Title:

Mutational biases contribute to the neutral complexification of protein interaction networks following gene duplication

Authors:

Angel F. Cisneros[1,2,3,4], Lou Nielly-Thibault[2,3,4,5], Saurav Mallik[6], Emmanuel D. Levy[6], Christian R. Landry[1,2,3,4,5]

Affiliations:

[1]Département de biochimie, de microbiologie et de bio-informatique, Faculté des sciences et de génie, Université Laval, G1V 0A6, Québec, Canada
[2]Institut de biologie intégrative et des systèmes, Université Laval, G1V 0A6, Québec, Canada
[3]PROTEO, Le regroupement québécois de recherche sur la fonction, l'ingénierie et les applications des protéines, Université Laval, G1V 0A6, Québec, Canada
[4]Centre de recherche sur les données massives, Université Laval, G1V 0A6, Québec, Canada
[5]Département de biologie, Faculté des sciences et de génie, Université Laval, G1V 0A6, Québec, Canada
[6]Department of Chemical and Structural Biology, Weizmann Institute of Science, 7610001, Rehovot, Israel

Correspondance:
Tel: +1 418 656 3954; E-mail: christian.landry@bio.ulaval.ca

Abstract

Biological systems can gain complexity over time. While some of these transitions are likely driven by natural selection, the extent to which they occur without providing an adaptive benefit is unknown. At the molecular level, one example is heteromeric complexes replacing homomeric ones following gene duplication. Here, we build a biophysical model and simulate the evolution of homodimers and heterodimers following gene duplication using distributions of mutational effects inferred from available protein structures. We keep the specific activity of each dimer identical, so their concentrations drift neutrally in the absence of new functions. We show that for more than 60% of tested dimer structures, the relative concentration of the heteromer increases over time due to mutational biases that favor the heterodimer. However, allowing mutational effects on synthesis rates and differences in the specific activity of homodimers and heterodimers can limit or reverse the observed bias toward heterodimers. Our results show that the accumulation of more complex protein quaternary structures is likely under neutral evolution, and that natural selection would be needed to reverse this tendency.

Keywords

Gene duplication / mutational biases / neutral evolution / protein biophysics / protein interaction networks

Introduction

Gene duplication is a potent mechanism for the generation of novelties (Conant & Wolfe, 2008; Kuzmin *et al*, 2022). Duplication introduces an additional copy of a gene, which may allow sister copies to accumulate mutations that may not have been accessible before. The long-term maintenance of the duplicates could depend on one or both evolving new functions (Deng *et al*, 2010; Bridgham *et al*, 2008). Alternatively, the two copies could split the functions of the ancestral protein (Force *et al*, 1999; Lynch & Conery, 2000; He & Zhang, 2005; Baker *et al*, 2013). Under this model, the two copies show partial degeneration of their ancestral function and complement each other in accomplishing the ancestral function. As a result, cellular biochemical or genetic networks become more complex without net functional or fitness gain.

One context in which this increase in complexity takes place is the assembly of protein complexes. Following the duplication, two paralogous genes encode the product of the ancestral gene. In the case of the duplication of a gene encoding a homomeric protein complex (self-interacting protein), the paralogous proteins might assemble in a mixture of states (Pereira-Leal *et al*, 2007): two homodimers and a heterodimer in the case of dimers, with further combinatorial complexity depending on the number of subunits in the ancestral homomer. For the sake of simplicity and because they are very frequent in protein interaction networks, we will focus on dimers here. Indeed, homodimers amount to almost 66% of homomers and no other multimeric assembly exceeds 15% of homomers, with some variation depending on the species (Lynch, 2012, 2013; Levy & Teichmann, 2013). An important consideration is that heteromers directly become the state with the highest concentration right after the duplication, and the energetic barrier needed for selective homomerization increases with the number of subunits (Hochberg *et al*, 2018). For example, the duplication of an ancestral homodimer would lead to 50% heterodimers and 25% of each homodimer at binding equilibrium if all their parameters are the same (synthesis rates, folding energies, binding affinities, etc.). As such, heteromers are expected to assume a significant proportion of the functional contributions directly after the duplication. This is a major biological change that occurs immediately following gene duplication.

Heteromers of paralogs play a major role in modern protein interaction networks. These new heteromeric interactions have been shown to be retained in more than 30% of paralogs in several species and are especially prevalent for duplicates with higher sequence identity (Marchant *et al*, 2019; Mallik & Tawfik, 2020). Interestingly, in some of these cases the heteromer becomes the only active complex (Wang *et al*, 2006; Boncoeur *et al*, 2012) or diverges functionally (Bridgham *et al*, 2008; Baker *et al*, 2013; Paul & Hristova, 2019; Li *et al*, 2022). A heteromer taking over the function of the ancestral homomer is an example of degeneration with complementation whereby a function encoded by a single gene is now encoded by two genes. A consequence of such transitions toward heterodimers is that protein interaction networks do not benefit from the robustness usually brought about by redundancy introduced by duplication (Diss *et al*, 2017; Dandage & Landry, 2019). One major question that these observations have led to is how and why these transitions would occur.

One main question therefore becomes: is natural selection favoring the transition from homodimers to heterodimers, or could neutral processes be sufficient? Many arguments are in favor of adaptive changes. Changes in assembly mode can be adaptive, since different homomeric and heteromeric symmetries facilitate specific properties such as allostery and transport better than monomers (Bergendahl & Marsh, 2017). There are some compelling and potentially adaptive examples of retained heteromers. For instance, Pillai et al. (2020) thoroughly described the emergence of allosteric cooperativity in heterotetrameric haemoglobin, which is limited in homotetramers of the beta (Kurtz *et al*, 1981) or gamma subunits (Kidd *et al*, 2001). Other examples include the BfER steroid receptor that negatively regulates its paralog BfSR (Bridgham *et al*, 2008), the Mcm1 and Arg80 transcription factors whose homomers and heteromers bind specific targets (Baker *et al*, 2013), and half-ABC transporters that are only capable of transporting specific molecules out of the cell when forming heterodimers (Boncoeur *et al*, 2012). In these three examples, heteromerization between paralogs appears to be an important element contributing to the new function or feature. However, recent research has shown that protein interaction interfaces that do not provide an adaptive benefit can become entrenched as mutations accumulate (Finnigan *et al*, 2012; Hochberg *et al*, 2020; Schulz *et al*, 2022). A notable example is the fungal V-ATPase ring, which after a duplication shifted from a 5:1 stoichiometry to a 4:1:1 stoichiometry, resulting in increased complexity without clear changes in function (Finnigan *et al*, 2012). Thus, natural selection may not be necessary to explain why many extant heteromers were retained.

The above examples suggest that a fraction of extant heteromers of paralogs might have provided a limited adaptive benefit. Furthermore, adaptive benefits could occur through the emergence of new functions over long timescales, while the initial fixation of a heteromer of paralogs may not have been adaptive. Indeed, different models have been proposed to explain how complexity could arise neutrally (Stoltzfus, 1999; Lynch, 2007; Muñoz-Gómez *et al*, 2021). In these models, the increase in complexity does not provide a fitness advantage or a disadvantage, but instead tends to become prevalent due to drift and mutational biases. Mutational biases refer to particular types of mutations (or their effects) occurring more often than others, which influences the supply of available mutations and ultimately adaptation (Svensson & Berger, 2019; Cano *et al*, 2022). In the case of homomers and heteromers, there could be mutational biases with respect to whether mutations favor the assembly of one type of complex over the other.

Theoretical models could contribute to setting a null expectation, allowing us to distinguish between the effects of natural selection and neutral evolution. Intuitively, if mutations naturally disrupt heteromers or favor homomerization, heteromers would tend to disappear neutrally. On the contrary, if mutations tend to favor the maintenance of the heteromers, then natural selection against these mutations or other mechanisms like compartmentalization and differential regulation would be necessary to separate the proteins in space or time and remove the heteromer (Marchant *et al*, 2019). The effect of individual mutations is expected to be amplified in homomers since the mutant is present in all subunits, whereas only some of the subunits of heteromers would be mutant copies. Hence, since mutational effects on binding

affinity are more often destabilizing than stabilizing, one could expect the heteromer to prevail over the homomers. The inherent difference in expected effects of mutations could be potentiated by any mutational biases in terms of which types of mutations happen more often. Overall, the interplay between the effects of mutations, along with natural selection, should determine the fate of the paralogs. As such, we hypothesize that, in a neutral scenario, heteromers will slowly become more abundant while homomers wither away due to the accumulation of slightly destabilizing mutations.

Here, we use simulations to examine the fate of heterodimeric complexes in a context where selection does not differentiate between homodimers and heterodimers. Natural selection acts on the total amount of protein complexes formed so all three molecular complexes are free to explore a large space of possible binding equilibrium neutrally within the limit set by selection. We characterize the relative impact of several parameter values that can modify the proportions of homodimers (HM) and heterodimers (HET), including changes in the synthesis rates and folding energies of individual subunits, as well as the binding affinities and specific activities of their homodimers and heterodimers. Finally, we examine the relative advantages that homodimers and heterodimers would need to provide to become the dominant type of complex after the duplication.

Results

The equilibrium between homomers and heteromers is highly sensitive to mutational effects on binding energy

We examined protein-protein interactions after gene duplication. We first built a model that allows us to estimate the concentrations of both homodimers and the heterodimers at equilibrium depending on several parameters that can diverge between paralogs as mutations accumulate (Figure 1A). These parameters are their synthesis rates (which encompass all steps leading to protein synthesis, including transcription and translation), their folding energy, their degradation and their association into homodimers and heterodimers. The folding energies of the proteins are used to determine the fraction of folded protein copies (Sailer & Harms, 2017), which can later form homodimers or heterodimers based on their binding affinity. The concentration of the three dimers is our focal trait of interest, and we can estimate their concentration at binding equilibrium using synthesis and decay rates, folding and binding energies. The full derivations of the systems of equations used to estimate concentrations at equilibrium before and after the duplication are shown in Supplementary Files 1 and 2, respectively. Later, in the evolutionary modelling, these parameters are allowed to evolve through mutations.

To examine how changes in these parameters will map to fitness, we defined a function that links fitness to the concentration of the total amount of protein complex formed (Figure 1B). Since we are interested in examining how neutral forces contribute to the transition from homomers to heteromers without the evolution of new functions, selection only acts on the total amount of complex formed (which satisfies the function) and does not distinguish between homodimers and heterodimers. These assumptions will be relaxed later, as we include more parameters that could directly favor one kind of dimer over the other.

We first characterized the sensitivity of the post-duplication equilibrium concentrations of homodimers and heterodimers to changes in the various parameter values. While preserving identical parameter values for both genes after the duplication, the relative concentrations of the three complexes will be 1AA:2AB:1BB. We then examined the post-duplication equilibrium concentration of the three complexes by testing different values for some of the parameters (Figure 1C-J). We first studied the effects of different combinations of folding energy values (Figure 1C-F) around the typical value of -5 kcal/mol (Pace, 1975; Plaxco *et al*, 2000). As expected, destabilizing the fold of only one of the paralogs leads to a decrease in the number of available copies of that protein. As a result, the concentrations of its corresponding homodimer and the heterodimer decrease as well, with an increase in the concentration of the other homodimer (Figure 1C-D). We note that the percentage of heterodimers never exceeds 50% when only folding energies change, and that it is quickly reduced as either of the two subunits is destabilized. Destabilizing either or both of the protein subunits beyond a ΔG_{fold} of 0 leads to a reduction in the total concentration of dimers, and thus in their total activity (Figure 1E) and fitness (Figure 1F). Overall, the effect of folding energy on the balance between homodimers

and heterodimers is limited as long as $\Delta G_{fold} < 0$ for both subunits. Once ΔG_{fold} for one of the subunits becomes greater than zero, the number of properly folded subunits of that protein decreases rapidly, and thus equilibrium shifts toward the other homodimer.

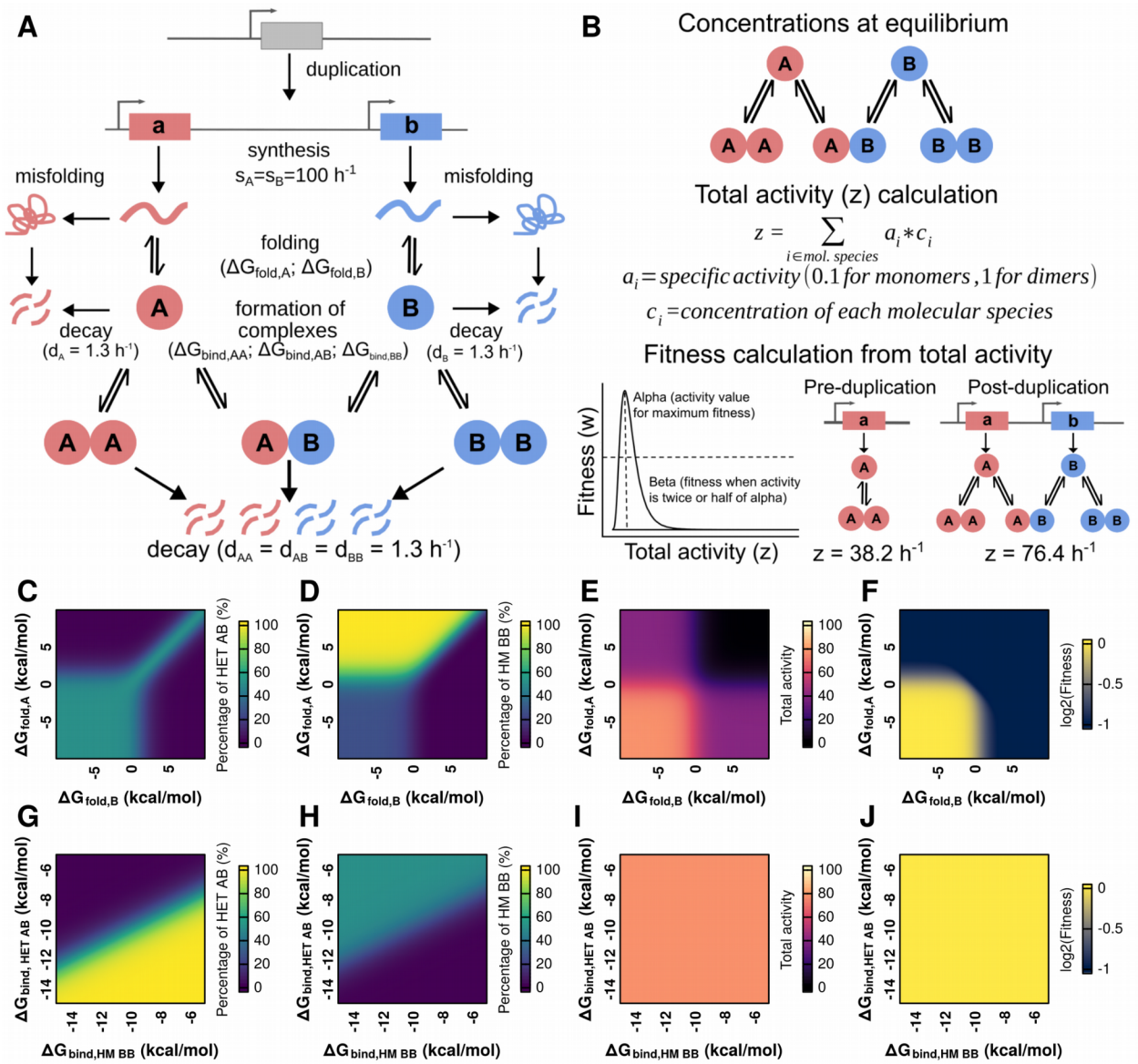

Figure 1. Changes to folding energy and binding affinity modify the concentration of each dimer and the fitness of the system.

A) Post duplication system. Paralogous proteins are produced with synthesis rates s_A and s_B. These parameters are meant to encompass all steps of protein synthesis (transcription and translation, which are not considered separately here). The probability of correct folding is calculated based on the folding energy (ΔG_{fold}) following Sailer & Harms (2017) (see Methods). Misfolded proteins are removed from the system. Correctly folded proteins assemble into homodimers and heterodimers based on binding energies ($\Delta G_{bind,HM\ AA}$, $\Delta G_{bind,HET\ AB}$, $\Delta G_{bind,HM\ BB}$).

B) The concentrations of each molecular species (monomers, homodimers, and heterodimers) are estimated at

equilibrium based on these variables and their decay rates. The total activity of the system is calculated by multiplying the equilibrium concentrations of each molecular species by their specific activity (0.1 for monomers, 1 for both homodimers and the heterodimer). Fitness is estimated based on the total activity of the system using a lognormal function given by two parameters: alpha and beta. With the parameters for synthesis rates and decay rates shown in figure 1A, the estimated values of total activity for the system before and after the duplication corresponded to 38.2 h^{-1} and 76 h^{-1}.

C-F) Effect of variation of the ΔG_{fold} for each monomer on the percentage of HET (C), percentage of the HM BB (D), the total activity (E), and fitness (F) of the system. Fitness is estimated using a lognormal distribution with the optimum (alpha value) set to a total activity of 80 (see Figure 1B).

G-J) Effect of variation of $\Delta G_{bind,HET\ AB}$ and $\Delta G_{bind,HM\ BB}$ on the percentage of HET (G), percentage of the HM BB (H), and the total activity (I), and fitness (J) of the system. For panels G-J, $\Delta G_{bind,HM\ AA}$ was kept constant at -10 kcal /mol.

We then studied the effects of changes to binding affinity in their typical range of values, centered around -10 kcal/mol (Choi *et al*, 2015; Jankauskaitė *et al*, 2018). Figures 1G-J show the effect of variation in the binding affinities of the heterodimer and one of the homodimers ($\Delta G_{bind,HET\ AB}$ and $\Delta G_{bind,HM\ BB}$), while keeping the binding affinity of the other homodimer ($\Delta G_{bind,HM\ AA}$) constant at -10 kcal/mol to represent the data in two dimensions. As expected, the heterodimer is enriched if it becomes the complex with the strongest binding affinity and depleted if it has the weakest binding affinity (Figure 1G-H, Figure S1A-D). Interestingly, there are regions in the solution space in which most of the available subunits assemble into homodimers although one of the homodimers has a slightly weaker binding affinity than the heterodimer (Figure S1E). In this case, the available copies of one of the proteins would be depleted by fast assembly of their homodimer, so the copies of the second protein would assemble into their respective homodimers. On the contrary, heterodimers are enriched when their binding affinity is slightly weaker than that of one of the homodimers but much stronger than that of the second homodimer (Figure S1F). Here, any exchange of subunits between the heterodimer and the stronger homodimer would release subunits of the weaker homodimer, leading to an imbalance of concentrations of free subunits. Thus, equilibrium would favor strong heterodimers over a mixture of strong and weak homodimers. Importantly, in the range of values tested and assuming the specific activities of homodimers and heterodimers are identical, the total activity and fitness of the system remain unchanged despite shifts in the concentration of each dimer (Figures 1I-J). Overall, the post-duplication equilibrium is highly sensitive to changes in binding affinity when all other parameters remain constant: there is a narrow region where the concentrations of homodimers and heterodimers stay close to the original 1AA:2AB:1BB ratios, but small changes can lead to drastic changes in the concentration of each dimer.

The solution space of the system of equations shows that directly after duplication, the equilibrium concentrations of each dimer are particularly sensitive to mutational effects on binding affinity. Differences between the binding affinity of homodimers and heterodimers in the range of 0.5 - 1.0 kcal/mol are sufficient to drastically alter the representation of the heteromer in the system (Figure 1G). This is supported by empirical observations. Studies show that homomeric specificity can be achieved with just 1-3 substitutions (Ashenberg *et al*, 2011; Emlaw *et al*, 2021) and marginal differences in binding affinity (Hochberg *et al*, 2018). On the other hand, there are regions of the solution space that are much less sensitive to these mutational effects, for example, when the difference in ΔG_{bind} between the homodimers and the

heterodimer is already very high. These results are in agreement with the models on the energetic cost of homomerization derived by Hochberg et al. (2018) using a different approach. An important consideration is that these changes in binding affinity alter the representation of each dimer in the system, but do not modify the total concentration of complexes, as seen in Figure 1I. Thus, there would be many ways for the cell to provide the same total concentration of complexes, all of which could have the same fitness if every dimer has the same specific activity (Figure 1J). While the landscapes presented on Figure 1 help visualize the effect of changes to one parameter at a time, the fate of modern paralogs (Marchant *et al*, 2019; Mallik & Tawfik, 2020) is a product of the co-evolution and correlation of many parameters (synthesis, folding, binding, activity, etc.). Simulating evolution of these landscapes therefore needs to be based on the actual distributions of effects, which can vary from one protein to another. Importantly, these observations show that the system could theoretically deviate strongly from the starting relative concentrations while keeping the same overall performance and this, in a directional manner, if mutations have biased effects in one direction or the other.

Evolutionary simulations show a general bias toward heterodimers replacing the ancestral homodimer

We used the above model to simulate the evolution of the post-duplication equilibrium between homodimers and heterodimers as mutations accumulate (Figure 2A). Mutations lead to changes in the protein sequence, which can affect the binding affinity and the folding energy of proteins. To derive the distributions of mutational effects, we sampled a set of 104 homodimers with crystal structures from the Protein Data Bank (Berman *et al*, 2000) and simulated the effects of all possible single amino acid substitutions on folding energy and binding affinity with FoldX (Delgado *et al*, 2019) (see Methods). Briefly, entries from the PDB were chosen if their structures were solved by X-ray crystallography, annotated as homomers, having exactly the same or no missing residues, and with sequence lengths between 60 and 450 residues. From the resulting set of structures, we selected entries from the major ECOD architecture types (Cheng *et al*, 2014) so that we would have chosen at least seven for each of them. Similarly, we clustered these proteins at 40% sequence identity with CD-Hit (Li & Godzik, 2006; Fu *et al*, 2012), and selected one representative protein from each cluster.

We used our model and the dataset of PDB structures to simulate the post-duplication evolution of pairs of paralogs. We considered two different regimes for our simulations: parametric, in which we sampled mutational effects from multivariate normal distributions with different parameters based on the pooled structural data; and single structure, in which mutational effects are sampled directly from the distribution of effects of the corresponding structure. Since mutations can affect binding affinity and folding energy at the same time, the multivariate normal distributions used in our parametric simulations maintain the correlations observed on aggregate for the data with structures, namely the correlations between $\Delta\Delta G_{bind, HET}$ and $\Delta\Delta G_{bind, HM}$ (r = 0.9) (Figure 2B), between $\Delta\Delta G_{bind, HET}$ and $\Delta\Delta G_{fold}$ (r = 0.3) (Figure 2C), and between $\Delta\Delta G_{bind, HM}$ and $\Delta\Delta G_{fold}$ (r = 0.3) (Figure 2D). We annotated figure panels for results for parametric simulations with a computer cartoon and results for single structure simulations with a protein dimer cartoon throughout the rest of the text.

Each step of the simulation represents a mutational event. Importantly, the underlying distribution from which the effects of mutations are sampled is kept constant. As discussed later, we apply this simplifying assumption to keep the simulations scalable and because, while the effects of new individual mutations are contingent on previous ones, it is unclear whether the shape of the global distribution changes. Once a mutational effect is sampled, we update all relevant binding affinities and folding energies and recalculate the concentrations at equilibrium for monomers and dimers. Next, we use the equilibrium concentrations to obtain the total activity of the system. Fitness is then estimated using a lognormal function given by two parameters: alpha (the value of total activity that maximizes fitness) and beta (the fitness corresponding to activity values of half and twice the value of alpha, beta was set to 0.5 for all simulations). We tested our simulations with two different values for alpha to consider two different scenarios: one in which the total activity after the duplication (76.4 h^{-1}) is very close to the optimal value (alpha = 80 h^{-1}), and one in which it overshoots the optimal value (alpha = 60 h^{-1}). Finally, fitness is used to calculate a probability of fixation for that mutation. Once a mutational effect fixes or is rejected, the simulation continues with the next sampling step. Each simulation was run with 50 replicates and allowed to continue until 200 mutations were fixed. Although some of the proteins tested are shorter than 200 residues, protein sequences of the two paralogs tended to keep around 50% sequence identity with the WT sequence and more than 30% sequence identity with one another by the end of the simulations. These measures are consistent with real-world scenarios since more than half of all *S. cerevisiae* and *E. coli* paralogous pairs also exhibit ≥30% identity with each-other (Mallik & Tawfik, 2020). Since the results for the two scenarios (alpha = 60 h^{-1}, alpha = 80 h^{-1}) generally agreed, the results with alpha = 80 h^{-1} are shown in the main text and those with alpha = 60 h^{-1} are shown in the supplementary materials.

We tested how changes in the parameters of the distribution of mutational effects of the homodimer affected the results of the simulation (Figure 2E, Figure S2). We allowed the parameters for the distribution of effects on $\Delta\Delta G_{bind, HM}$ to change while the distribution for $\Delta\Delta G_{bind, HET}$ was given by N(0.2, 1.2) and the distribution for $\Delta\Delta G_{fold}$ was given by N(2.6, 4.6). As expected, homodimers dominate when the average mutation is more destabilizing for the heterodimer than for the homodimer. Homodimers also dominate when the variance in mutational effects on $\Delta\Delta G_{bind, HM}$ is high, since this allows for extreme stabilizing effects on the homodimer. High variance also allows extreme destabilizing effects to be sampled often but they would be likely eliminated by selection because they also affect the total concentration of the complexes. Interestingly, simulations with parameters along the diagonal of Figure 2E tend to maintain both homodimers and heterodimers, with two conditions of particular interest indicated with a circle and a triangle. The circle indicates the simulation in which mutational effects are the same for homodimers and heterodimers, while the triangle represents the intuitive expectation of mutational effects on the homodimer doubling those on the heterodimer. The preservation of the mixture of homodimers and heterodimers in these two conditions suggests that the mere doubling of mutational effects on the homodimer does not lead to an increase in the concentration of heterodimers, as long as effects are sampled from correlated ideal normal distributions. Thus, our parametric simulations show that when none of the dimers

offers an advantage over the others, biases in the distribution of available mutational effects alone can determine which dimer dominates at the end of the simulation.

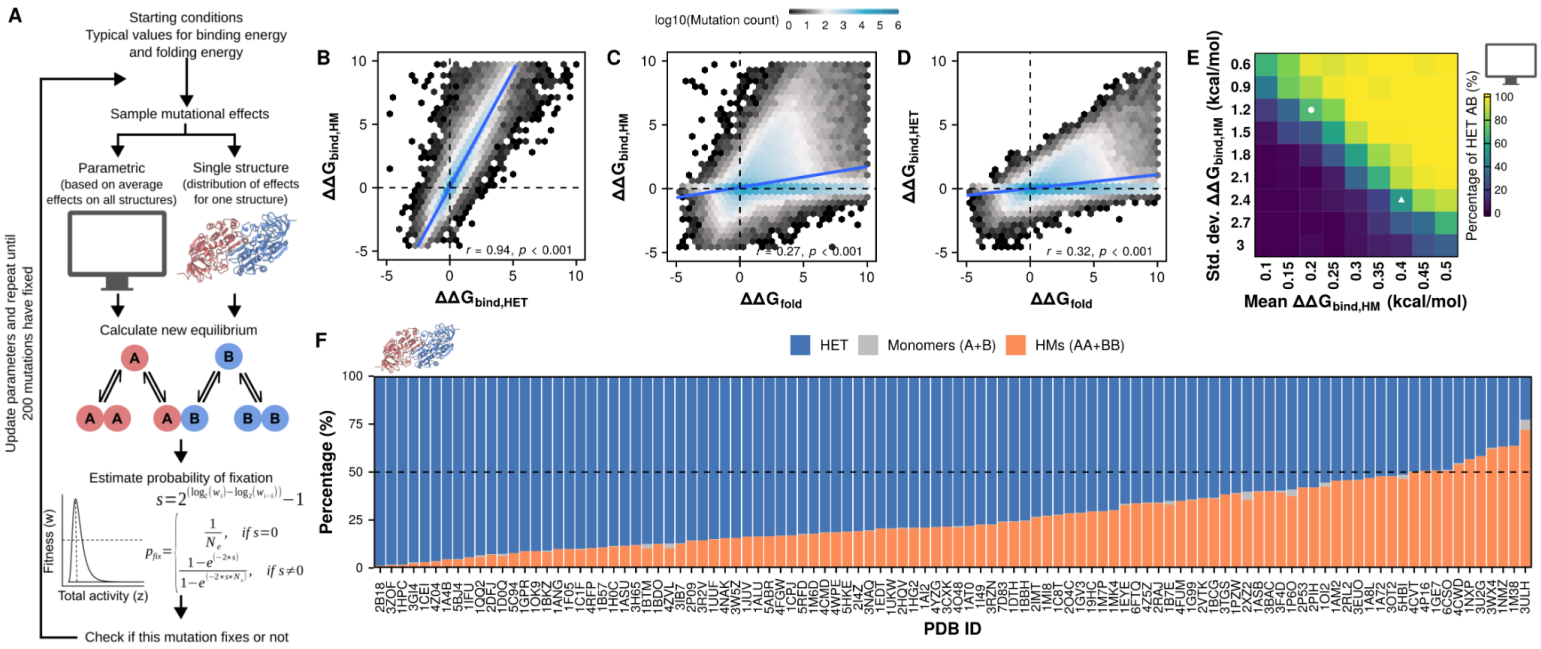

Figure 2. Biases in the distribution of effects on mutations on binding affinity can drive enrichments of homomers or heteromers.

A) Overview of the simulation workflow. Briefly, homodimers were sampled from the PDB and the effects of all possible mutations on their structures were simulated using FoldX. We use two different regimes of simulations: parametric, in which mutational effects are sampled from a multivariate normal distribution derived from the pooled data for all structures; and single structure, in which mutational effects are sampled directly from the available effects on a single structure. Once a mutation is sampled, we update the folding energy and binding affinities in the system to recalculate the new concentrations at equilibrium. These new concentrations then are used to infer fitness and the probability of fixation for the sampled mutation. The simulation continues until 200 mutations have been fixed. The structure for 1A72, used as the cartoon for the single structure simulations, was visualized with ChimeraX (Pettersen *et al*, 2021).

B-D) Distributions of mutational effects on $\Delta G_{bind,HET}$, $\Delta G_{bind,HM}$, and ΔG_{fold}. Hexagonal bins indicate the density of mutational effects for the pool of mutations for all structures. Two variables and their correlations are shown in each panel: $\Delta G_{bind,HET}$ and $\Delta G_{bind,HM}$ (B), $\Delta G_{bind,HM}$ and ΔG_{fold} (C), $\Delta G_{bind,HET}$ and ΔG_{fold} (D).

E) Percentage of heteromers at the end of the simulations with each set of parameters for the homodimers (average of 50 replicates, 200 mutations fixed for each). The white circle indicates the condition in which $\Delta\Delta G_{bind,HET}$ and $\Delta\Delta G_{bind,HM}$ are equally distributed (N(0.2, 1.2)). The white triangle indicates the condition in which the distribution for $\Delta\Delta G_{bind,HM}$ uses the same parameters calculated when considering all the sampled structures (N(0.4, 2.4)).

F) Average percentages of heterodimers, homodimers, and monomers for each of the PDB structures at the end of the simulations. The dashed line indicates the starting point at 50% heterodimers and 50% homodimers (25% of each homodimer).

We next set out to test our simulation system with the direct distributions of mutational effects

for each of the PDB structures sampled. Simulations were run with the estimated distributions of mutational effects for 104 structures of homodimeric proteins available in the Protein Data Bank. While replicates of the same simulation followed different trajectories due to the inherently random sampling, we observed that some structures seemed to favor particular kinds of complexes (Figure 2F, Figure S3). Strikingly, for most of the protein structures tested the average concentration of heterodimers at the end of the simulations is higher than the starting 50%. Indeed, for 66 structures (63.4%) the final concentration of heterodimers is higher than 70%, as opposed to only 1 structure (1%) for which the final concentration of homodimers is higher than 70%. For the remaining 37 structures (35.5%), the final concentrations stay closer to the initial equilibrium, with neither heterodimers nor homodimers enriched beyond 70%. The enrichment of heteromers tends to occur relatively early in the simulations, with some structures showing high heterodimer concentrations after fewer than 100 mutations had been fixed (Figure S3). This is consistent with our previous finding that the heteromerization also dominates in recently duplicated yeast paralogous pairs that have retained high sequence identity (Marchant *et al*, 2019; Mallik & Tawfik, 2020). This suggests there is a neutral bias toward the replacement of ancestral homodimers by heterodimers, although independent identical events might reach different outcomes.

In our previous simulations, the optimum value for total activity (alpha = 80 h^{-1}) was slightly higher than the total activity directly after the duplication. The results of simulations where the duplication overshoots the optimum (alpha = 60 h^{-1}) correlate quite well with them, although simulations with alpha = 60 h^{-1} consistently yielded slightly lower percentages of heterodimers (Figure S4A-B). Since the simulations with alpha = 60 h^{-1} needed to decrease the total activity of the system to approach the optimum, in these simulations at least one of the subunits is quickly destabilized (Figure S4C), which does not happen in the simulations with alpha = 80 h^{-1} (Figure S4D). Indeed, the quick destabilization observed in the simulations with alpha = 60 h^{-1} occurs at the same time as an initial decrease in the concentration of heterodimers since the available concentrations of free subunits would be different (Figure S4E). As the simulations advance, the concentrations of heterodimers increase, both when alpha = 60 h^{-1} and when alpha = 80 h^{-1} (Figure S4E-F). Thus, our simulations show the inherent bias toward the increase in the concentration of heterodimers is robust to the position of the fitness optimum.

We tested associations between structural parameters and the outcome of the simulations. First, we focused on the interfaces of our set of PDB structures. We looked at the relative size of the interface core (Figure S5A), the overall stickiness of the interface (Figure S5B), the proportion of interface core residues that are involved in homotypic contacts (Figure S5C), and the overall secondary structure of the interface (Figure S5D). Neither of these interface parameters showed a clear influence on the outcome of the simulations. We extended our secondary structure analysis to the ECOD architectures (Cheng *et al*, 2014) of our protein set (Figure S5E, Table S1). Outcomes of simulations with structures with similar ECOD architectures were highly variable, suggesting that structure alone is not enough to make an association with the bias toward heterodimers. Similarly, we observed that modifying the starting parameters of subunit folding energy and binding affinity tended to have little effect on the most frequent outcome of simulations with a given structure (Figure S6). These results suggest that

the outcome of the simulations depends on properties of the distribution of mutational effects that are not directly derived from these structural features.

Weak but constant mutational biases influence the outcome of the simulations

We reasoned that, although the distributions for all proteins in our set are very similar, they might differ with respect to the availability of outliers or pervasive biases with small magnitude in terms of folding or binding energies, which would then drive evolution in one particular direction. Considering that changes in ΔG_{fold} never lead to an increase in the concentration of heterodimers, we focused our analysis on the effects on ΔG_{bind}. For simplicity, we first compared the distributions of available and fixed mutational effects on $\Delta G_{bind,HET}$ and $\Delta G_{bind,HM}$ for the structures with the two most extreme outcomes (2B18 as HET-dominant and 3ULH as HM-dominant) (Figure 3A-B). As noted above, mutational effects on $\Delta G_{bind,HET}$ and $\Delta G_{bind,HM}$ are highly correlated for both structures, and fixed mutations were not restricted to any single part of the distribution. Thus, we hypothesized that proteins that are enriched for heterodimers in the simulations might have access to more mutations with slight deviations from the expected diagonal that favor the heterodimer, which would gradually accumulate as the simulation progresses. For example, a mutation that is below the diagonal in figure 3A-B would destabilize the heterodimer less (or stabilize it more) than expected based on the effect on the homodimer, whereas mutations above the diagonal would favor the homodimer. We calculated these deviations from the diagonal as follows:

$$residual = \Delta\Delta G_{bind,HET\,AB} - 0.5 * \Delta\Delta G_{bind,HM\,AA}$$

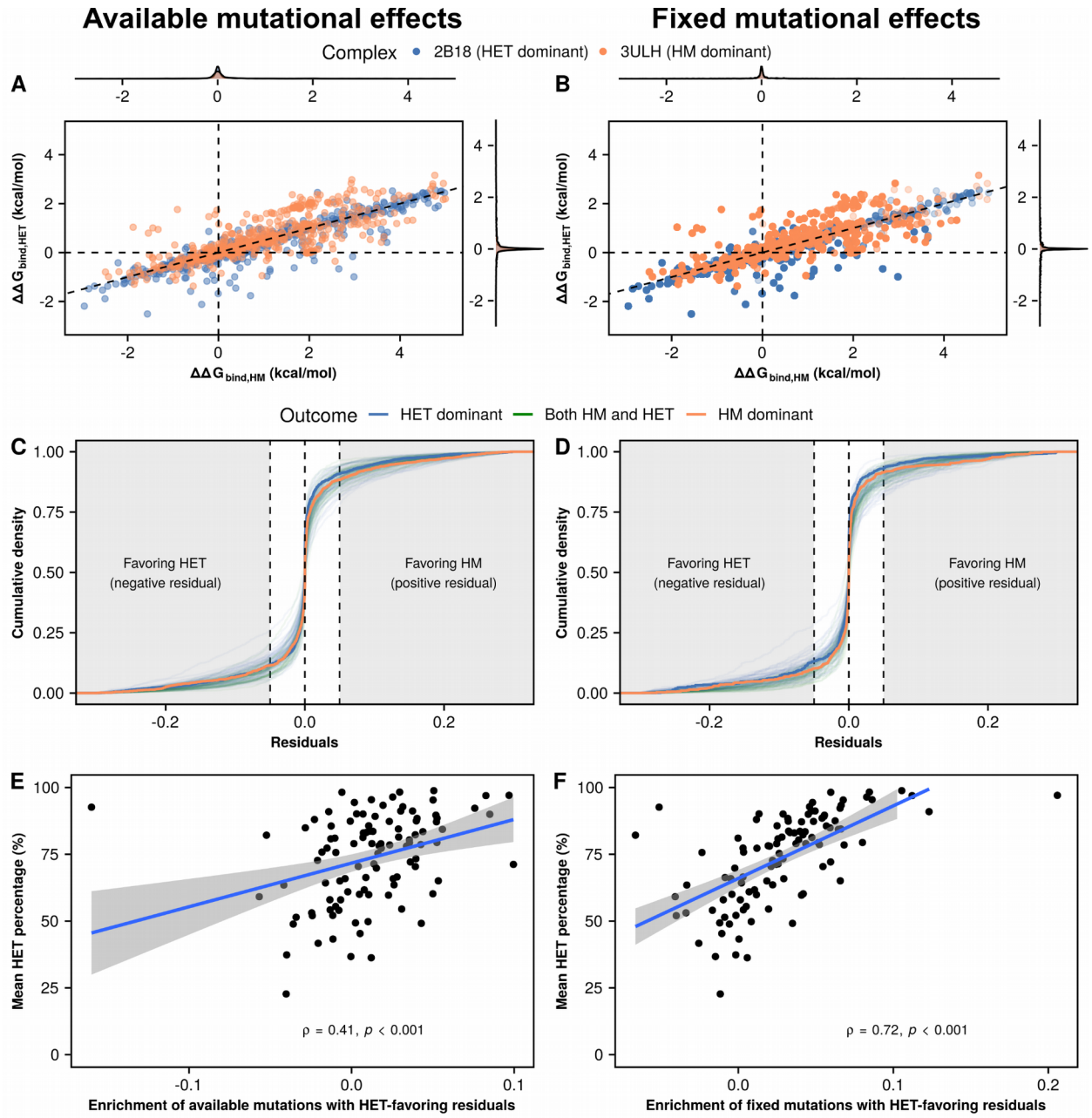

Figure 3. Slight deviations from the diagonal (residuals) are associated with the outcome of the simulations.

A-B) Distributions of available (A) and fixed (B) mutational effects on $\Delta G_{bind,HET}$ and $\Delta G_{bind,HM}$ for the two proteins with the most extreme outcomes in figure 2F: 2B18 (HET-dominant) and 3ULH (HM-dominant). The diagonal ($\Delta\Delta G_{bind,HET} = 0.5 * \Delta\Delta G_{bind,HM}$) is shown to illustrate the expectation of mutations on the heterodimer having half of the effect of mutations on the homodimer.

C-D) Distributions of available (C) and fixed (D) residuals with respect to the expectation for each of the 104 structures. Colors indicate mean outcomes for each structure: HET dominant (HET > 70% of complexes), HM dominant (HM > 70% of complexes), both HM and HET (rest of cases). The thicker blue and orange lines represent 2B18 and 3ULH, respectively. Shaded regions on the left and right indicate mutations whose residuals have a magnitude of at least 0.05 favoring either the heterodimer (negative values) or the homodimer (positive values).

E-F) Correlation between enrichment of available (E) and fixed (F) mutations with residuals favoring the heterodimer and the mean percentage of heterodimers at the end of the simulations. The enrichment of HET-favoring residuals is calculated as the density of mutations with residuals smaller than -0.05 minus the density of mutations with residuals greater than 0.05.

The cumulative distributions of residuals of available and fixed mutations show minor differences overall (Figure 3C-D). While comparing 2B18 (the most HET-dominant structure) to 3ULH (the most HM-dominant structure) reveals that 3ULH has a slightly higher density of mutations with positive (HM-favoring) residuals, the effect is not as clear for the rest of the structures. To have a finer resolution, we defined an enrichment of HET-favoring residuals as the percentage of residuals with magnitudes below -0.05 (HET-favoring) minus the percentage of residuals with magnitudes above 0.05 (HM-favoring). Structures with a higher enrichment of HET-favoring mutations are the ones with the highest concentrations of heterodimers at the end of the simulations (Figure 3E-F). We note that the available enrichment of HET-favoring mutations is amplified in the mutations that fix throughout the simulations. Indeed, the cumulative sum of residuals fixed up to a particular point in the simulations explains very well the observed percentage of heteromers (Figure S7). As a result, small but consistent mutational biases can lead to an increase in the concentration of the heterodimer, even without selection explicitly favoring it.

Differences in synthesis rate and specific activity can reverse the bias toward heterodimers

Our results so far show that the concentration of heterodimers relative to the homodimers after duplication is expected to increase even in the absence of selection favoring this complex over the homodimers. However, a large fraction of duplicated homodimeric proteins do not maintain their heterodimers. Other forces could thus counterbalance this inherent mutational trend. These include differences in compartmentalization, expression profiles, and specific activities. Paralogs that are separated into different compartments, such as yeast thioredoxin reductases TRR1 (cytosolic) and TRR2 (mitochondrial), do not interact *in vivo* despite still being capable of doing so *in vitro* (Oughtred *et al*, 2021; Mallik *et al*, 2022). For differentially expressed paralogs, such as E. coli LYSS (constitutive) and LYSU (heat-induced) (Brevet *et al*, 1995), equilibrium shifts toward the homodimer of the more highly abundant protein. Differences in the specific activity of homomers and heteromers would help selection distinguish between the two of them to maintain a particular total activity in the system.

In this section, we focus on the effects of differential synthesis rates and specific activities on the system. We do not model the case where one of the complexes could evolve a novel function, because it would require multiple additional considerations. For example, the system would need to describe and deconvolute mutational effects on each of the two functions. Also, the two functions could have different weights in the fitness calculations. Additionally, those considerations should be specific to each protein, unlike folding and binding which are common physical properties of all the dimers we consider.

We first used our equations to characterize the effect of changes in relative synthesis rates

between paralogs. We observed that the concentration of the heterodimer is maximized when synthesis rates are identical (Figure 4A). Conversely, the concentration of either homodimer is maximized as the synthesis rate of its subunits increases beyond that of its paralog (Figure 4B). The total activity of the system increases as the synthesis rate of one of the paralogs increases (Figure 4C), which initially leads to an increase in fitness as total activity approaches the optimum of our fitness function (alpha = 80 h^{-1}), but then decreases as total activity goes beyond it (Figure 4D). Since many paralogous proteins are differentially expressed (Gout *et al*, 2010; Gout & Lynch, 2015; Aubé *et al*, 2023), expression divergence could counteract mutational biases that favor heterodimers. However, the extent of this effect would depend on the relative frequencies of mutational effects on synthesis rates versus the coding sequence, as well as the fitness effects of changes in the total activity of the system due to the total amount of protein copies produced.

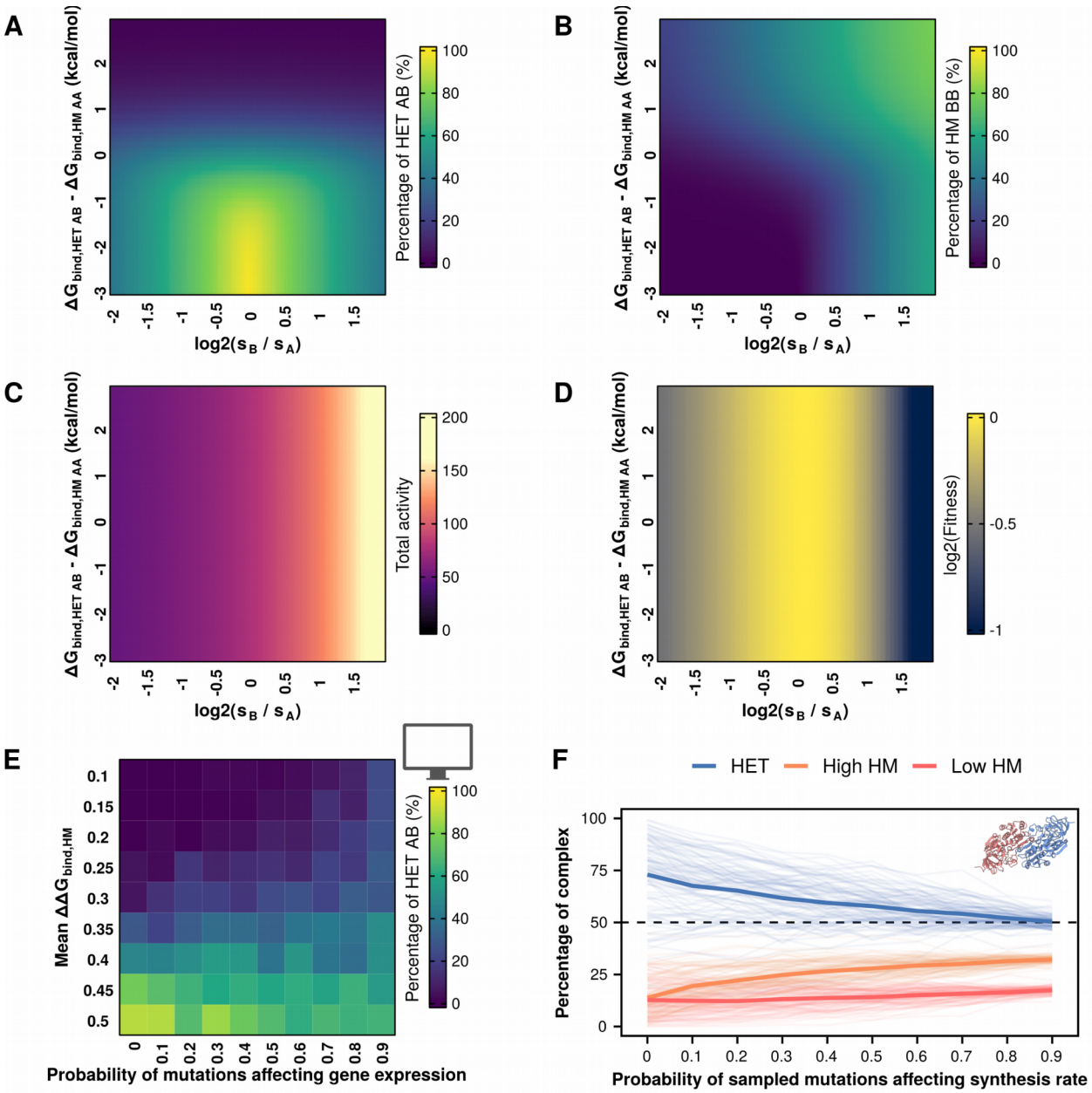

Figure 4. Allowing changes in synthesis rates can limit the bias toward heterodimers.

A-D) Percentage of heterodimers (A), one of the homodimers (B), total activity (C), and fitness (D) as a function of the ratio of synthesis rates and differences between $\Delta\Delta G_{bind,AB}$ and $\Delta\Delta G_{bind,AA}$. In panels A-D, s_A, $\Delta\Delta G_{bind,AA}$, and $\Delta\Delta G_{bind,BB}$ remain constant throughout. Changes in the y-axis indicate changes in $\Delta\Delta G_{bind,AB}$, with all three values ($\Delta\Delta G_{bind,AA}$, $\Delta\Delta G_{bind,AB}$, $\Delta\Delta G_{bind,BB}$) being identical at zero.

E) Percentages of heterodimers at the end of parametric simulations allowing changes in synthesis rates and varying mean effects of mutations on homodimers.

F) Percentages of heterodimers and each of the two homodimers for each of the 104 tested structures when simulations include different probabilities of mutations affecting synthesis rates. The two homodimers were distinguished based on their abundance, such that high HM refers to the one with the higher concentration and low

HM to the one with a lower concentration.

Mutations can affect synthesis rates in multiple ways. For example, they could affect binding of transcription factors to a promoter or translation rates (Hausser *et al*, 2019; Verma *et al*, 2019; Kemble *et al*, 2020; Roos & de Boer, 2021; Aubé *et al*, 2023). To include changes in synthesis rates in our simulations, we added an additional parameter representing the probability of mutations affecting synthesis rates (p_{exp}), focusing on mutations that occur in the promoter region. In these cases, we sampled the effect of mutations on synthesis rates from a separate skew-normal distribution (mean = 0, variance = 0.025, skew = -0.125) based on the effects of mutations on the TDH3 promoter on expression level (Metzger *et al*, 2016). Following the data from Metzger et al. (2016), these effects were considered to be multiplicative, i.e. the current synthesis rate at the corresponding step in the simulations would be multiplied by the value sampled from the distribution of mutational effects on synthesis rates.

Since promoter sequences can have widely variable rates of evolution (Young *et al*, 2015, 2022), we proceeded to repeat our simulations with different probabilities of mutational effects affecting synthesis rates. In these new parametric simulations, we observed that higher probabilities of mutations affecting synthesis rates generally led to less extreme outcomes in terms of HM or HET dominance (Figure 4E, Figure S8), even when mutational effects on binding affinity are skewed against either of them. At higher probabilities of mutations affecting synthesis rate, the concentration of the heterodimer remains in the range between 40 - 60% of the complexes. Similarly, the bias toward heterodimers observed for the simulations with the distributions of mutational effects of individual PDB structures became considerably weaker as the probability of mutations affecting synthesis rates increased (Figure 4F). We note that the synthesis rates of the two paralogs did not diverge considerably in the simulations (in the most extreme case, with mutations having p_{exp} = 0.9, median divergence in synthesis rates for the same replicate = 26.1 h^{-1}, interquartile range = [12.07 h^{-1}, 44.6 h^{-1}]) (Figure S9). We repeated these simulations setting the optimal activity value (alpha) to 60 h^{-1} so that the duplication would overshoot the optimum and obtained similar results, albeit with slightly lower concentrations of heterodimers (Figure S10).

Overall, allowing changes in synthesis rates tended to preserve both homodimers and heterodimers. This effect can be understood as two components. First, it opens the possibility of divergence in synthesis rates, which would naturally shift equilibrium away from the heterodimer and toward the homodimer of the more abundant protein (Figure 4A-B). Since the fitness of the system was sensitive to synthesis rates (Figure 4D), it is likely that an increase in the synthesis rate of one protein would be followed by a decrease in the synthesis rate of its paralog, even if these differences in synthesis rates are moderate (Figure S9). Second, because mutations in our simulations either affect synthesis rates or the protein sequence, accelerating the evolution of synthesis rates implied decelerating the evolution of the coding sequence and limiting the accumulation of mutational biases that led to the enrichment of the heterodimer. Our result that allowing changes in synthesis rates limits the bias toward heterodimers can help explain why extant paralog pairs often keep both homomers and heteromers (Marchant *et al*, 2019; Mallik & Tawfik, 2020), considering that the abundances of paralogs often diverge (Gout *et al*, 2010; Gout & Lynch, 2015; Aubé *et al*, 2023).

We then characterized the effect of changes in the specific activity of the homodimers or the heterodimers. In real biological systems, the expectation would be that these specific activities start out as being identical and then slowly diverge as mutations accumulate. However, since the relationship between sequence and function is complex and specific to each protein, we decided to use the simplifying assumption that specific activities would differ directly from the beginning. In these simulations, the more active dimer always has a specific activity of 1, whereas the less active dimer has a specific activity between 0.2 and 0.9. We define a HET activity bias as the difference in specific activities. For example, a -10% HET activity bias indicates the heterodimer is less active than the homodimers (specific activities of 0.9 for heterodimers and 1 for homodimers), whereas a 10% HET activity bias indicates the heterodimer is more active (specific activities of 1 for heterodimers and 0.9 for homodimers). As expected, changes in the specific activity do not affect the assembly of complexes (Figure 5A-B), since this is determined by binding affinities and the concentrations of properly folded subunits. However, they do have an impact on the total activity of the system (Figure 5C) and fitness (Figure 5D). Total activity is maximized when the most active dimer is the most abundant one (top left and bottom right corners of Figure 5C). Since the specific activity of homodimers or heterodimers never exceeds 1 with our definition of HET activity bias, fitness is maximized in the same regions where activity is maximized (Figure 5D), although the landscape would be different if we allowed values higher than 1. Thus, the solution space of the system of equations shows that allowing differences in specific activity can promote selection for either homodimers or heterodimers.

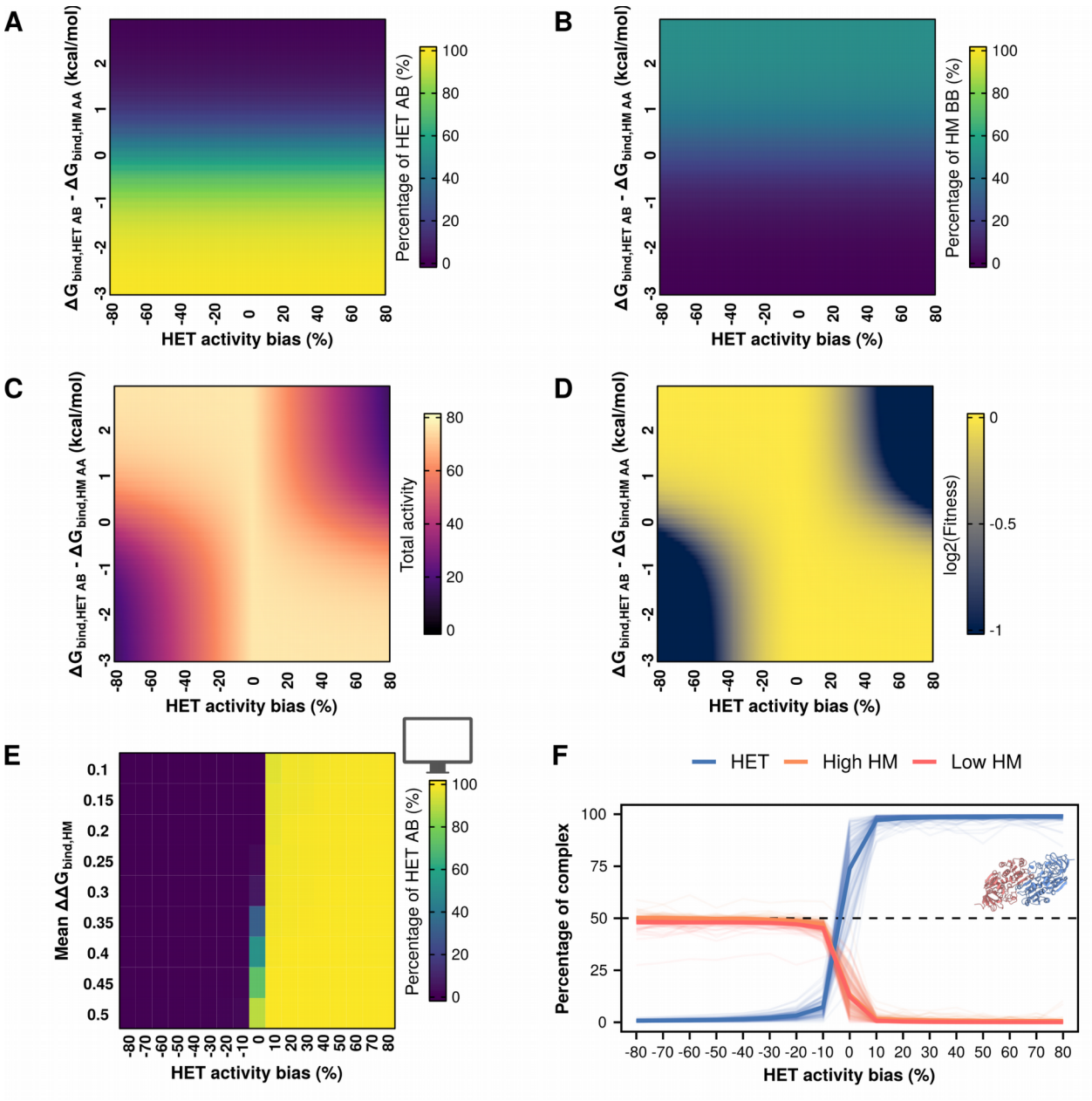

Figure 5. Differences in specific activities of homodimers and heterodimers can revert the observed bias towards heterodimers.

A-D) Percentage of heterodimers (A), one of the homodimers (B), total activity (C), and fitness (D) as a function of differences in specific activity and differences between between $\Delta\Delta G_{bind,HET\ AB}$ and $\Delta\Delta G_{bind,HM\ AA}$. In panels A-D, $\Delta\Delta G_{bind,HM\ AA}$ and $\Delta\Delta G_{bind,HM\ BB}$ remain constant throughout. Changes in the y-axis indicate changes in $\Delta\Delta G_{bind,AB}$, with all three values ($\Delta\Delta G_{bind,HM\ AA}$, $\Delta\Delta G_{bind,HET\ AB}$, $\Delta\Delta G_{bind,HM\ BB}$) being identical at zero.

E) Parametric simulations using differences in the specific activities of homomers and heteromers and varying mean effects of mutations on homomers. HET activity bias indicates a percentage decrease (negative values) or increase (positive values) of the HET activity with respect to the HM activity.

F) Percentages of heterodimers and each of the two homodimers for each of the 104 tested structures when

simulations consider differences in specific activity between homodimers and heterodimers. The two homodimers were distinguished based on their abundance, such that high HM refers to the one with the higher concentration and low HM to the one with a lower concentration.

We then tested the effect of allowing different specific activities for homodimers and heterodimers in our simulations of evolution. Our parametric simulations show that differences in specific activity can overturn the bias toward heterodimers when the optimum is set to 80 h^{-1} (Figure 5E, Figure S11). We observed a similar pattern in our simulations with structures (Figure 5F). Large differences in specific activity lead to 100% homodimers or 100% heterodimers, depending on the complex with higher specific activity. However, when these differences become very small, the system seems slightly less efficient at eliminating the heterodimers. When the heterodimer is 10% less active than the homodimer, many structures still maintain a small fraction of heterodimers, whereas homodimers are essentially eliminated when the heterodimer is 10% more active. Interestingly, moving the optimal total activity from 80 h^{-1} to 60 h^{-1} led to important changes in the fitness landscape (Figure S12A) that altered drastically the results of the parametric simulations (Figure S12B). Whereas extreme biases in specific activity predictably led to the corresponding enrichment of homodimers or heterodimers depending on the more active dimer, HET activity values between -20% and 20% yielded more complex patterns. In this range of HET activity bias values, the simulations can approach the optimum activity value of 60 h^{-1} either by keeping a low total concentration of highly active complexes or by keeping a high total concentration of less active complexes, with the likelihood of each scenario depending on the distribution of mutational effects. Indeed, we see both scenarios in our simulations with single structures (Figure S12C). However, when looking at the global trends for our simulations with structures the heteromer is still enriched even when it is around 20% less active than the homodimers (Figure S12D) since their distributions have more destabilizing effects on the homomers (see white triangle in Figure 2E) and slight mutational biases favoring the heterodimer (Figure 3). Thus, differences in specific activity can easily revert the bias toward heterodimers, but the most likely outcome would still be affected by the underlying distribution of mutational effects and the fitness function.

Discussion

There are many different potential fates for duplicate genes. These often involve divergence of the paralogs in terms of function, regulation, and interactions with other proteins (Conant & Wolfe, 2008). Much work has been dedicated to describing the different outcomes (Hochberg *et al*, 2018; Marchant *et al*, 2019; Mallik & Tawfik, 2020) and developing models of the fate of duplicate genes (Force *et al*, 1999; Lynch & Conery, 2000; Gout & Lynch, 2015; Johri *et al*, 2022). However, we lack a thorough understanding of the underlying molecular determinants and evolutionary forces. The way mutations themselves can drive evolution through their biases towards specific phenotypic outcomes could play a role. For instance, subfunctionalization is more likely to occur than neofunctionalization because loss-of-function mutations are more likely than beneficial gains of function (Lynch & Force, 2000; Gibson & Goldberg, 2009). This alone could naturally bias the maintenance of gene duplicates in one direction even in the absence of natural selection favoring one outcome or the other. Here, we examine a special case of subfunctionalization which is the potential replacement of a homodimer by a heterodimer following gene duplication.

We characterized the equilibrium between homodimers and heterodimers after duplication. Small effects on binding affinity in the range of 0.5 - 1 kcal / mol can drastically alter the proportions of homodimers and heterodimers. Since our estimations of the distributions of mutational effects are centered close to that range (averages of 0.2 kcal / mol for heterodimers and 0.4 for homodimers), mutations that could shift the equilibrium are likely to be sampled often by evolution. Furthermore, although mutational effects on binding for homomers and heteromers are highly correlated, our FoldX predictions suggest that there is a small percentage of mutations that could directly stabilize one complex while destabilizing the other (Figure 2B). Our model shows that changes in the proportion of the three complexes could therefore arise through a few mutations, matching previous reports of obligate homomers (Ashenberg *et al*, 2011; Hochberg *et al*, 2018) and heteromers (Emlaw *et al*, 2021), in which the dominant dimer is only marginally more stable than the other.

In our simulated evolutionary trajectories, we observe that the most common outcome is for the proportion of heterodimers to increase. This result extends our previous observation that heterodimers of paralogs are often retained when only the homodimers are selected for (Marchant *et al*, 2019). However, a key difference is that in our previous study we consider selection acting on the ΔG_{bind} of one of the complexes, whereas here we consider selection acting on the total activity of the system. As such, the relative concentrations of homodimers and heterodimers are allowed to drift as long as the total activity is maintained. The main implication of our results is that heterodimers of paralogs are likely to become the major functional unit over the ancestral homodimeric protein unless the heterodimer is selected against. Overall, the dominance of heteromers in out simulations is in agreement with previous analyses of the homo and heteromeric fates of paralogs in high-throughput PPI data (Mallik & Tawfik, 2020). However, high-throughput PPI data are typically derived without considering

barriers against heteromerization (such as differential compartmentalization and expression) and the relative abundance of each interaction. As such, further experimental characterization of the fates of paralogs would require methods that address these issues.

The case for selection against heteromers might differ among protein families. For example, selection against heteromers of paralogs has been pointed out as a way to reduce cross-talk and promote specialization of transcription factors (Amoutzias *et al*, 2008), kinases (Ashenberg *et al*, 2011), and chaperones (Pareek *et al*, 2011). However, differences in specificity in transcription factors can also be achieved if one of the homomers and the heteromer are retained (Mamnun *et al*, 2002; Baker *et al*, 2013). Others have reported positive selection for heteromerization of transcription factors (Hernández-Hernández *et al*, 2007), as well as allosteric (Kafková *et al*, 2017) and cooperative (Pillai *et al*, 2020) contributions of heteromerization, which would act to preserve the heteromer. Interestingly, Pils and Schulz (2004) have suggested that catalytically inactive protein domains are common (between 10% and 15%) in eukaryotic proteomes. These catalytically inactive proteins are often conserved and seem to have evolved regulatory functions mediated by binding to their paralogs (Pils & Schultz, 2004; Adrain & Freeman, 2012), providing a niche to preserve these heteromers. A final consideration about the difficulty of removing heteromers is that the starting concentrations immediately after duplication become more skewed towards heteromers as the number of subunits in the oligomer increases. Because of this skew, higher order complexes require a greater difference in the binding affinities of homomers and heteromers to promote specificity for homomerization (Hochberg *et al*, 2018). Thus, it is unclear whether there would be selection against most heteromers of paralogs and if it could be efficient enough to eliminate them.

Our results add to a growing body of work that suggests that complexity can increase neutrally (Stoltzfus, 1999; Lynch, 2007; Hochberg *et al*, 2020; Muñoz-Gómez *et al*, 2021; Schulz *et al*, 2022; Abrusán & Foguet, 2023). In the case of heteromers of paralogs, two proteins would become necessary to provide the function of an ancestral homomeric protein. Such transitions increase the complexity of protein interaction networks, and they do not necessarily provide an adaptive benefit. A feature that can only be introduced by heteromers is mutational asymmetry, which allows distinct sequences for the different subunits. Mutants that would otherwise eliminate catalytic activity in a homomeric assembly can participate in heteromers with native-like activity (Ebert *et al*, 2015). Indeed, the distributions we obtained when simulating mutational effects *in silico* had smaller effects on heterodimers than on homodimers and often had small enrichments in HET-favoring mutations (Figure 3E-F). As a result, their fixation could be a case of “survival of the flattest” due to smaller effects of point mutations (Wilke *et al*, 2001). However, this comes with increases in network complexity that do not contribute to robustness (Diss *et al*, 2017; Dandage & Landry, 2019). Overall, non-adaptive processes play a significant role in the abundance of heteromers of paralogs in extant protein interaction networks, although there are known examples of heteromers that provide an adaptive benefit (Bridgham *et al*, 2008; Pillai *et al*, 2020).

Importantly, our model makes multiple simplifying assumptions. In order to keep the simulations scalable, the distribution of mutational effects on folding energy and binding affinity is kept

constant throughout, instead of being recalculated at every step. We make this assumption also because while it is well-known that the effects of each new mutation are contingent on previous mutations, whether the overall shape of the distribution stays the same is unclear. Similarly, any systematic errors in our FoldX predictions would propagate as we simulate more mutations, making the estimations progressively less accurate. However, given that we consider the heterogeneity of the effects of mutations among the starting complexes, it is likely correct to assume that the distribution of effects of mutations through time is captured in this initial heterogeneity.

We do not consider the potential effects of other covariates that would prevent interactions between paralogs, such as compartmentalization (Marchant *et al*, 2019) and cotranslational assembly (Badonyi & Marsh, 2022). In these cases, the interaction between the two paralogs would not be possible, and the system would be expected to preserve the homomers. Our model also does not consider that tight binding could provide cross-stabilizing contributions to protein folding, such as the ones modeled in Rotem et al. (2018), although the fraction of properly folded subunits in our simulations consistently stays above 90% in most of our simulations. Finally, as mutations accumulate, one might expect the specific activity of homomers and heteromers to evolve. However, the distribution of mutational effects on specific activity would be unique for each protein and without more knowledge on the distribution of mutational effects on specific functions, we could make arbitrary choices that could bring the balance in one direction or another. Relaxing our assumptions and including more cases would allow our model to more accurately represent the fate of the paralogs, but our model already shows the complex interplay between synthesis rates, folding energies, binding affinity, and specific activities.

Our results highlight the importance of exploring neutral hypotheses for the evolution of molecular systems. While clearly natural selection plays a role in shaping them, there appears to be an important role of neutral evolution and mutational biases as well. Our result that heteromers of paralogs are likely to replace ancestral homomers after duplication agrees with recent observations that heteromers are more abundant in eukaryotic proteomes than in prokaryotes (Mallik & Tawfik, 2020; Schweke *et al*, 2023). This is consistent with the higher population sizes of prokaryotes (Lynch, 2006; Charlesworth, 2009), since any kind of selection against heteromers would be more efficient in these species than in eukaryotes. Future work should continue to investigate the factors leading to the fate of paralogs, focusing on how both natural selection and neutral evolution have contributed to the modern levels of complexity of protein interaction networks.

Materials and Methods

1.- Protein structures

1.1.- Selection of protein structures

The complete set of structures deposited on the Protein Data Bank (Berman *et al*, 2000) were downloaded on March 29th, 2021. Structures were then filtered based on the following criteria:

- Experimental data: X-ray crystallography
- Author-assigned biological assembly: Dimeric
- Structures must have two identical subunits (no missing residues or both subunits miss the same residues)
- Sequence length was restricted to 60 - 450 residues

We mapped the remaining set of structures to their ECOD architecture annotations (Cheng *et al*, 2014). We selected structures from the major architecture types such that we had at least 7 of each one. Our final set comprised 104 PDB structures. These structures were clustered with CD-Hit version 4.8.1 (Li & Godzik, 2006; Fu *et al*, 2012) with the following parameters: -n 2 (word size), -c 0.4 (sequence identity threshold), -G 1 (global alignment), and -s 0.8 (length difference cutoff). All of our selected structures were identified in different clusters, confirming low sequence identity between them.

1.2.- Structural analyses

We studied different structural parameters for each protein structure (Figure S4). Secondary structures for each residue of each structure were annotated with DSSP version 2.2.1 (Kabsch & Sander, 1983). Interfaces were identified following Tsai et al. (1996). Briefly, residues were considered to be at the interface core if the distance between one of their non-hydrogen atoms and a non-hydrogen atom from any residue of the other subunit was smaller than the sum of their van der Waals radii plus 0.5 Angstroms. Residues within 6 Angstroms of the interface core but not in contact with the other subunit were classified as the interface rim. We analyzed the proportion of residues in contact with their counterparts from the other subunit as a proxy for symmetry at the interface. For this analysis, we used a cut-off distance of 4 Angstroms between any non-hydrogen atoms of the two residues on the two chains.

2.- Estimation of distribution of mutational effects with FoldX

The biological assemblies of selected PDB structures were generated using custom scripts (Marchant *et al*, 2019). Energy minimization was performed on the biological assemblies using the FoldX Repair function (Delgado *et al*, 2019) 10 times to ensure convergence (Usmanova *et al*, 2018). Repaired structures were then used to perform *in silico* mutagenesis using the FoldX BuildModel and AnalyseComplex functions (Delgado *et al*, 2019) with the MutateX workflow (Tiberti *et al*, 2022) to estimate mutational effects on binding affinity and folding energy. Mutations were simulated in two runs for each complex: one run in which mutations were applied on both subunits (simulating mutational effects for homomers) and one run in which mutations were applied on only one subunit (simulating mutational effects for heteromers). Mutational effects on binding affinity were taken from the corresponding distributions, while mutational effects on folding energy were taken from the distribution of effects on the heteromer since our model uses changes on folding energy to calculate the fraction of properly folded subunits available for complex assembly.

3.- Framework for simulations

3.1.- Pre-duplication model

Our simulations consider a system of a gene encoding a protein (A) that forms a homodimer (AA). Protein copies are synthesized with synthesis rate s_A and typical values for starting parameters of folding energy (ΔG_{fold} = -5 kcal / mol) (Pace, 1975; Plaxco *et al*, 2000) and binding affinity (ΔG_{bind} = -10 kcal / mol) (Choi *et al*, 2015; Jankauskaitė *et al*, 2018). The proportion of properly folded subunits is calculated based on equation 1 (Sailer & Harms, 2017):

$$w = \frac{1}{1+e^{\Delta G_{fold}}} \qquad (1)$$

where w is the fraction of folded proteins and ΔG_{fold} is the free energy of the native fold.

Properly folded subunits are kept in the system and considered for the formation of complexes, while the remaining misfolded subunits are removed (assumed to be degraded/not to participate in the formation of complexes). Decay rates for properly folded monomers (d_A) and homodimers (d_{AA}) are set to 1.3 h^{-1}, following the derivation by Hausser et al (2019) based on median protein half-life and cell division rate for yeast. The assembly of homodimers is driven by the association constant k_{AA}. Thus, using s_A, w, d_A, d_{AA}, and k_{AA} we can calculate the expected concentrations at equilibrium of A and AA (see full derivation in Supplementary File 1). The equilibrium concentrations of monomers and homodimers are multiplied by their specific activity (0.1 for monomers, 1 for the homodimer) to calculate the total activity of the system. Fitness is determined from a lognormal function (defined by alpha and beta parameters) from the total activity of the system. Since the total activity for the starting system was calculated at around 40 units, the optimum fitness was set at 60 units (alpha parameter) to favor duplication events. The beta parameter was set to 0.5 so that fitness would be reduced by half when the total activity was either half or twice the alpha parameter.

Mutations are then sampled randomly. At the start of the simulation, each mutation has a probability of being a mutation on the coding sequence, a duplication, or a mutation causing a change in synthesis rate. The effects of coding mutations on folding energy ($\Delta\Delta G_{fold}$) and binding affinity ($\Delta\Delta G_{bind}$) are sampled from distributions of mutational effects (see section 3.3) and are used to calculate the fraction of properly folded subunits and the equilibrium constants, respectively. Equation 2 shows how the association constants for each dimer are calculated once ΔG_{bind} at the i-th step is updated with the effect of the sampled mutations ($\Delta\Delta G_{bind}$):

$$K_{AA} = e^{\frac{-(\Delta G_{bind,AA} + \Delta\Delta G_{bind,AA})}{RT}} \qquad (2)$$

where R is the gas constant (1.987 x 10^{-3} cal K^{-1} mol^{-1}) and T is temperature at 298 K.
Once a mutation has been sampled, the equilibrium concentrations of all molecular species and

the total activity of the system are recalculated using equations 3 and 4 (full derivation in Supplementary File 1):

$$c_A = \frac{-d_A + \sqrt{d_A^2 + 8\, d_{AA} k_{AA} s_A}}{4\, d_{AA} k_{AA}} \qquad (3)$$

$$c_{AA} = k_{AA} c_A^2 \qquad (4)$$

where c_A and the c_{AA} are the concentrations of each molecular species, d_A and d_{AA} are the decay rates of each molecular species, s_A is the rate of synthesis of properly folded subunits of A, and k_{AA} is the association constant for AA.

Using the new equilibrium concentrations, the fitness of the new state can be compared to that of the previous state. Simulations continue until 200 mutations have been fixed, out of which only one can be a duplication. Duplications introduce new molecular species: monomer B, homodimer BB, and heterodimer AB. As a result, the calculations for the equilibrium concentration of each complex change (see next section).

3.2.- Post-duplication system

Once a duplication has been fixed, the system will have two genes encoding two proteins (A and B) that will assemble into complexes AA, AB, and BB. Mutations continue to be sampled from the underlying distribution of mutational effects, and the system no longer samples for further duplication events. Once a mutational effect is selected, the affected protein (A or B) is selected at random. Afterwards, the new ΔG_{fold} is calculated to estimate the fraction of properly folded subunits and the ΔG_{bind} for both the homodimeric and heterodimeric interactions are used to estimate the new equilibrium constants. Following Hochberg et al. (2018), a factor of 2 is used to correct for the mixing entropy difference that inherently favors heteromers in the mixture of heterodimers and homodimers. Once all the values are calculated, the 4th degree polynomial in equation 5 is used to estimate the concentration at equilibrium of A (full derivation in Supplementary File 2):

$$
\begin{aligned}
0 &= \\
& 2 d_{AA} k_{AA} \left(4 \frac{d_{AA} k_{AA} d_{BB} k_{BB}}{d_{AB} k_{AB}} - d_{AB} k_{AB} \right) c_A^4 \\
& + \\
& \left(d_A \left(8 \frac{d_{AA} k_{AA} d_{BB} k_{BB}}{d_{AB} k_{AB}} - d_{AB} k_{AB} \right) - 2 d_B d_{AA} k_{AA} \right) c_A^3 \\
& + \\
& \left(\frac{2 d_{BB} k_{BB} d_A^2}{d_{AB} k_{AB}} + s_A \left(d_{AB} k_{AB} - 8 \frac{d_{AA} k_{AA} d_{BB} k_{BB}}{d_{AB} k_{AB}} \right) - d_A d_B - d_{AB} k_{AB} s_B \right) c_A^2 \\
& + \\
& s_A \left(d_B - 4 d_A \frac{d_{BB} k_{BB}}{d_{AB} k_{AB}} \right) c_A \\
& + \\
& 2 s_A^2 \frac{d_{BB} k_{BB}}{d_{AB} k_{AB}}
\end{aligned}
\tag{5}
$$

where dA, dB, dAA, dAB, and dBB are decay rates for each molecular species; sA and sB are the synthesis rates of properly folded subunits of A and B; kAA, kAB, and kBB are the association constants for each of the dimers.

Once the four potential solutions to the concentration of A are known, the concentrations of the rest of the molecular species are calculated using equations 6-9:

$$c_B = \frac{s_A / c_A - 2 d_{AA} k_{AA} c_A - d_A}{d_{AB} k_{AB}} \tag{6}$$

$$c_{AA} = k_{AA} c_A^2 = \frac{r_2}{d_{AA} + r_1} c_A^2 \tag{7}$$

$$c_{BB} = k_{BB} c_B^2 = \frac{r_4}{d_{BB} + r_3} c_B^2 \tag{8}$$

$$c_{AB} = k_{AB} c_A c_B = \frac{r_6}{d_{AB} + r_5} c_A c_B \tag{9}$$

where r_1, r_3, and r_5 are dissociation rates of AA, BB, and AB, respectively; r_2, r_4, and r_6 are association rates of AA, BB, and AB, respectively.

Among the four possible solutions to the complete system of equations, the physically correct solution is the one that results in positive values for the concentrations of all five molecular

species.

3.3.- Parametric distributions of mutational effects

We performed simulations by sampling from multivariate normal distributions of mutational effects. Each multivariate normal distribution was used to sample effects on ΔG_{fold} for each monomer, $\Delta G_{bind, AA}$ for the homodimer, and $\Delta G_{bind, AB}$ for the heteromer. We set the means of the multivariate normal distributions and the correlations between variables to values similar to those observed in the distributions simulated with FoldX:

- Effects on ΔG_{fold}: N(2.6, 4.6)
- Effects on $\Delta G_{bind, AB}$: N(0.2, 1.2)
- Effects on $\Delta G_{bind, AA}$: The mean was allowed to vary between 0.06 and 0.6, while the standard deviation varied between 0.6 and 3 (Figure 2B). For all other figures showing results of simulations with parametric distributions, the standard deviation was kept constant at 2.4.
- Correlation between effects on ΔG_{fold} and $\Delta G_{bind, AB}$: 0.3
- Correlation between effects on ΔG_{fold} and $\Delta G_{bind, AA}$: 0.3
- Correlation between effects on $\Delta G_{bind, AA}$ and $\Delta G_{bind, AB}$: 0.9

3.4.- Simulations allowing changes in gene expression

A subset of the simulations were run allowing mutations to have an effect on gene expression. For each sampled mutation, a given probability (ranging from 0.1 to 0.9) was given that a mutation would have an effect on synthesis rates instead of the coding sequence. The underlying distribution of mutational effects on gene expression was based on a fit to previously published work on the TDH3 promoter (Metzger *et al*, 2016). Briefly, the data were fit to a skew normal distribution, which gave the following parameters: mean = 0, standard deviation = 0.025, skew = -0.125. Since the mutational effects reported by Metzger et al. (2016) were normalized with respect to the WT, they were applied by multiplying the sampled values times the current synthesis rate.

3.5.- Simulations with biases in specific activity

A subset of the simulations were run to explore the outcomes when the homomers and heteromers had different specific activities. To bias a simulation towards homomers, the specific activity for homomers was kept at 1 while the specific activity of heteromers was given a value ranging from 0.1 to 0.9. Conversely, to bias a simulation towards heteromers, the specific activity for heteromers was kept at 1 while the specific activity of homomers was given a value ranging from 0.1 to 0.9.

4.- Data analysis

For each structure, we analyzed the results of 50 replicate simulations. Each replicate continues

until 200 mutations are fixed. We tracked the equilibrium concentrations of all molecular species (monomers A and B; homodimers AA and AB; and heteromer BB) at every time point. We define terms for the relative concentration of each molecular species at the end of the simulation according to equation 10:

$$p_X = \frac{100 * c_X}{c_A + c_B + c_{AA} + c_{AB} + c_{BB}} \qquad (10)$$

where p_X indicates the relative concentration of X (X being any of A, B, AA, AB, or BB) and c_X indicates the concentration of X.

We analyzed enrichments in homomers or heteromers by looking for deviations from the initial conditions post-duplication (25% AA, 50% AB, 25% BB). As such, replicates were classified based on the relative concentrations of each species at the end of the simulation:

- HET dominant: 70 <= pAB
- HM dominant: 70 <= (pAA + pBB)
- Both HM and HET: 70 <= (pAB + pAA + pBB) AND 70 >= pAB AND 70 >= (pAA + pBB)
- Monomers: 70 <= (pA + pB)
- Ambiguous: 70 >= (pAB + pAA + pBB) AND 70 >= (pA + pB)

We consider two measures of the result of the simulation: the concentration of each molecular species at the end of the simulation averaged over all replicates, and the percentage of replicates for each simulation that are classified in each of the preceding groups.

All scripts are available on Github (https://github.com/Landrylab/Homomer_duplication_2023) with supporting data on Zenodo (https://zenodo.org/records/10048861).

Attributions:

AFC, LNT, and CRL designed research. LNT derived the system of equations used in the model. LNT and AFC wrote the framework for simulations. AFC ran simulations and characterized the solution space for the system of equations. AFC, SM and CRL analyzed results. AFC and CRL wrote the paper with input from LNT and SM. EDL and CRL supervised research.

Acknowledgments:

We thank François Rouleau, Soham Dibyachintan, Adrian Serohijos, Sarah Otto, and members of the Landry and Levy labs for their feedback and help on the project.

Funding:

This work was supported by Canadian Institutes of Health Research (CIHR) Foundation grant 387697 (to C.R.L.), Canada Research Chair in Cellular Systems and Synthetic Biology (to C.R.L.), FRQNT Merit Scholarship Program for Foreign Studies (PBEEE) (to A.F.C.), joint funding from MEES and AMEXCID (to A.F.C.), Mitacs Globalink Research Award (to A. F. C.).

Conflict of interest:

The authors declare no conflict of interest.

References

Abrusán G & Foguet C (2023) An Assessment of Quaternary Structure Functionality in Homomer Protein Complexes. *Mol Biol Evol* 40

Adrain C & Freeman M (2012) New lives for old: evolution of pseudoenzyme function illustrated by iRhoms. *Nat Rev Mol Cell Biol* 13: 489–498

Amoutzias GD, Robertson DL, Van de Peer Y & Oliver SG (2008) Choose your partners: dimerization in eukaryotic transcription factors. *Trends Biochem Sci* 33: 220–229

Ascencio D, Ochoa S, Delaye L & DeLuna A (2017) Increased rates of protein evolution and asymmetric deceleration after the whole-genome duplication in yeasts. *BMC Evol Biol* 17

Ashenberg O, Rozen-Gagnon K, Laub MT & Keating AE (2011) Determinants of Homodimerization Specificity in Histidine Kinases. *J Mol Biol* 413: 222–235

Aubé S, Nielly-Thibault L & Landry CR (2023) Evolutionary trade-off and mutational bias could favor transcriptional over translational divergence within paralog pairs. *PLoS Genet* 19: e1010756

Badonyi M & Marsh JA (2023) Buffering of genetic dominance by allele-specific protein complex assembly. *Sci Adv* 9: eadf9845

Baker CR, Hanson-Smith V & Johnson AD (2013) Following gene duplication, paralog interference constrains transcriptional circuit evolution. *Science* 342: 104–108

Bergendahl LT & Marsh JA (2017) Functional determinants of protein assembly into homomeric complexes. *Sci Rep* 7: 4932

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN & Bourne PE (2000) The Protein Data Bank. *Nucleic Acids Res* 28: 235–242

Boncoeur E, Durmort C, Bernay B, Ebel C, Di Guilmi AM, Croizé J, Vernet T & Jault J-M (2012) PatA and PatB form a functional heterodimeric ABC multidrug efflux transporter responsible for the resistance of Streptococcus pneumoniae to fluoroquinolones. *Biochemistry* 51: 7755–7765

Brevet A, Chen J, Lévêque F, Blanquet S & Plateau P (1995) Comparison of the enzymatic properties of the two Escherichia coli lysyl-tRNA synthetase species. *J Biol Chem* 270: 14439–14444

Bridgham JT, Brown JE, Rodríguez-Marí A, Catchen JM & Thornton JW (2008) Evolution of a new function by degenerative mutation in cephalochordate steroid receptors. *PLoS Genet* 4: e1000191

Cano AV, Rozhoňová H, Stoltzfus A, McCandlish DM & Payne JL (2022) Mutation bias shapes the spectrum of adaptive substitutions. *Proc Natl Acad Sci U S A* 119

Charlesworth B (2009) Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nat Rev Genet* 10: 195–205

Cheng H, Schaeffer RD, Liao Y, Kinch LN, Pei J, Shi S, Kim B-H & Grishin NV (2014) ECOD: an evolutionary classification of protein domains. *PLoS Comput Biol* 10: e1003926

Choi J-M, Serohijos AWR, Murphy S, Lucarelli D, Lofranco LL, Feldman A & Shakhnovich EI (2015) Minimalistic predictor of protein binding energy: contribution of solvation factor to protein binding. *Biophys J* 108: 795–798

Conant GC & Wagner A (2003) Asymmetric sequence divergence of duplicate genes. *Genome Res* 13: 2052–2058

Conant GC & Wolfe KH (2008) Turning a hobby into a job: how duplicated genes find new functions. *Nat Rev Genet* 9: 938–950

Dandage R & Landry CR (2019) Paralog dependency indirectly affects the robustness of human cells. *Mol Syst Biol* 15: e8871

Delgado J, Radusky LG, Cianferoni D & Serrano L (2019) FoldX 5.0: working with RNA, small molecules and a new graphical interface. *Bioinformatics* 35: 4168–4169

Deng C, Cheng C-HC, Ye H, He X & Chen L (2010) Evolution of an antifreeze protein by neofunctionalization under escape from adaptive conflict. *Proc Natl Acad Sci U S A* 107: 21593–21598

Diss G, Gagnon-Arsenault I, Dion-Coté A-M, Vignaud H, Ascencio DI, Berger CM & Landry CR (2017) Gene duplication can impart fragility, not robustness, in the yeast protein interaction network. *Science* 355: 630–634

Ebert MCCJC, Maximilian C C J, Morley KL, Volpato JP, Schmitzer AR & Pelletier JN (2015) Asymmetric mutations in the tetrameric R67 dihydrofolate reductase reveal high tolerance to active-site substitutions. *Protein Science* 24: 495–507

Emlaw JR, Tessier CJG, McCluskey GD, McNulty MS, Sheikh Y, Burkett KM, Musgaard M & daCosta CJB (2021) A single historical substitution drives an increase in acetylcholine receptor complexity. *Proc Natl Acad Sci U S A* 118

Finnigan GC, Hanson-Smith V, Stevens TH & Thornton JW (2012) Evolution of increased complexity in a molecular machine. *Nature* 481: 360–364

Force A, Lynch M, Pickett FB, Amores A, Yan YL & Postlethwait J (1999) Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151: 1531–1545

Fu L, Niu B, Zhu Z, Wu S & Li W (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28: 3150–3152

Gibson TA & Goldberg DS (2009) Questioning the ubiquity of neofunctionalization. *PLoS Comput Biol* 5: e1000252

Gout J-F, Kahn D, Duret L & Paramecium Post-Genomics Consortium (2010) The relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution. *PLoS Genet* 6: e1000944

Gout J-F & Lynch M (2015) Maintenance and Loss of Duplicated Genes by Dosage Subfunctionalization. *Mol Biol Evol* 32: 2141–2148

Hausser J, Mayo A, Keren L & Alon U (2019) Central dogma rates and the trade-off between precision and economy in gene expression. *Nat Commun* 10: 68

Hernández-Hernández T, Martínez-Castilla LP & Alvarez-Buylla ER (2007) Functional diversification of B MADS-box homeotic regulators of flower development: Adaptive evolution in protein-protein interaction domains after major gene duplication events. *Mol Biol Evol* 24: 465–481

He X & Zhang J (2005) Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics* 169: 1157–1164

Hochberg GKA, Liu Y, Marklund EG, Metzger BPH, Laganowsky A & Thornton JW (2020) A hydrophobic ratchet entrenches molecular complexes. *Nature* 588: 503–508

Hochberg GKA, Shepherd DA, Marklund EG, Santhanagoplan I, Degiacomi MT, Laganowsky A, Allison TM, Basha E, Marty MT, Galpin MR, *et al* (2018) Structural principles that enable oligomeric small heat-shock protein paralogs to evolve distinct functions. *Science* 359: 930–935

Jankauskaitė J, Jiménez-García B, Dapkūnas J, Fernández-Recio J & Moal IH (2018) SKEMPI 2.0: an updated benchmark of changes in protein–protein binding energy, kinetics and thermodynamics upon mutation. *Bioinformatics* 35: 462–469

Johri P, Gout J-F, Doak TG & Lynch M (2022) A Population-Genetic Lens into the Process of Gene Loss Following Whole-Genome Duplication. *Mol Biol Evol* 39

Kabsch W & Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22: 2577–2637

Kafková L, Debler EW, Fisk JC, Jain K, Clarke SG & Read LK (2017) The Major Protein Arginine Methyltransferase in Trypanosoma brucei Functions as an Enzyme-Prozyme Complex. *J Biol Chem* 292: 2089–2100

Kemble H, Eisenhauer C, Couce A, Chapron A, Magnan M, Gautier G, Le Nagard H, Nghe P & Tenaillon O (2020) Flux, toxicity, and expression costs generate complex genetic interactions in a metabolic pathway. *Sci Adv* 6: eabb2236

Kidd RD, Baker HM, Mathews AJ, Brittain T & Baker EN (2001) Oligomerization and ligand binding in a homotetrameric hemoglobin: two high-resolution crystal structures of hemoglobin Bart's (gamma(4)), a marker for alpha-thalassemia. *Protein Sci* 10: 1739–1749

Kurtz A, Rollema HS & Bauer C (1981) Heterotropic interactions in monomeric βSH chains from human hemoglobin. *Arch Biochem Biophys* 210: 200–203

Kuzmin E, Taylor JS & Boone C (2022) Retention of duplicated genes in evolution. *Trends Genet* 38: 59–72

Levy ED & Teichmann S (2013) Structural, evolutionary, and assembly principles of protein oligomerization. In *Progress in Molecular Biology and Translational Science*, Giraldo J & Ciruela F (eds) pp 25–51. Academic Press

Li W & Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22: 1658–1659

Li Y, Zhang R, Wang C, Forouhar F, Clarke OB, Vorobiev S, Singh S, Montelione GT, Szyperski T, Xu Y, *et al* (2022) Oligomeric interactions maintain active-site structure in a noncooperative enzyme family. *EMBO J* 41: e108368

Lynch M (2006) Streamlining and simplification of microbial genome architecture. *Annu Rev Microbiol* 60: 327–349

Lynch M (2007) The frailty of adaptive hypotheses for the origins of organismal complexity. *Proc Natl Acad Sci U S A* 104 Suppl 1: 8597–8604

Lynch M (2012) The evolution of multimeric protein assemblages. *Mol Biol Evol* 29: 1353–1366

Lynch M (2013) Evolutionary diversification of the multimeric states of proteins. *Proc Natl Acad Sci U S A* 110: E2821–8

Lynch M & Conery JS (2000) The evolutionary fate and consequences of duplicate genes. *Science* 290: 1151–1155

Lynch M & Force A (2000) The probability of duplicate gene preservation by subfunctionalization. *Genetics* 154: 459–473

Mallik S & Tawfik DS (2020) Determining the interaction status and evolutionary fate of duplicated homomeric proteins. *PLoS Comput Biol* 16: e1008145

Mallik S, Tawfik DS & Levy ED (2022) How gene duplication diversifies the landscape of protein oligomeric state and function. *Curr Opin Genet Dev* 76: 101966

Mamnun YM, Pandjaitan R, Mahé Y, Delahodde A & Kuchler K (2002) The yeast zinc finger regulators Pdr1p and Pdr3p control pleiotropic drug resistance (PDR) as homo- and heterodimers in vivo. *Mol Microbiol* 46: 1429–1440

Marchant A, Cisneros AF, Dubé AK, Gagnon-Arsenault I, Ascencio D, Jain H, Aubé S, Eberlein C, Evans-Yamamoto D, Yachie N, *et al* (2019) The role of structural pleiotropy and regulatory evolution in the retention of heteromers of paralogs. *eLife* 8: e46754

Metzger BPH, Duveau F, Yuan DC, Tryban S, Yang B & Wittkopp PJ (2016) Contrasting Frequencies and Effects of cis- and trans-Regulatory Mutations Affecting Gene Expression. *Mol Biol Evol* 33: 1131–1146

Muñoz-Gómez SA, Bilolikar G, Wideman JG & Geiler-Samerotte K (2021) Constructive Neutral Evolution 20 Years Later. *J Mol Evol* 89: 172–182

Oughtred R, Rust J, Chang C, Breitkreutz B-J, Stark C, Willems A, Boucher L, Leung G, Kolas N, Zhang F, *et al* (2021) The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci* 30: 187–200

Pace CN (1975) The stability of globular proteins. *CRC Crit Rev Biochem* 3: 1–43

Pareek G, Samaddar M & D'Silva P (2011) Primary sequence that determines the functional overlap between mitochondrial heat shock protein 70 Ssc1 and Ssc3 of Saccharomyces cerevisiae. *J Biol Chem* 286: 19001–19013

Paul MD & Hristova K (2019) The RTK Interactome: Overview and Perspective on RTK Heterointeractions. *ACS Chem Rev* 119: 5881–5921

Pereira-Leal JB, Levy ED, Kamp C & Teichmann SA (2007) Evolution of protein complexes by duplication of homomeric interactions. *Genome Biol* 8: R51

Pettersen EF, Goddard TD, Huang CC, Meng EC, Couch GS, Croll TI, Morris JH & Ferrin TE (2021) UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci* 30: 70–82

Pillai AS, Chandler SA, Liu Y, Signore AV, Cortez-Romero CR, Benesch JLP, Laganowsky A, Storz JF, Hochberg GKA & Thornton JW (2020) Origin of complexity in haemoglobin evolution. *Nature* 581: 480–485

Pils B & Schultz J (2004) Inactive enzyme-homologues find new function in regulatory processes. *J Mol Biol* 340: 399–404

Plaxco KW, Simons KT, Ruczinski I & Baker D (2000) Topology, stability, sequence, and length: defining the determinants of two-state protein folding kinetics. *Biochemistry* 39: 11177–11183

Roos D & de Boer M (2021) Mutations in cis that affect mRNA synthesis, processing and translation. *Biochim Biophys Acta Mol Basis Dis* 1867: 166166

Rotem A, Serohijos AWR, Chang CB, Wolfe JT, Fischer AE, Mehoke TS, Zhang H, Tao Y, Lloyd Ung W, Choi J-M, *et al* (2018) Evolution on the Biophysical Fitness Landscape of an RNA Virus. *Mol Biol Evol* 35: 2390–2400

Sailer ZR & Harms MJ (2017) Molecular ensembles make evolution unpredictable. *Proc Natl Acad Sci U S A* 114: 11938–11943

Schulz L, Sendker FL & Hochberg GKA (2022) Non-adaptive complexity and biochemical function. *Curr Opin Struct Biol* 73: 102339

Schweke H, Levin T, Pacesa M, Goverde CA, Kumar P, Duhoo Y, Dornfeld LJ, Dubreuil B, Georgeon S, Ovchinnikov S, *et al* (2023) An atlas of protein homo-oligomerization across domains of life. *bioRxiv*: 2023.06.09.544317

Stoltzfus A (1999) On the possibility of constructive neutral evolution. *J Mol Evol* 49: 169–181

Svensson EI & Berger D (2019) The Role of Mutation Bias in Adaptive Evolution. *Trends Ecol Evol* 34: 422–434

Tiberti M, Terkelsen T, Degn K, Beltrame L, Cremers TC, da Piedade I, Di Marco M, Maiani E & Papaleo E (2022) MutateX: an automated pipeline for in silico saturation mutagenesis of protein structures and structural ensembles. *Brief Bioinform* 23: bbac074

Tsai CJ, Lin SL, Wolfson HJ & Nussinov R (1996) Protein-protein interfaces: architectures and interactions in protein-protein interfaces and in protein cores. Their similarities and differences. *Crit Rev Biochem Mol Biol* 31: 127–152

Usmanova DR, Bogatyreva NS, Bernad JA, Eremina AA, Gorshkova AA, Kanevskiy GM, Lonishin LR, Meister AV, Yakupova AG, Kondrashov FA, *et al* (2018) Self-consistency test reveals systematic bias in programs for prediction change of stability upon mutation. *Bioinformatics* 34: 3653–3658

Verma M, Choi J, Cottrell KA, Lavagnino Z, Thomas EN, Pavlovic-Djuranovic S, Szczesny P, Piston DW, Zaher HS, Puglisi JD, *et al* (2019) A short translational ramp determines the efficiency of protein synthesis. *Nat Commun* 10: 5774

Wang Z, Stalcup LD, Harvey BJ, Weber J, Chloupkova M, Dumont ME, Dean M & Urbatsch IL (2006) Purification and ATP hydrolysis of the putative cholesterol transporters ABCG5 and ABCG8. *Biochemistry* 45: 9929–9939

Wilke CO, Wang JL, Ofria C, Lenski RE & Adami C (2001) Evolution of digital organisms at high mutation rates leads to survival of the flattest. *Nature* 412: 331–333

Young RS, Hayashizaki Y, Andersson R, Sandelin A, Kawaji H, Itoh M, Lassmann T, Carninci P, FANTOM Consortium, Bickmore WA, *et al* (2015) The frequent evolutionary birth and death of functional promoters in mouse and human. *Genome Res* 25: 1546–1557

Young RS, Talmane L, Marion de Procé S & Taylor MS (2022) The contribution of evolutionarily volatile promoters to molecular phenotypes and human trait variation. *Genome Biol* 23: 89

Supplementary Figures

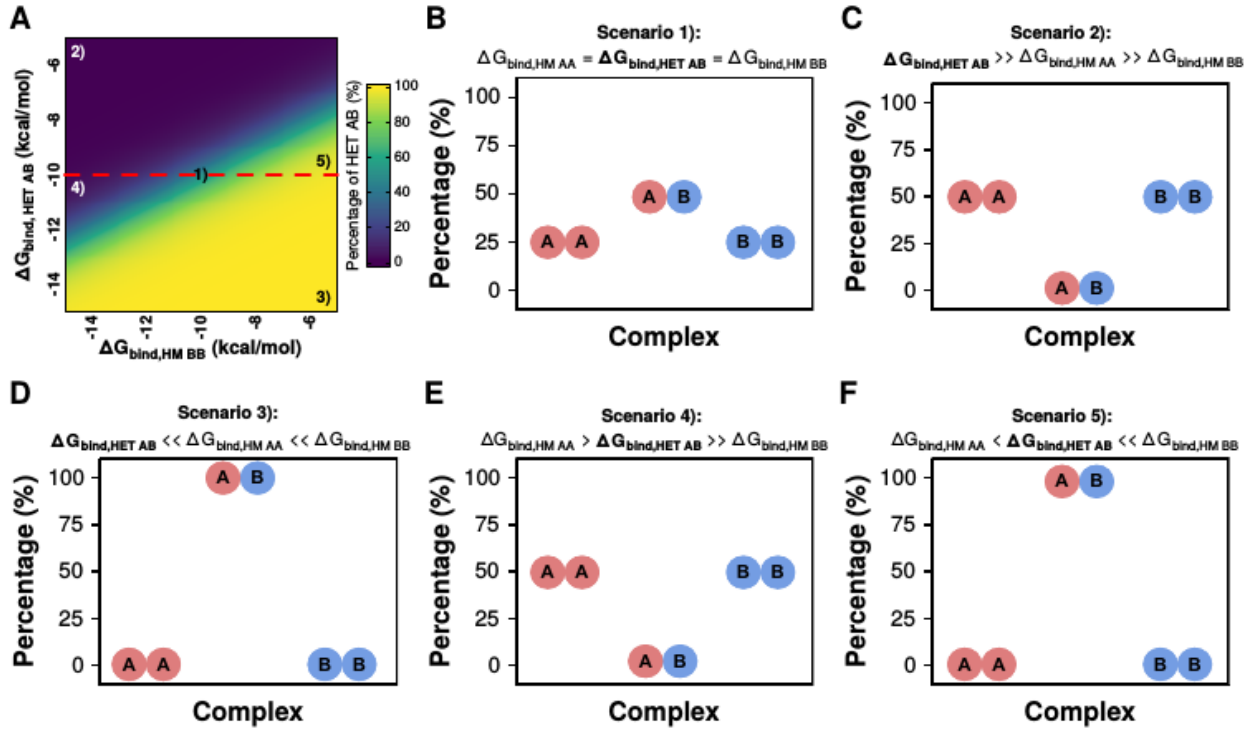

Figure S1. The equilibrium between homodimers and heterodimers is determined by binding affinities.

A) Effect of variation of $\Delta G_{bind,AB}$ and $\Delta G_{bind,BB}$ on the percentage of HET, similar to Figure 1G. The dashed red line indicates the value of $\Delta G_{bind,AA}$, which is kept constant at -10 kcal/mol. Numbered labels inside the heatmap indicate the different scenarios described in the rest of the panels.

B) Scenario 1): Directly after the duplication, the binding affinities of the three dimers are identical, leading to a 25% AA:50% AB:25% BB equilibrium.

C) Scenario 2): Homomers dominate when the heterodimer has the weakest binding affinity.

D) Scenario 3): Heterodimers dominate when they have the strongest binding affinity.

E) Scenario 4): Homomers dominate when the binding affinity of the heterodimer is slightly stronger than that of one of the homodimers but much weaker than the other one.

F) Scenario 5): Heterodimers dominate when the binding affinity of the heterodimer is slightly weaker than that of one of the homodimers but much stronger than the other one.

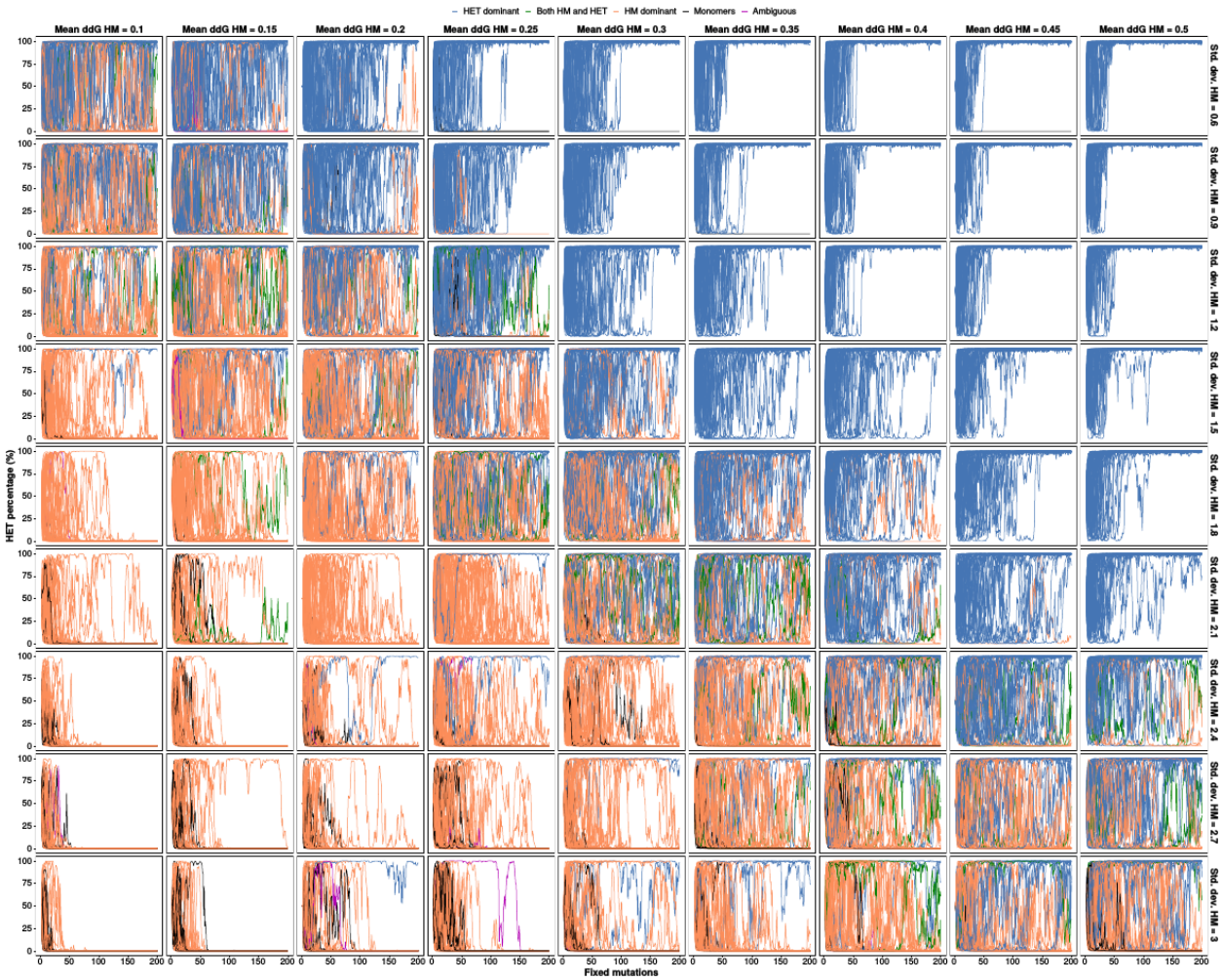

Figure S2. Trajectories of individual replicates of parametric simulations. Panels show trajectories of 50 individual replicates of parametric simulations with different means and standard deviations for the distribution of effects on $\Delta G_{bind,HM}$ (same data used in figure 2C). Replicates were colored according to the equilibrium concentrations of complexes at the end of each simulation following table 1.

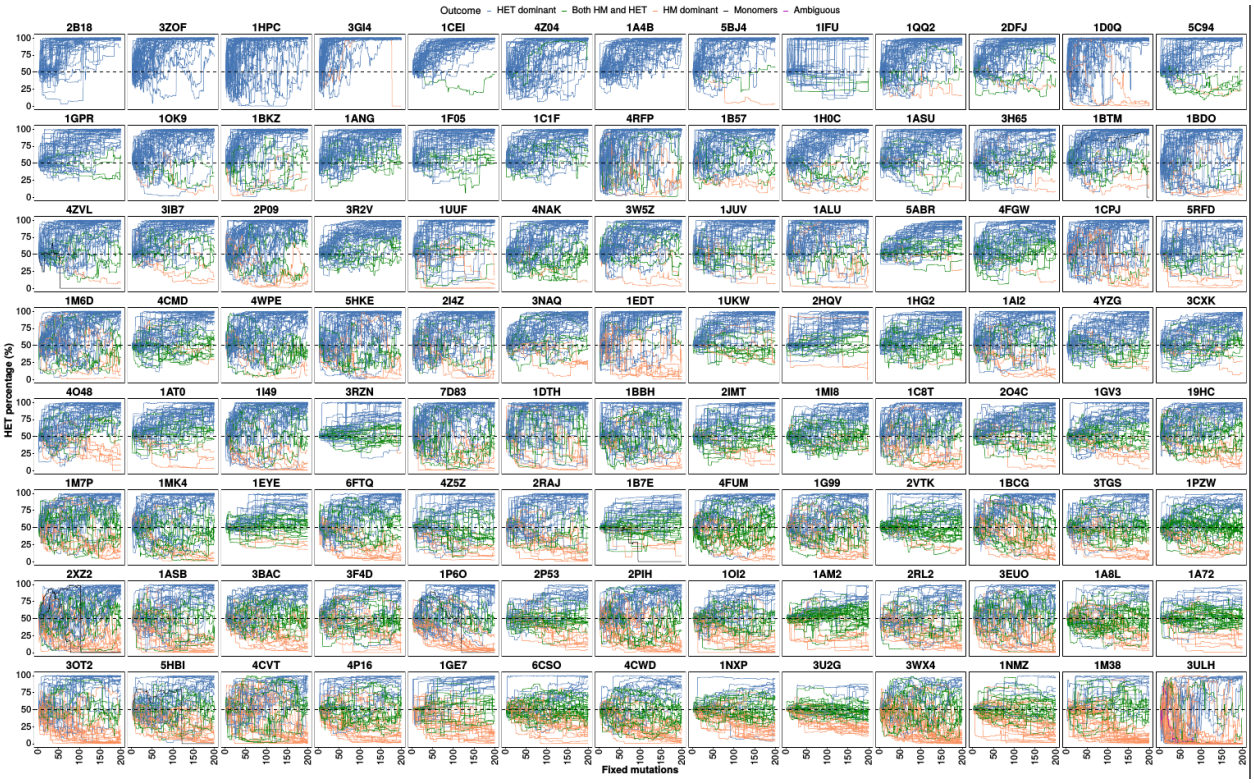

Figure S3. Trajectories of individual replicates of simulations with structures. Panels show trajectories of 50 individual replicates of simulations using the distributions of mutational effects estimated for each PDB structure (same data used for figures 3A-3B). Replicates were colored according to the equilibrium concentrations of complexes at the end of each simulation following table 1.

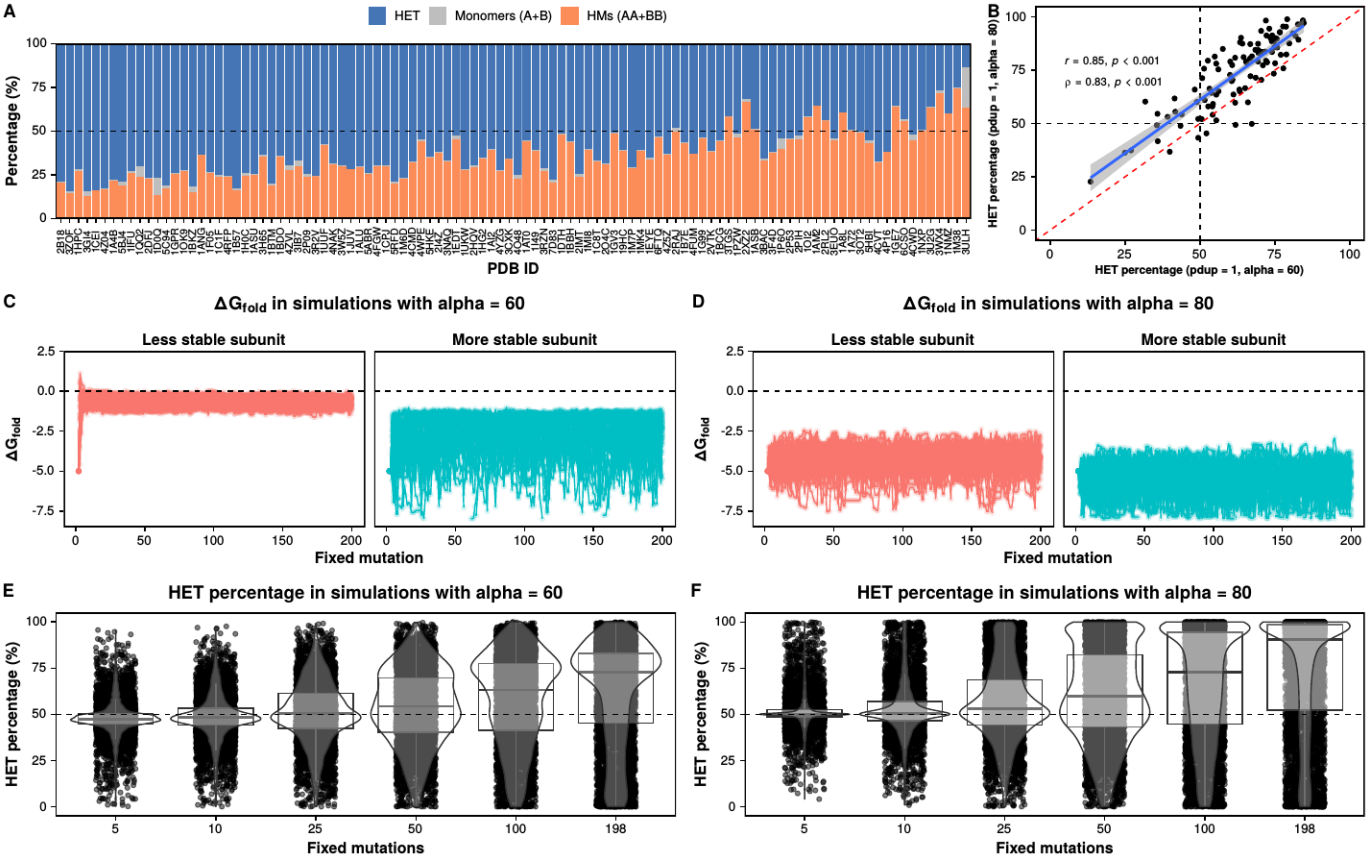

Figure S4. The general bias observed for heterodimers is also observed when the optimal activity is set to a slightly lower value (alpha = 60).

A) Percentage of heterodimers, homomers, and monomers at the end of simulations with distributions of mutational effects structures when the optimal activity (alpha) is set to 60. Structures are listed in the same order as in figure 3A.

B) Correlation between the percentages of heterodimers observed at the end of simulations with alpha = 60 (data from figure S3A) and simulations with alpha = 80 (data from figure 3A).

C-D) Evolution of ΔG_{fold} values in simulations with alpha = 60 (C) and alpha = 80 (D). For simplicity, only the data for PDB: 1GPR are shown.

E-F) Evolution of the percentage of heteromers in simulations with alpha = 60 (E) and alpha = 80 (F). Replicates for all simulations with all structures are shown.

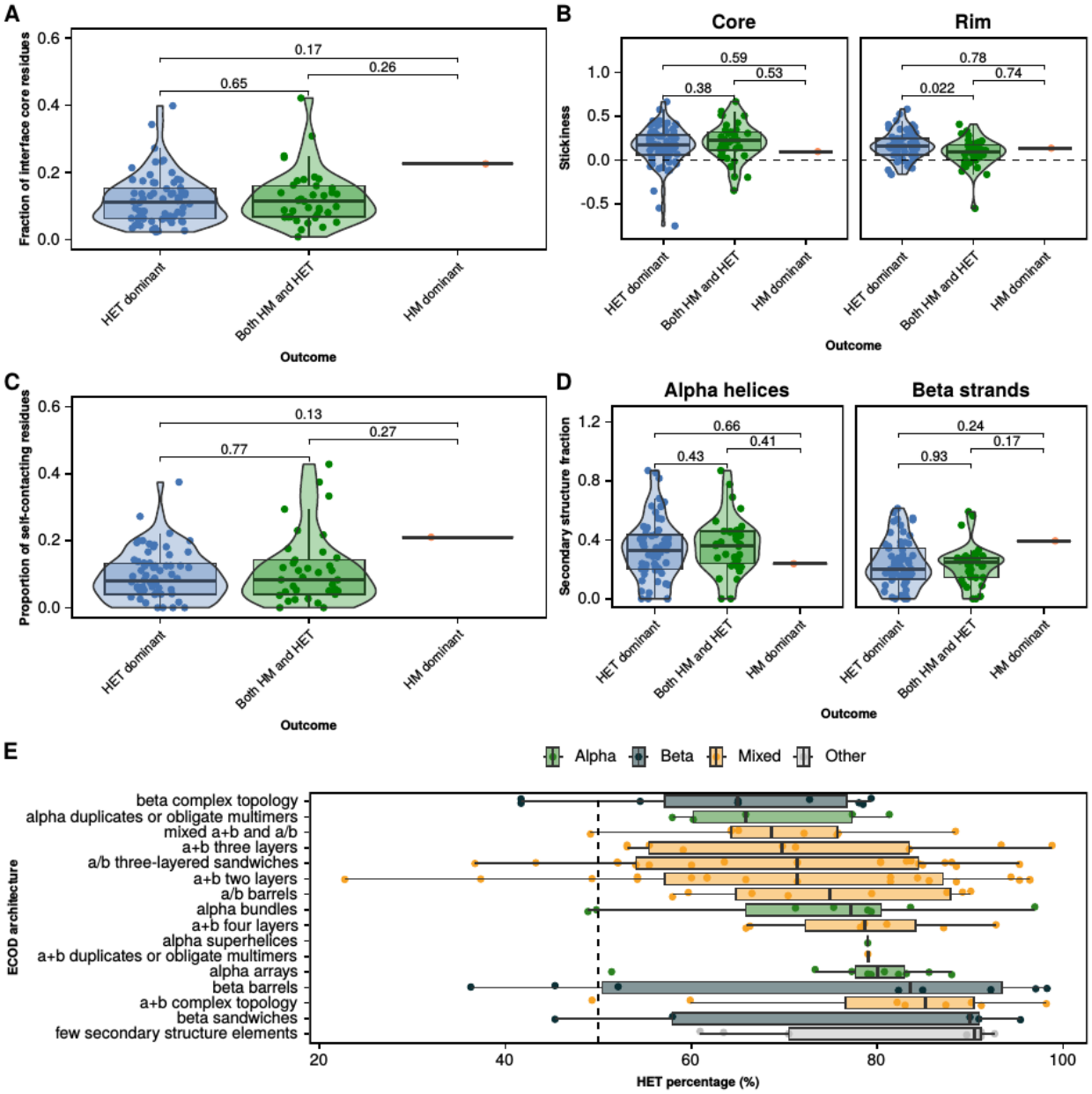

Figure S5. No clear effects of structural properties of protein interaction interfaces on the observed outcomes.

A-D) Distributions for all 104 structures tested for fraction of residues of the whole protein that are located in the interface core (A), the average stickiness of the interface core and rim (B), the proportion of residues in the interface core and rim that are within 4 Angstroms of their counterparts in the opposite subunit (C), and the percentage of alpha helices and beta strands at the interface (D). Structures were classified in panels A-D according to the outcome of their simulations from figure 3A. p-values were calculated using Wilcoxon tests.

E) Percentages of heteromers at the end of the simulation for dimers with different ECOD architectures. Architectures are classified according to their dominant secondary structure (alpha helices, beta strands, or mixed).

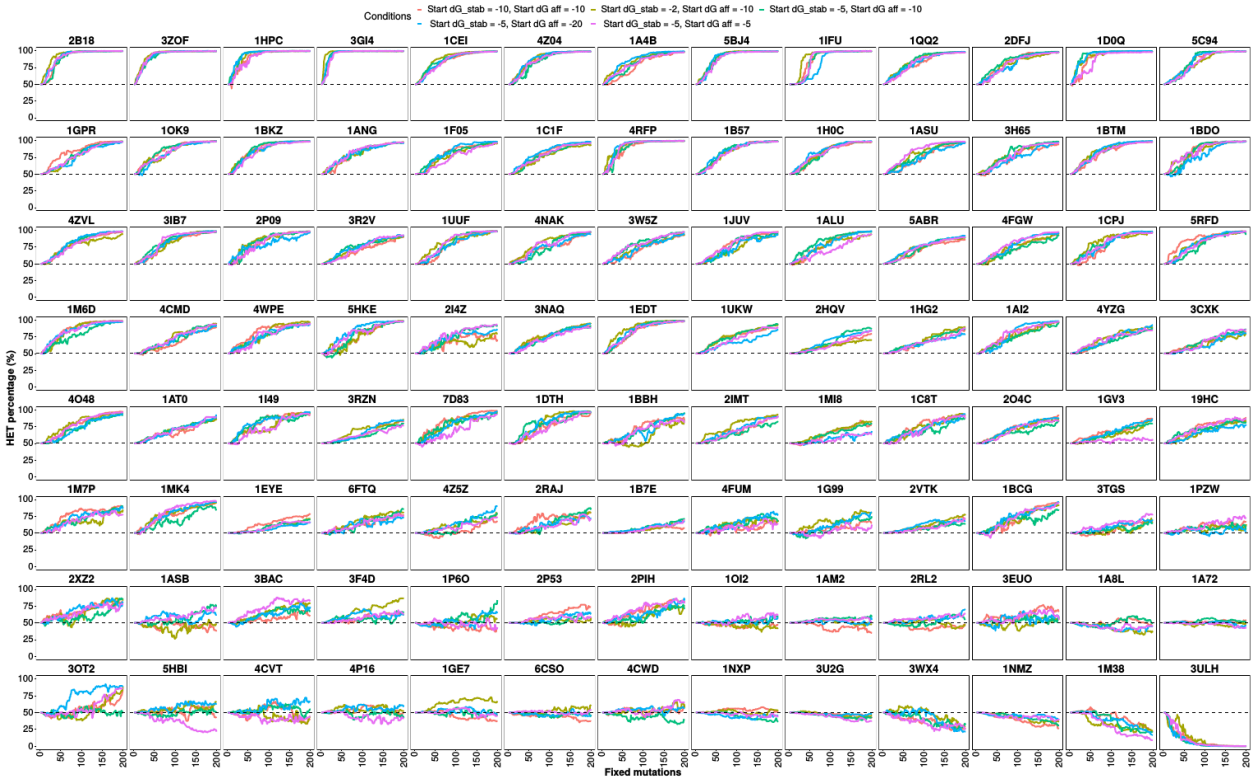

Figure S6. Changes in the starting values for ΔG_{fold} and ΔG_{bind} do not alter the outcome significantly. Simulations were run using the distributions of mutational effects for each PDB structure and changing the starting parameters for ΔG_{fold} and ΔG_{bind}. Each combination of values was selected by doubling or halving one of the reference parameters (ΔG_{fold} = -5 kcal/mol, ΔG_{bind} = -10 kcal/mol) while keeping the other one constant.

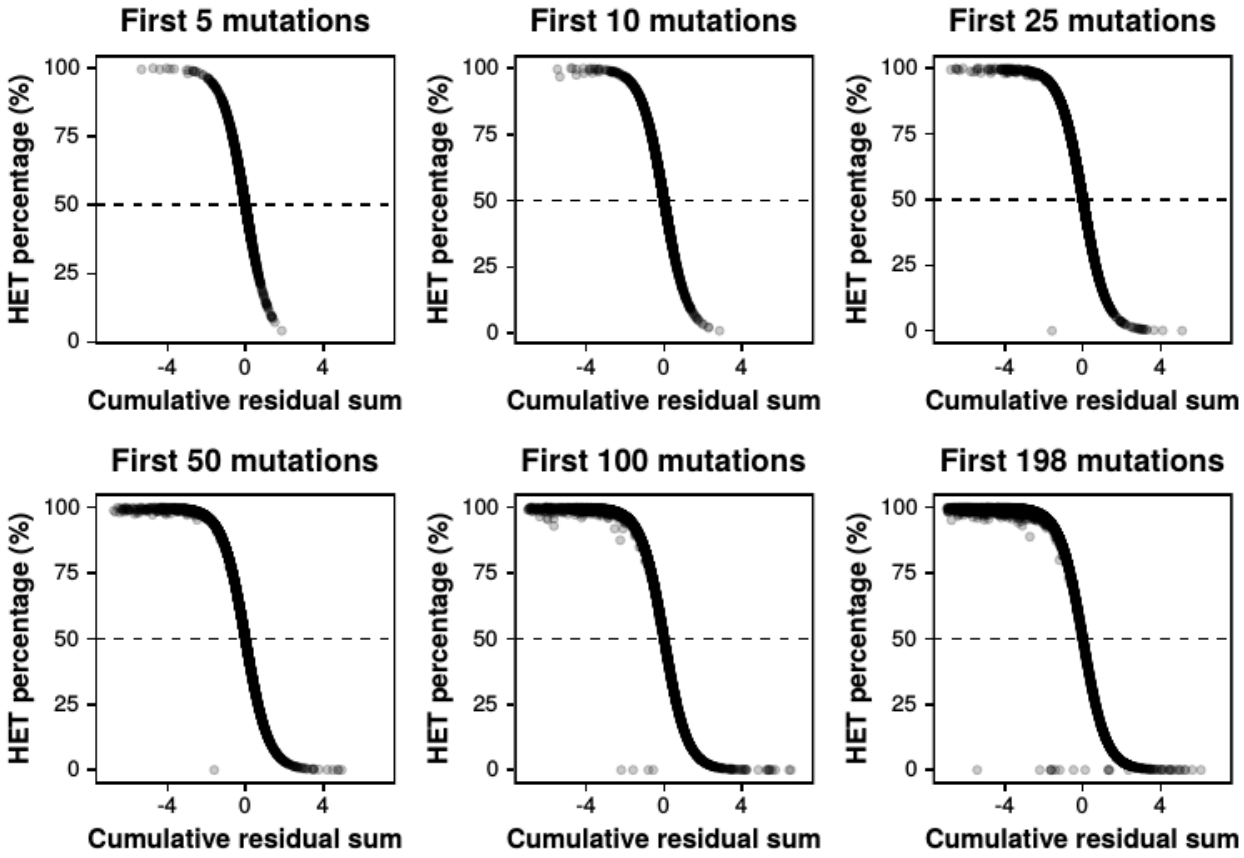

Figure S7. The cumulative sum of residuals of fixed mutations reflects the percentage of heteromers. For each replicate of simulations with each structure, the residuals of fixed mutations up to specific points in the simulation (5, 10, 25, 50, 100, 198 mutations) were added and compared to the percentage of heterodimers at that point in the simulation.

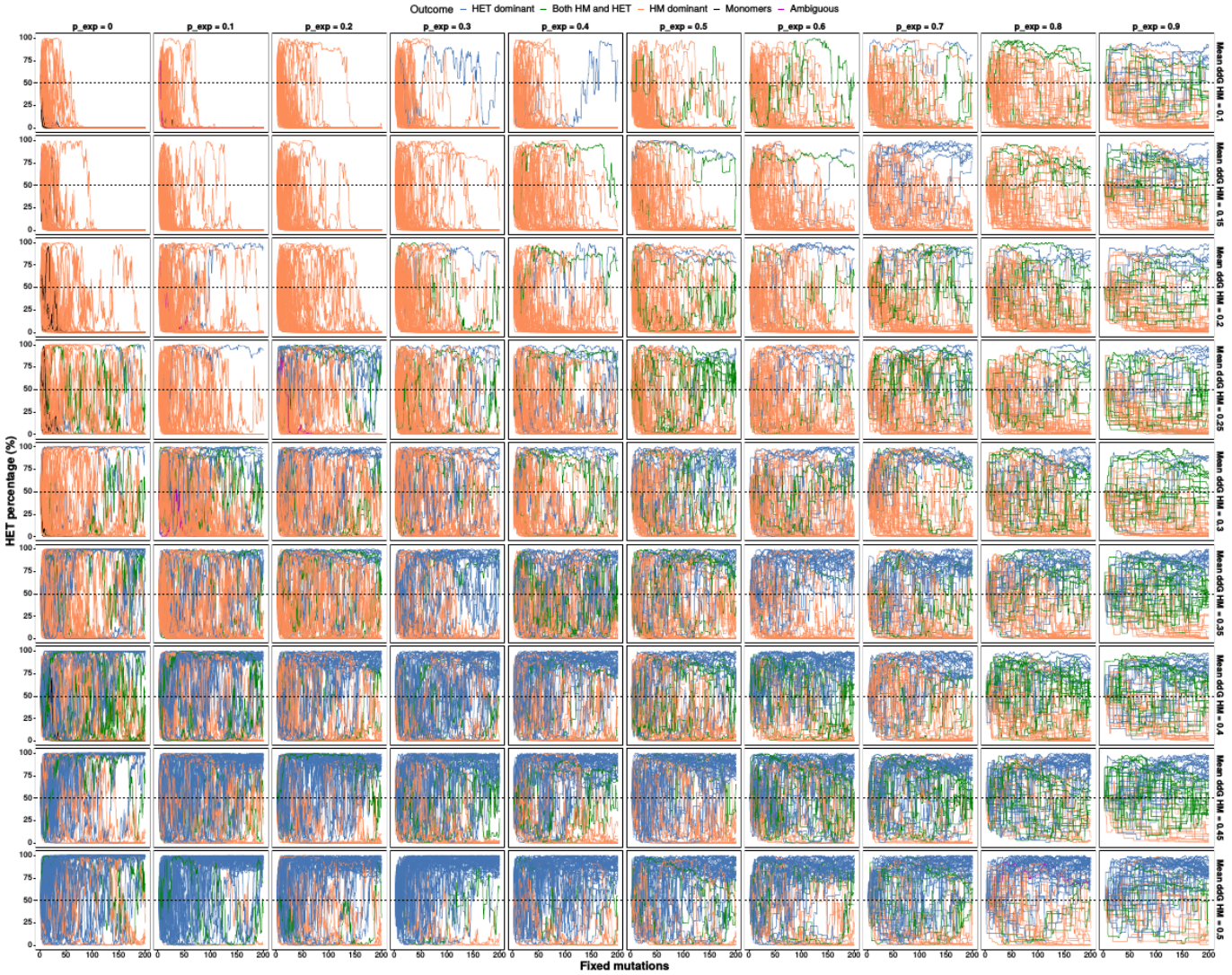

Figure S8. Trajectories of individual replicates for parametric simulations that allow changes in synthesis rates. Individual replicates were colored following table 1 based on the percentage of heteromers at the end of the simulation. The data shown are the same as in figure 5E.

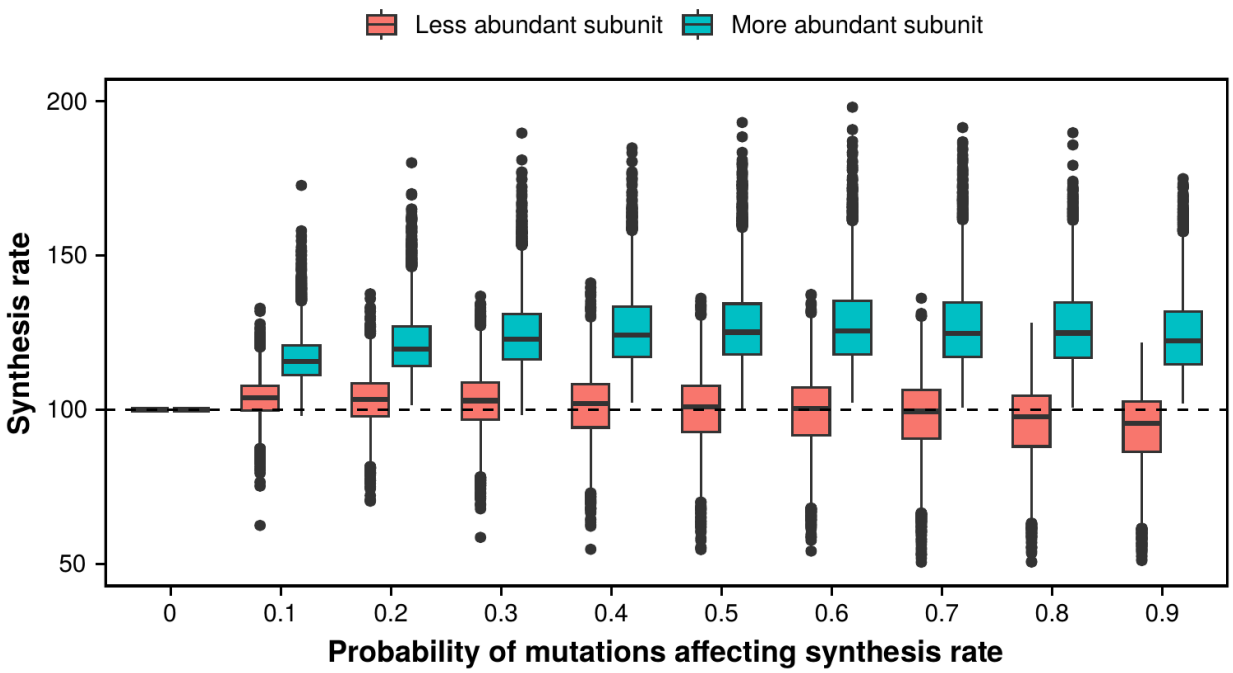

Figure S9. Divergence in synthesis rates between paralogs in simulations with structures. Simulations were run using the distributions of mutational effects for all replicates of each of the 104 PDB structures and different probabilities of mutations affecting synthesis rates. For each simulation, the more abundant subunit and the least abundant subunit were distinguished. The dashed line at 100 indicates the starting synthesis rate for both subunits.

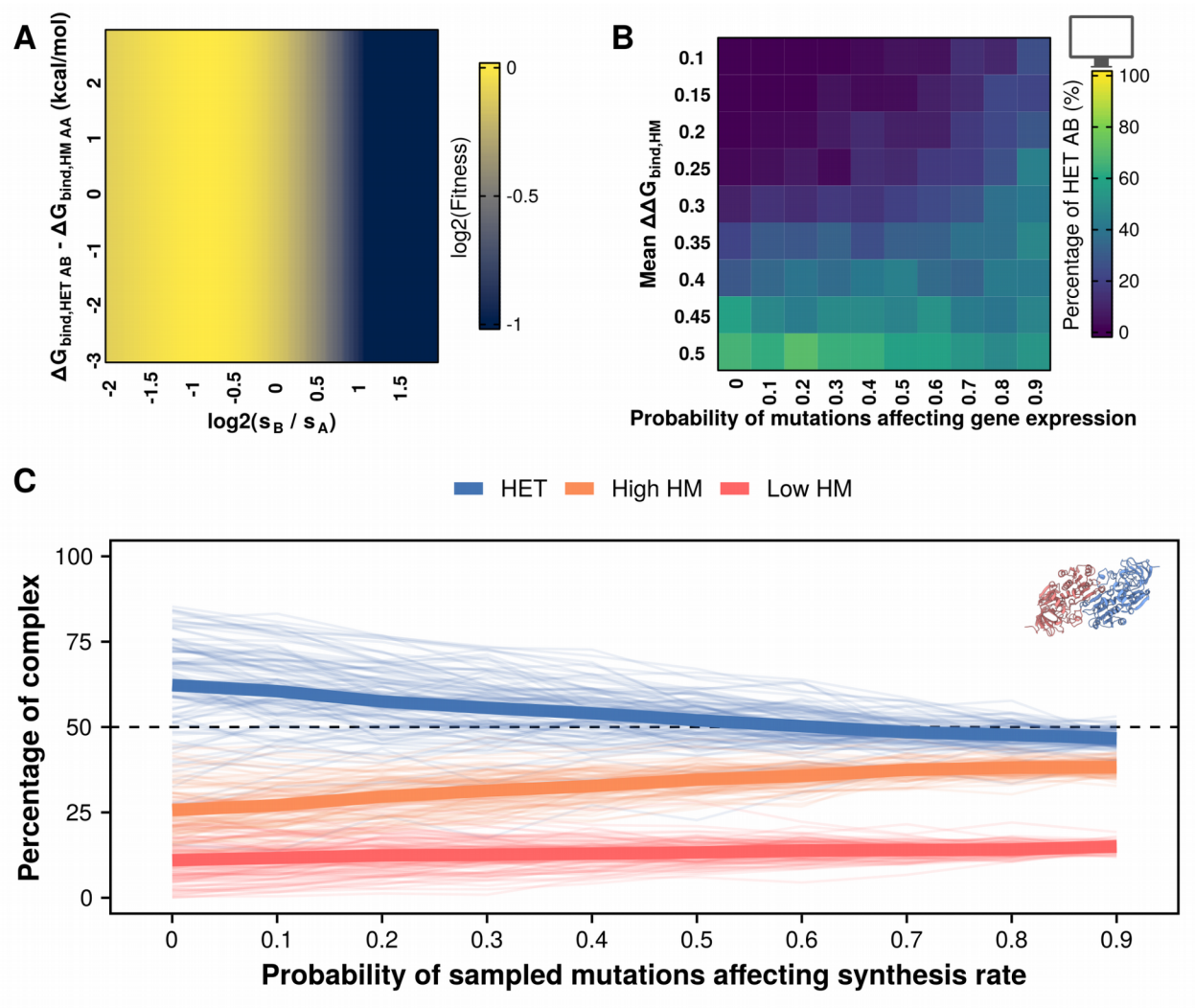

Figure S10. Simulations allowing changes in synthesis rates with lower optimal activity (alpha = 60).

A) Fitness landscape when the optimal activity is set to 60. Note that s_A, $\Delta\Delta G_{bind,AA}$, and $\Delta\Delta G_{bind,BB}$ remain constant throughout.

B) Percentages of heterodimers at the end of parametric simulations allowing changes in synthesis rates and varying mean effects of mutations on homodimers.

C) Percentages of heterodimers and each of the two homodimers for each of the 104 tested structures when simulations consider different probabilities of mutations affecting synthesis rates. The two homodimers were distinguished based on their abundance.

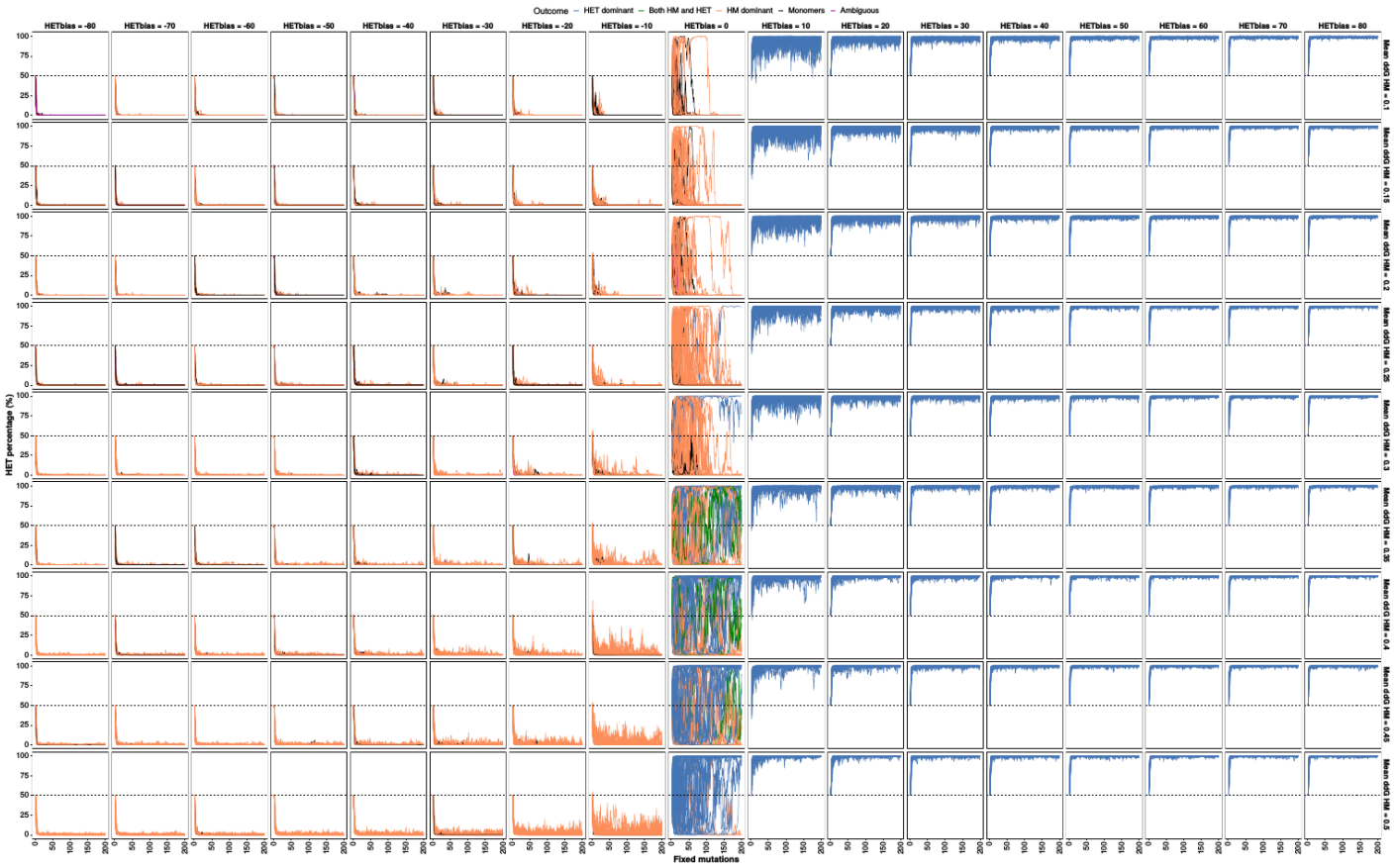

Figure S11. Individual trajectories for parametric simulations with differences in specific activities. Individual replicates were colored following table 1 based on the percentage of heteromers at the end of the simulation. The data shown are the same as in figure 6E.

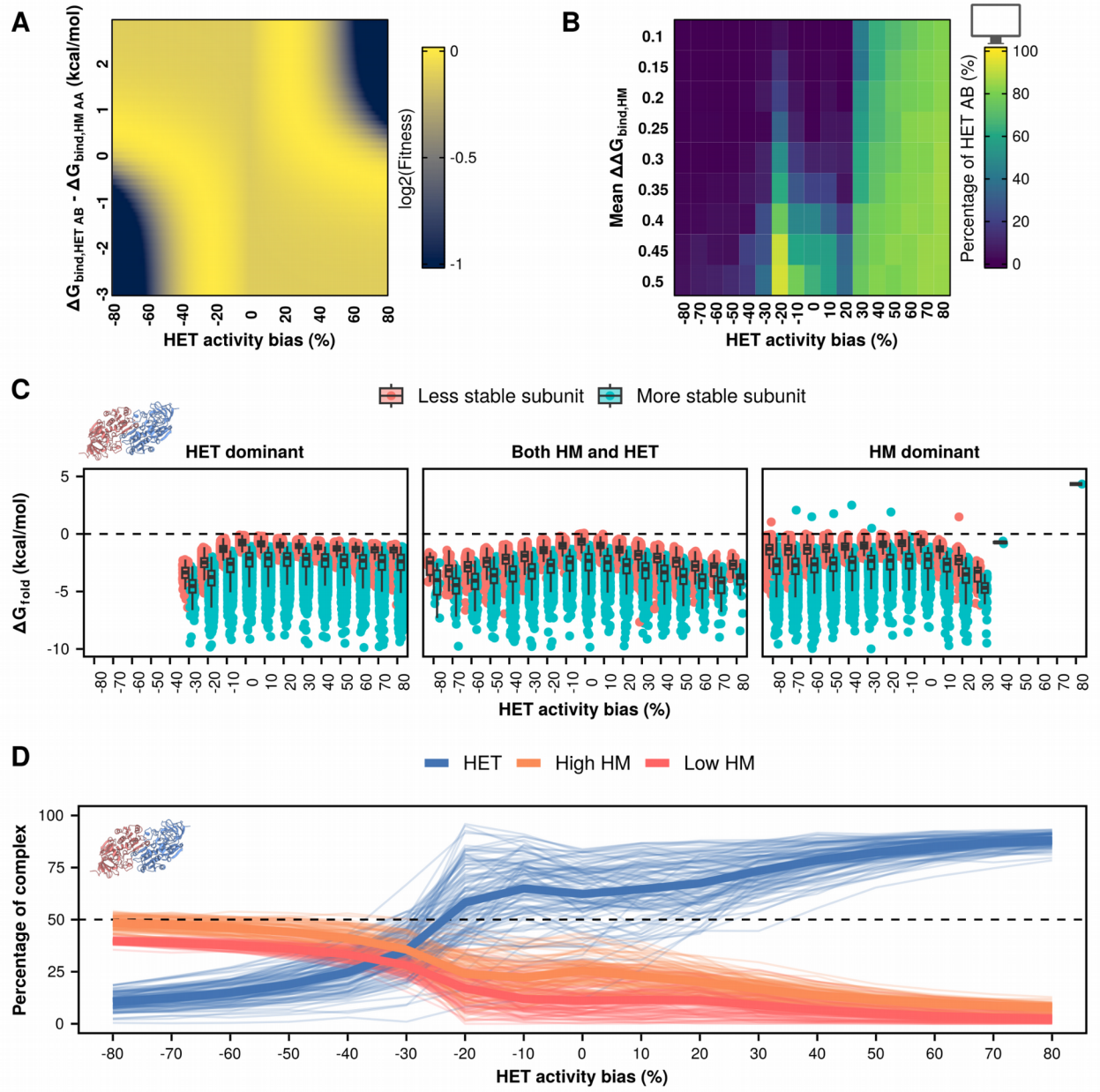

Figure S12. Simulations with differences in specific activity and lower optimal total activity (alpha = 60).

A) Fitness landscape when the optimal activity is set to 60. Note that $\Delta\Delta G_{bind,AA}$, and $\Delta\Delta G_{bind,BB}$ remain constant throughout.

B) Percentages of heterodimers at the end of parametric simulations with different values of specific activity and varying mean effects of mutations on homodimers.

C) ΔG_{fold} values at the end of simulations with alpha = 60 and different specific activities for homodimers and heterodimers. Disfavored outcomes, such as HET dominant when the heterodimer is less active, can still be reached if the subunits remain stable.

D) Percentages of heterodimers and each of the two homodimers for each of the 104 tested structures when

simulations consider different specific activities of homodimers and heterodimers. The two homodimers were distinguished based on their abundance.