theory for the folding and stability of globular proteins. Biochemistry. 1985;24:1501–1509.

[4] Aharoni A, Gaidukov L, Khersonsky O, Gould SM, Roodveldt C, Tawfik DS. The'evolvability'of promiscuous protein functions. Nat Genet. 2005;37:73–76.

[5] Amitai G, Gupta RD, Tawfik DS. Latent evolutionary potentials under the neutral mutational drift of an enzyme. HFSP J. 2007;1:67–78.

[6] Khersonsky O, Tawfik DS. Enzyme promiscuity: a mechanistic and evolutionary perspective. Ann Rev Biochem. 2010;79:471–505.

[7] Hayden EJ, Ferrada E, Wagner A. Cryptic genetic variation promotes rapid evolutionary adaptation in an RNA enzyme. Nature. 2011;474:92–95.

[8] Hoque T, Chetty M, Sattar A. Extended HP model for protein structure prediction. J Comput Biol. 2009;16:85–103.

[9] Piatigorsky J. Gene Sharing and Evolution: the Diversity of Protein Functions. Harvard University Press Cambridge MA:; 2007.