A general theory of specific binding: insights from a genetic, mechano-chemical protein model

John M. McBride,[1] Jean-Pierre Eckmann,[2] and Tsvi Tlusty[1,3]

[1] *Center for Soft and Living Matter, Institute for Basic Science, Ulsan 44919, South Korea*
[2] *Département de Physique Théorique and Section de Mathématiques, University of Geneva, Geneva, Switzerland*
[3] *Departments of Physics and Chemistry, Ulsan National Institute of Science and Technology, Ulsan 44919, South Korea*[*]

Proteins need to selectively interact with specific targets among a multitude of similar molecules in the cell. Despite a firm physical understanding of binding interactions, we lack a general theory of how proteins evolve high specificity. Here, we present a model that combines chemistry, mechanics and genetics, and explains how their interplay governs the evolution of specific protein-ligand interactions. The model shows that there are many routes to achieving discrimination – by varying degrees of flexibility and shape/chemistry complementarity – but the key ingredient is precision. Harder discrimination tasks require more collective and precise coaction of structure, forces and movements. Proteins can achieve this through correlated mutations extending far from a binding site, which fine tune the localized interaction with the ligand. Thus, the solution of more complicated tasks is aided by increasing the protein, and proteins become more evolvable and robust when they are larger than the bare minimum required for discrimination. Our model makes testable, specific predictions about the role of flexibility in discrimination, and how to independently tune affinity and specificity. Thus, the proposed theory of molecular discrimination addresses the natural question "why are proteins so big?" A possible answer is that molecular discrimination is often a hard task best performed by adding more layers to the protein.

Significance statement. Proteins excel at discriminating between molecules. Otherwise, cells would not be able to accurately translate mRNA into protein sequences, chemical signals would get mixed up, and enzymes would catalyze unwanted interactions. To understand how proteins can achieve this, we develop a model of protein binding to study evolution of molecular discrimination through the interplay of shape, flexibility, chemical binding and entropy. The model reveals that mutations far from the binding site can fine-tune binding interactions, and larger proteins are better discriminators. Proteins thus solve the discrimination problem by being large so that they have many ways of fine-tuning binding.

INTRODUCTION

Proteins are the main molecular workforce inside cells, and the tasks they perform invariably rely on short-range binding interactions. The cell is filled with thousands of molecular species, some differing by only a single atom. Yet somehow, most proteins can specifically bind to only a few select species.[1–3] Despite knowing much about pairwise binding,[4–13] we have little mechanistic understanding of how proteins *selectively* bind to targets, while avoiding interactions with similar, but non-cognate, molecules.[10] Such undesirable interactions can, at best, lead to inefficiencies through inhibition,[14] and at worst, result in aggregation[15] or cross-talk between signals.[16]

Unwanted interactions can be minimized by designing mismatch between ligands and binding pockets,[17,18] such that deformation costs energy – a form of 'conformational proofreading'.[19] Other work has suggested that residues not directly involved in binding may play a role in discrimination and allostery.[20–22] Here, taking inspiration from the many experimental examinations of molecular discrimination by proteins, such as enzymes, tRNA synthetases, transcription factors, and antibodies ,[1,2,23–30] we propose a simple yet general theory of specific binding.

We study the evolution of discrimination by proteins using a genetic, mechano-chemical model of binding. We find that, although the discrimination problem is in general difficult, it has many possible solutions: shape mismatch, chemical complementarity, and flexibility can all be manipulated to tune interaction specificity. The important common thread is that it requires precision – *e.g.*, just the right amount of shape mismatch for a given flexibility. We show that residues distant from the binding site enable this fine-tuning of mechanical deformation upon binding – demonstrating that larger proteins benefit from having more degrees of freedom. It follows that harder discrimination challenges require larger proteins. Furthermore, larger proteins are more evolvable and robust since their set of possible functional sequences is larger and more connected. We further explain the mechanisms through which affinity and specificity can be tuned independently, and discuss the role of flexibility and entropy. Altogether, the simple, coarse-grained model shines light on the difficult problem of how proteins achieve such superlative molecular discrimination. At the same time, by linking protein size to a property as ubiquitous as binding, we offer a possible answer to the question, "why are proteins so big?".[31,32]

RESULTS

We introduce the model in two stages. To gain insight into the interplay between deformation

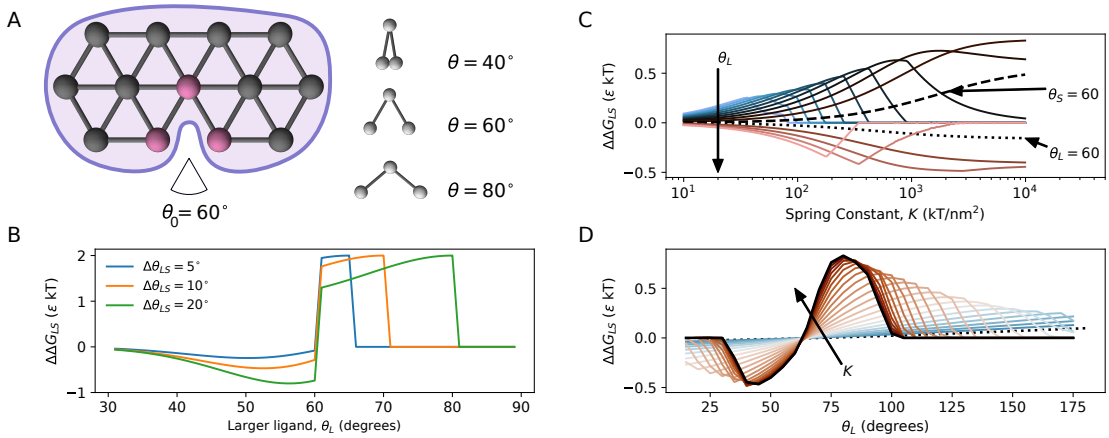

FIG. 1. A: A mechano-chemical model of molecular recognition. A protein is modelled as a 2-d spring network with a chemical binding site (three interaction sites, shown in pink); the Λ-shaped ligands have three interaction sites, and are uniquely defined by an angle θ. B: Binding energy gap in a rigid protein $\Delta\Delta G_{SL} = \Delta G_L - \Delta G_S$ as a function of the size of the larger ligand θ_L, for various ligand size differences $\Delta\theta_{LS}$. C: $\Delta\Delta G$ as a function of the spring constant K for $\Delta\theta_{LS} = 5^\circ$, for $10^\circ \leq \theta_L \leq 175^\circ$. D: $\Delta\Delta G$ as a function of θ_L for $\Delta\theta_{LS} = 5^\circ$, for $10 \leq K \leq 10^4$ kT/nm^2.

and binding energy, we consider first a "mechano-chemical" model – a protein whose bonds all have the same spring constant K, and K can be continuously tuned to improve discrimination. Later, we will introduce genetics by considering heterogeneous bond strengths encoded in a gene, resulting in more realistic evolution that proceeds via discrete steps.

Mechano-chemical model of molecular recognition. The model protein is a 2-dimensional elastic network[33–35] with harmonic bonds of length ∼1 nm and three chemical binding sites, each with binding energy constant ε (Fig. 1A). To make sense of this coarse-grained representation, one may envision the nodes as groups of tightly-connected amino acids that have highly-correlated motion.[36]

The protein binds to a rigid triangular ligand specified by an angle θ. We use the following terminology: Ligands that do not fit in the binding pocket ($\theta > \theta_0 = 60^\circ$) are called 'fat', otherwise 'thin'. When we compare two ligands, the larger of the two is denoted as ligand L (θ_L), and the smaller as ligand S (θ_S). We also refer to the cognate ligand (θ_C), *i.e.* the target ligand of the protein, and the non-cognate ligand (θ_{NC}), the functionally undesirable one. For each protein-ligand combination we find a minimum energy configuration (Methods) by moving and deforming the protein.

Recognition via lock-and-key binding. How can specific molecular recognition be achieved? Binding depends only on the free energy of binding ΔG. Discrimination, however, depends on the difference in binding energy between the cognate ligand and one or more non-cognate ligands, $\Delta\Delta G = \Delta G_{NC} - \Delta G_C$. Here, ligands vary along a single dimension – shape mismatch with the protein binding pocket, $\Delta\theta_0 = \theta - \theta_0$ – so typically we only need to consider one non-cognate ligand.

As a starting point, we examine the limiting case of a completely rigid protein, which results in *lock-and-key* binding. As no deformation occurs upon binding, the binding energy is simply

$$E_{\text{rigid}} = \begin{cases} -\varepsilon, & \text{if } \theta > \theta_0 \\ -\varepsilon\left(2 + e^{-\delta r^2/\sigma^2}\right), & \text{if } \theta \leq \theta_0 \end{cases},$$

where $\delta r = 2\sin((\theta_0 - \theta)/2)$ and $\sigma \sim 0.3$ nm is the interaction range (Methods). This is because fat ligands can only interact via one site, and thin ligands minimize binding energy by fully binding to two sites, and partially binding to the third.

In this limiting case, the best binding gap, $\Delta\Delta G = 2\varepsilon$, is achieved via *steric exclusion*: when one ligand is a perfect match for the binding site ($\theta_0 - \theta_S = 0$), and the other ligand is fat, for any $\Delta\theta_{LS} = \theta_L - \theta_S$ (Fig. 1B). If both ligands are thin, the potential binding gap is much lower ($\Delta\Delta G = \varepsilon$),[30] and can only be obtained for sufficiently dissimilar ligands ($\Delta\theta_{LS} \geq 40^\circ$). If the two thin ligands are similar, the optimal binding gap is only achievable with some *mismatch* between the cognate ligand and the protein ($\theta_L < \theta_0$). We thus see that binding with a rigid protein is in principle a feasible strategy for molecular discrimination. However, even in this hypothetical case shape mismatch is often necessary to promote binding of the cognate ligand while avoiding binding of the non-cognate.

Preferential binding by flexible proteins exploits ligand shape mismatch. Rigidity in proteins is limited by the nature of the non-covalent bonds holding them together. To explore the effect of rigidity

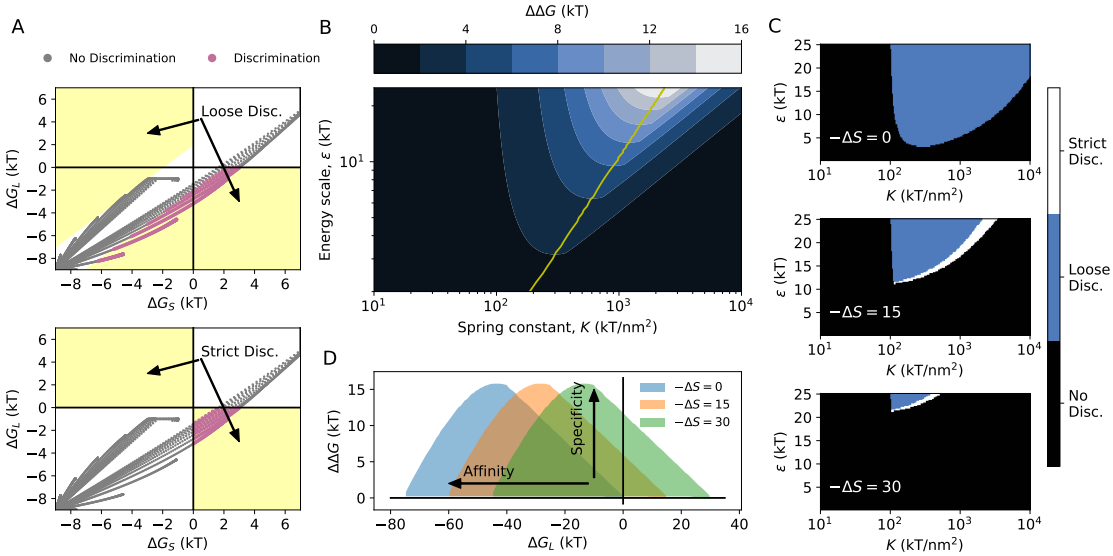

FIG. 2. A: Binding energy for ligand L vs. ligand S ($\varepsilon = 8\,\mathrm{kT}$, $-\Delta S = 15$). Shaded regions indicate "loose discrimination" ($\Delta G_\mathrm{C} < 0$ and $\Delta G_\mathrm{C} - \Delta G_\mathrm{NC} < -2\,\mathrm{kT}$), or "strict discrimination" ($\Delta G_\mathrm{C} < 0$ and $\Delta G_\mathrm{NC} > 0$). Binding energy is shown for the proteins from Fig. 1C-D ($\Delta\theta_\mathrm{LS} = 5°$). B: Binding energy gap as a function of spring constant K, and chemical energy constant ε. Optimal K for a given ε is indicated by the yellow line. C: Identification of regions that satisfy the conditions for loose and strict discrimination. D: Envelope of possible values of affinity (ΔG) and specificity ($\Delta\Delta G$) for the parameter space sampled in (B). Different envelopes are shown for values of entropic cost $-\Delta S$.

on discrimination, we first set wide bounds on the spring constants ($10 \leq K \leq 10^4\,\mathrm{kT/nm^2}$), and later narrow these to more biologically-plausible values of spring constants.[37–39] When K is low, binding is not very specific as the protein can perfectly match any ligand with little deformation energy (Fig. 1C). As K increases, specificity improves, until the protein is so rigid that neither cognate nor non-cognate ligands can induce the protein to deform. The maximum binding energy gap for fat ligands ($\Delta\Delta G = 0.84\varepsilon$, $\Delta\theta_\mathrm{LS} = 5°$, $K = 10^4\,\mathrm{kT/nm^2}$) is considerably lower than what can be achieved by rigid lock-and-key binding ($\Delta\Delta G = 2\varepsilon$).

For thin ligands, on the other hand, the maximum gap ($\Delta\Delta G = 0.49\varepsilon$, $\Delta\theta_\mathrm{LS} = 5°$, $K \sim 3 \cdot 10^3\,\mathrm{kT/nm^2}$) is double (Fig. 1C) of what can be achieved by a completely rigid protein ($\Delta\Delta G = 0.24\varepsilon$, $\Delta\theta_\mathrm{LS} = 5°$). In general, higher specificity is still achieved with greater rigidity across biologically-relevant values of K; but this requires ever-greater shape match between the protein and the cognate ligand, as $\Delta\Delta G$ becomes a steeper function of ligand shape (Fig. 1D). This coupling between flexibility and shape mismatch highlights the need for precise concerted control over both protein structure and dynamics.

Discrimination is more difficult than recognition. Binding energy gap is a key determinant of molecular discrimination. But binding will only occur spontaneously if the corresponding free energy change ΔG is negative, which depends on binding entropy.[40,41] The binding entropy ΔS is the logarithm of the relative change in the volume of the configuration space accessible by thermal fluctuations, and contributes to free energy as, $\Delta G = \Delta E - T\Delta S$, where ΔE is the sum of the mechanical and chemical binding energies, T is temperature, and ΔS is the entropy change upon binding. Taking this into account, we now formally define molecular discrimination: 'loose' discrimination is defined by $\Delta G_\mathrm{C} - \Delta G_\mathrm{NC} < -2\,\mathrm{kT}$, and $\Delta G_\mathrm{C} < 0\,\mathrm{kT}$; we use $-2\,\mathrm{kT}$ as a reasonable threshold that corresponds to a sevenfold difference in binding affinity; 'strict' discrimination is defined by $\Delta G_\mathrm{C} < 0$ kT, and $\Delta G_\mathrm{NC} > 0$ kT.

We replot the data from Fig. 1C-D in terms of discrimination, finding that specific binding is substantially more difficult than recognition alone (Fig. 2A). Even for our permissive threshold of $-2\,\mathrm{kT}$, only a fraction of cases result in discrimination. The disparity between the ability to distinguish between sets of fat ($\Delta G_\mathrm{S} < \Delta G_\mathrm{L}$) or thin ($\Delta G_\mathrm{L} < \Delta G_\mathrm{S}$) ligands illustrates the utility of steric exclusion as a discrimination mechanism (discrimination is possible over a large range of $-\Delta S$ for fat ligands, but if the ligands are thin then it is difficult to achieve discrimination).

Affinity and specificity are correlated. To understand the role of mechanical and chemical binding energy in discrimination, we now vary the chemical energy constant so that the chemical binding energy is within a reasonable biological range ($1 \leq \varepsilon \leq 25$ kT) .[42–46] Once again, increasing K aids specificity,

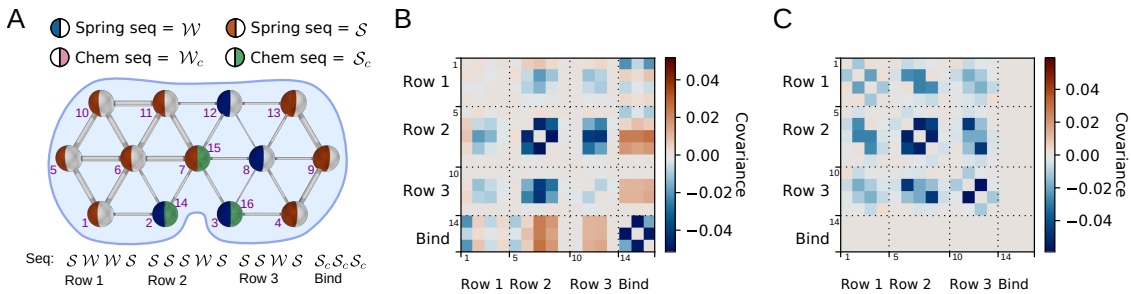

FIG. 3. A: Linking mechanics and binding chemistry to genetic sequence in the model: amino acids are either **w** (blue, forming weak bonds), or **s** (orange, forming strong bonds), as indicated by the colour of the left-half of the spheres; edge width indicates the three possible values of bond strength of the bonds as determined by the sequence; chemical energy is similarly determined by sequence as $\mathbf{w}_c$, or $\mathbf{s}_c$ as indicated by the colour of the right-half of the spheres. Sequence is read from left-to-right, from bottom-to-top (as indicated by the numbers). B-C: Sequence covariance for all loose-discrimination solutions (B), and solutions with $|\Delta\Delta G| > 3.6$ kT (C). For this example, the parameters are: $K_s = 1{,}000$, $K_m = 400$, $K_w = 100\,\mathrm{kT/nm^2}$; $\varepsilon_s = 6$, $\varepsilon_w = 1.75$ kT; $\theta_\mathrm{S} = 80°$, $\theta_\mathrm{L} = 90°$.

up until a point where the protein is too rigid to accommodate either ligand (Fig. 2B). Beyond this point, increases in K must also be accompanied by higher ε, so that there is sufficient chemical driving force for mechanical deformation. ε sets an upper bound on specificity ($\Delta\Delta G$), which is achieved at some optimal K, and this optimal K is proportional to ε. As a result, affinity and specificity are naturally correlated within the parameter space.

Entropy can be used to modulate affinity and specificity. Binding entropy is a key determinant of molecular discrimination (Fig. 2C).[43,46–49] For the same discrimination task as in Fig. 2B, only loose discrimination is possible for $-\Delta S = 0$ since both ligands bind to the protein. As we increase the entropy cost $-\Delta S$, loose discrimination becomes more difficult since stronger chemical bonds are needed for binding. However, increasing $-\Delta S$ also enables strict binding, up to the point where the entropic cost is too high for any ligand to bind. Entropy thus offers a way to decouple affinity and specificity (Fig. 2D).

Sequence variation reveals epistasis. We have so far ignored the quintessential feature of proteins – proteins are heteropolymers, composed of distinct amino acids whose sequence is encoded in genes. We now expand our model to include these two important components – heterogeneity in flexibility, and change via discrete mutations – by coupling the mechanics and chemistry to the protein sequence.

Thus, we examine a model protein consisting of $N_\mathrm{A} = 13$ amino acid letters, and $N_\mathrm{B} = 3$ binding letters, using a 2-letter alphabet for each (Fig. 3A). Amino acids are either **w** or **s**, such that harmonic bonds between neighbours depend on their identities: bonds are either strong (**s**–**s**, K_s), medium (**s**–**w** or **w**–**s**, K_m), or weak (**s**–**s**, K_w). Similarly, chemical binding strength is determined by the letters $\mathbf{w}_c$ and $\mathbf{s}_c$, respectively, resulting in weak (ε_w) and strong bonds (ε_s). We also take into account the fact that the change in conformational entropy is greater for more flexible proteins by calculating the conformational entropy of the protein before and after binding (Methods).

For a representative set of parameters, we calculate the binding free energy ΔG for all possible $2^{N_\mathrm{A}+N_\mathrm{B}} = 2^{16}$ sequences, and identify those sequences that result in discrimination. On average, there is no bias towards a particular amino acid at any position (SI Fig 1), with the exception of strong chemical binding at the tip of the Λ-shaped binding pocket (position 15). However, detailed examination of sequence covariance reveals substantial, non-random patterns of *epistasis* (Fig. 3B). There are positive correlations between the chemical binding sites and the central amino acids in rows 2 and 3 – increasing the stiffness of these amino acids allows for stronger chemical binding. One could anticipate this, given the correlations between ε, K and $\Delta\Delta G$ found in Fig. 2B. Thus, the subset of solutions with $|\Delta\Delta G| > 3.6\,\mathrm{kT}$ (Fig. 3C) contains only those sequences that have taken advantage of this positive correlation – all sequences have strong chemical bonds, and the central amino acids typically have strong bonds (SI Fig 1).

Far-away residues enable fine-tuning of protein binding. All epistasis visible in Fig. 3C is the outcome of negative correlations in the sequence. This can be attributed to a process where protein flexibility is fine-tuned, to achieve the optimum amount of deformation at the binding site. For example, there are negative correlations between all of the central seven amino acids, such that a mutation at any of these points is likely to be accompanied by an opposite mutation at one of the other sites. This is because many different sequences can encode similar dynamical modes of hinge-like motion (SI Fig 1). Note, however, that this effective fine-tuning of

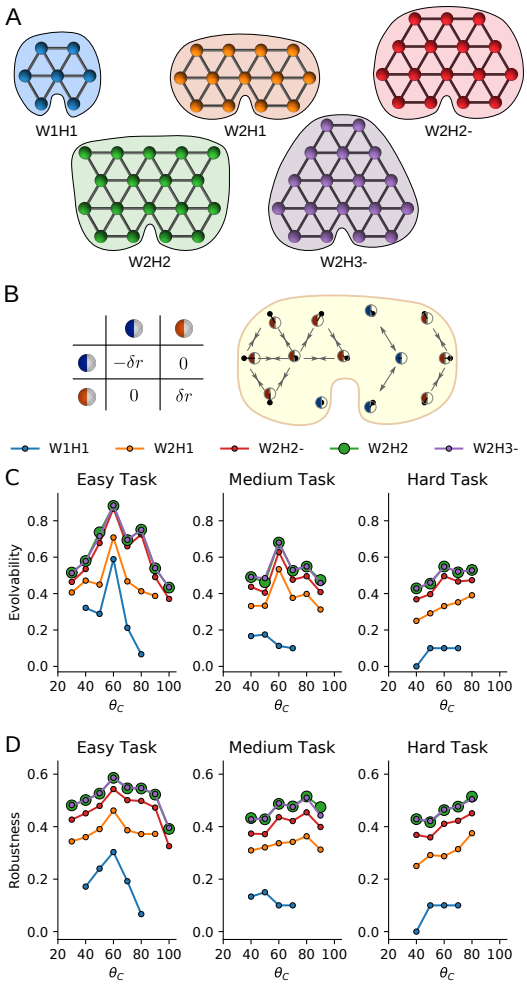

FIG. 4. A: Protein models of different size. B: We map protein sequence to perturbations in the protein structure according to an interaction matrix. If the interaction is positive, neighbouring amino acids experience an attractive force with displacement magnitude $\delta r = 0.1$ nm and vice versa; the perturbation vector (black line) is the average of all neighbouring interactions with neighbouring amino acids. The small black circles are the original positions, and the large circles show both the identities of the amino acids and the new equilibrium configuration. C-D: Evolvability (C) and robustness (D) (both normalized by genome size) as a function of the angle of the cognate ligand, θ_C, for easy, medium, and hard discrimination tasks. Colours indicate protein models of different size. Green circles are large to highlight that they overlap with purple circles. Symbols are only shown for ligands for which solutions were found.

the protein mechanics is not limited to the binding site and is evident throughout the full length of the protein, between pairs of residues at distances of 2 nm (SI Fig 2).

Larger proteins are more evolvable and robust. Residues far from the binding site give the protein extra degrees of freedom that can be used to fine-tune binding. We thus study how these degrees of freedom are related to the ability to evolve functional sequences that are robust to mutations, by modelling proteins of different size (Fig. 4A). To study protein evolution, we link genotype to fitness via a simple measure of binding specificity,

$$F = e^{-\Delta G_C} - \sum_i e^{-\Delta G^i_{NC}} , \quad (1)$$

where ΔG_C is the binding energy of the cognate ligand, and ΔG^i_{NC} is the binding energy of non-cognate ligand i.

Real fitness landscapes have numerous local minima,[50,51] but our model produces very smooth fitness landscapes because shape mismatch ($\Delta\theta_0$) varies only along a single dimension. However, when we allow the protein *equilibrium* structure to change depending on sequence variation – a more realistic approximation – the generate free energy landscape is much more rugged. We do not explicitly model protein folding, but instead account for sequence variation resulting in small perturbations to the native structure (Fig. 4B, Methods). Thus, we study the fitness landscapes of our proteins using a set of ligands, $\theta \in \{20°, 30°, \ldots 100°\}$. For each cognate ligand θ_C the corresponding non-cognate ligands are those for which $|\theta_C - \theta_{NC}| \leq \phi$. We consider three discrimination tasks: easy, $\phi = 10°$, medium, $\phi = 20°$, and hard, $\phi = 30°$.

We find that larger proteins are able to solve discrimination tasks for a greater range of cognate ligands, and longer sequences are more evolvable and robust,[52–54] even after controlling for genome size (Fig. 4C-D, Methods). Increasing protein size gives diminishing returns which eventually saturate, indicating that finite degrees of freedom are sufficient to achieve a maximally evolvable and robust protein. For more difficult tasks, proteins of different size exhibit a larger difference in evolvability and robustness, and larger proteins are needed to reach the saturation point (Fig. 4C-D, SI Fig 3). Thus, the required degrees of freedom depend on the difficulty of the discrimination problem.

DISCUSSION

Through analysis of a conceptually simple, yet multi-faceted model of molecular discrimination, we start to understand the mechanisms, and evolution, of molecular discrimination.

How difficult is molecular discrimination? It depends on the context. For example, proteins that bind to nucleic acids or lipids must find their target out of a dizzying gallery of lookalikes.[25,55,56] Enzymes need to be able to release their product after catalysis,[57] which can be more (*e.g.*, isomerization reactions) or

less (*e.g.*, proteolysis) challenging depending on how similar the product is to the substrate. In principle, the difficulty, Ψ, can be expressed as a function of the set of ligands, $\{\ell\}$, and the required degree of specificity,

$$\Psi\left[\ell^{\mathrm{C}}, \Delta G_{\mathrm{C}}, \ell_1^{\mathrm{NC}}, \Delta\Delta G_1, \ldots, \ell_i^{\mathrm{NC}}, \Delta\Delta G_i\right] ,$$

where ΔG_{C} is the optimum binding affinity of the cognate ligand, and $\Delta\Delta G_i$ is the required minimal binding energy gap between the cognate ligand ℓ^{C} and non-cognate ligand i, ℓ_i^{NC}; the degree of specificity required for each non-cognate ligand depends on the *in vivo* concentration,[58] and the cost of incorrect binding.

The difficulty Ψ is a non-linear function, which we expect to generally increase with: the required specificity ($\Delta\Delta G_i$), the number of similar, non-cognate ligands ℓ_i^{NC}, and how easy it is to distinguish them from the cognate ligand ℓ^{C}. It is easier to distinguish: if the cognate ligand is smaller than the non-cognate, as it allows steric exclusion of the larger ligand (Fig. 1B); if the cognate ligand can form more energetic bonds (*e.g.*, an extra hydroxyl group effectively increases ε);[59] or has distinct chemical differences (*e.g.*, positive versus negative charge) to the non-cognate ligands.[60] We lack a general, robust method of quantifying Ψ; creating a metric of ligand discriminability (not the same as similarity[61,62]) would be immensely useful for, *e.g.*, predicting sites for specific inhibition of proteins. For now, we propose that an ad hoc approximation for Ψ could be taken to be the degrees of freedom needed to achieve functional discrimination (Fig. 5A).

Molecular discrimination by a hypothetical mechano-chemical machine. Lock-and-key binding is most specific, but only under very strict conditions (Fig. 1B): The non-cognate ligands must be larger than the cognate ligand (steric exclusion), the cognate ligand must perfectly match the binding site, and both protein and ligand must be rigid. Otherwise, some mismatch between cognate ligand and binding site is always needed to optimise specificity (Fig. 1D) – conformational proofreading is needed to minimise binding to the non-cognate ligand(s).[19] The path to better discrimination still lies with higher rigidity, but increasing control over shape mismatch is needed, since deviations from the optimal mismatch are tolerated less and less as rigidity increases (Fig. 1D). Ultimately, the key feature of good discrimination is precision: the right amount of flexibility, coupled with the right shape, results in optimal deformation.

Molecular discrimination by proteins. Proteins, as genetic, mechano-chemical machines, have some inherent features that strongly constrain their molecular discrimination ability.[59,63] Most bonds in proteins are non-covalent, so proteins cannot be very rigid. Moreover, protein geometry is limited to the topology of a folding chain, composed of discrete units (amino acids) of approximately 1 nm in size, so perfect shape match is practically impossible.

In light of these constraints, mutations very close to a binding site are bound to have a large effect on the flexibility and geometry of the binding site.[64] On the other hand, mutations further from the binding site can have ever smaller effects,[65,66] enabling fine-tuning of mechanics (Fig. 3),[67–71] and structure.[72,73] While we do not model protein folding in this study, we studied the effect of single mutations in proteins in the protein data bank (PDB), finding structural perturbations far from the mutated residue (SI Fig 4). This is consistent with our assertion that far-away mutations can influence molecular discrimination.

These observations, taken together with the finding that good discrimination necessitates precision, lead us to propose that larger proteins – as they have more degrees of freedom (potential mutations) – can achieve better discrimination through finer control over protein dynamics and structure at the binding site (Fig. 5B). For a given discrimination task there is a minimum protein size (Fig. 4C-D), but such an efficient protein may be difficult to evolve. As protein size grows beyond the bare minimum, there are ever more sequences capable of solving the problem, which results in sequences that are evolvable and robust (Fig. 4C-D). The proposed theory thus predicts that proteins are just large enough (to achieve discrimination), and then some more (due to evolution) (Fig. 5A).

Tuning affinity and specificity via energy and entropy. Proteins have an optimal affinity range for cognate ligand(s),[74] and require some degree of specificity for functionality. Here we discuss how these can be controlled by varying protein flexibility, shape, chemistry, and entropy.

To this end, we plot a phase diagram of our results (Fig. 5C-D). Affinity (ΔG) depends non-linearly on chemical bond strength, shape mismatch, pocket geometry (open / closed), flexibility, and entropy (Fig. 5C). The number, and strength of potential chemical bonds (represented in the model by ε) sets an upper limit to affinity (see "Limitations" section for counter-examples not covered here), which is attained when a ligand fits perfectly into a rigid protein pocket. For rigid proteins, shape mismatch sharply reduces affinity, but still results in zero-deformation partial binding, which is stronger in proteins with open binding pockets. Flexibility can both increase and decrease affinity: flexible proteins have higher conformational entropy, which results in higher binding entropy;[75] if there is some shape mismatch between the ligand and the binding pocket, flexibility increases affinity by enabling the protein to adopt an optimal conformation for binding. This behaviour is clearly demonstrated by our model (Fig. 5C), and can be further rationalised in a simple phenomenological equation,

$$\Delta G\left(\varepsilon, \Delta\theta_{\mathrm{LS}}, K\right) = -\varepsilon_{\max}\left[\alpha + (1-\alpha) f_{\mathrm{deform}}\right] - \Delta S ,$$

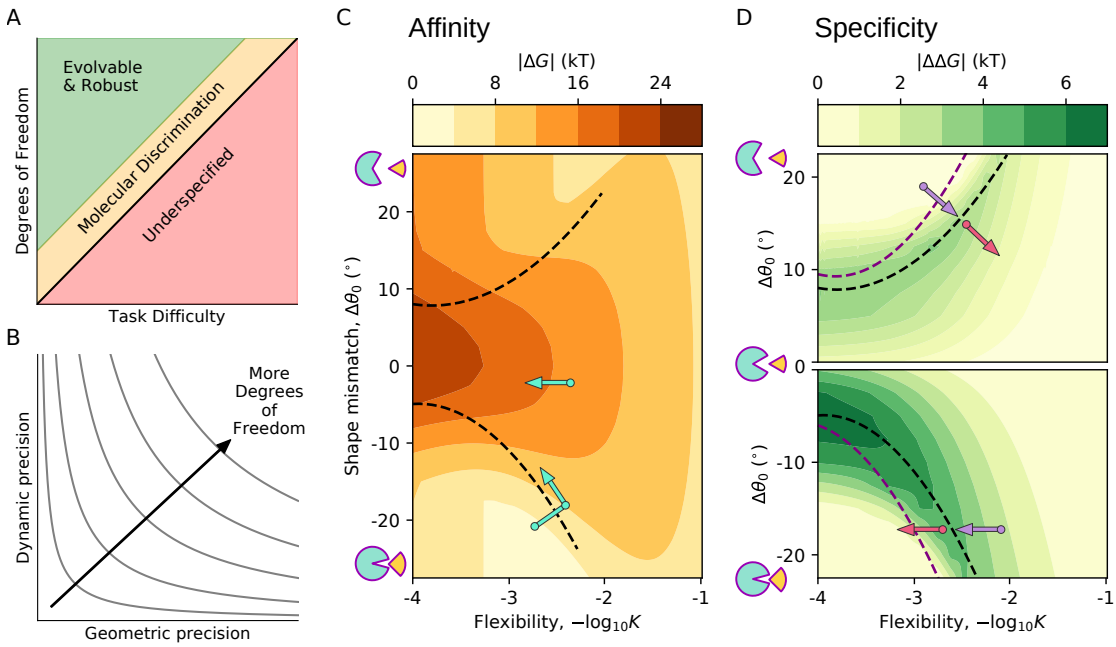

FIG. 5. Theory of biomolecular discrimination. A: More difficult discrimination tasks require more degrees of freedom, which in proteins corresponds to longer sequences. Sequences that are longer than the minimum necessary size are more robust and evolvable. B: Larger proteins have more degrees of freedom with which to fine-tune structure and dynamics at the binding site, since mutations at distal residues can have small yet significant effects. C-D: Binding affinity (C) and specificity (D) as a function of protein flexibility, and shape mismatch between cognate ligand and binding pocket ($\varepsilon = 8\,\mathrm{kT}$, $|\Delta\theta_{\mathrm{LS}}| = 5^\circ$, $-\Delta S = 5.5\log_{10}(K_s/K)$, $K_s = 10^4\,\mathrm{kT/nm^2}$). In the top plot (D) the cognate ligand is smaller than the non-cognate, and the binding pocket is open to the ligand; in the bottom plot (D) the cognate ligand is larger than the non-cognate, and the binding pocket is narrow compared to the ligand. The optimal mismatch for a given flexibility is shown by the black line; as ε increases, the new optimum shifts to the left (purple line; more mismatch). Cyan arrows (C) indicate experimental trajectories of antibodies during affinity maturation. Purple arrows (D) indicate experimental trajectories of enzyme evolution (top – esterase, bottom – cytochrome P450), and red arrows indicate predicted trajectories of suggested experiments.

where $\varepsilon_{\max}$ is the maximum possible energy from intermolecular bonds, α accounts for how much energy is attained through zero-deformation binding, and $0 \leq f_{\mathrm{deform}} \leq 1$ is a function of ε, $\Delta\theta_{\mathrm{LS}}$, and K, which equals zero when there is no deformation, and one when the protein completely binds the ligand with negligible deformation energy. Choosing simple functional forms for α and f_{deform}, we can recover the trends shown in Fig. 5C (SI Fig 5).

Specificity ($\Delta\Delta G$) is limited by affinity, as the binding energy gap can only be as large as the binding energy. Maximum specificity for a given ε is achieved by matching flexibility and shape mismatch (Fig. 5D), so that deformation energy is maximized (SI Fig 6).[76] Proteins need to be just flexible enough for the cognate ligand to bind, but not the non-cognate ligand. Too much flexibility (reduces deformation energy) and both will bind;[58] too rigid and either both will partially bind with negligible deformation.

The ability to selectively modulate binding affinity while maintaining specificity is key to controlling genetic circuits.[74] Our model shows that this can be achieved when binding energy and binding entropy can be controlled independently (Fig. 2D). However this can be difficult; for example, flexibility simultaneously affects binding energy and conformational entropy.[75] One example of how this issue is sidestepped is by using *intrinsically-disordered* regions, either through coupled folding and binding,[77,78] or through a disordered region far from the binding site affected allosterically.[79] Other routes to decoupling specificity and affinity relate to the entropy of the ligand, oligomeric complex,[80] or solvent.[81,82]

Examining and predicting experimental trends. Affinity and specificity vary non-monotonically with $\Delta\theta_{\mathrm{LS}}$, K, and ε, so there are no simple, universal trends. We can, however, discuss trends found in the experimental literature in the context of Fig. 5C-D. For each of these we offer explanations that are consistent with our model, leading to testable predictions.

Many studies report that germline antibodies are flexible, and become more rigid in a process known as affinity maturation.[83,84] A recent *in silico* directed evolution study corroborated this, and also showed that some antibodies first become more flexible, before

later becoming more rigid.[85] According to our model (cyan arrows, Fig. 5C), the former case should occur when antibodies have close to optimal shape mismatch (low $|\Delta\theta_{LS}|$); the latter case should occur when antibodies have high shape mismatch (high $|\Delta\theta_{LS}|$). This prediction can be tested by measuring shape mismatch: one can use measures based on static structure,[86–88] but for better results one should calculate how often the optimal binding configuration is sampled in the free antibody, using molecular dynamics simulations.[89–92]

Our model supports the notion that ancestral enzymes were both flexible and promiscuous,[93] as flexibility often correlates with specificity in our model. However, we find that the correlation between flexibility and specificity depends on shape and chemical binding energy (Fig. 5D). This can be illustrated by two case studies that relate conflicting accounts of the role of flexibility in enzyme promiscuity: Flexibility and promiscuity are correlated in a group of 57 human cytochrome P450 (CYT) enzymes,[94] while the opposite trend is observed in a group of 147 esterases.[95] In the case of CYT, we know that the binding pocket is quite small ($\Delta\theta_0 < 0$), so we can infer that the proteins fall to the right of the optimal line (bottom purple arrow, Fig. 5D). Thus, our model predicts that as rigidity increases, both specificity and affinity will eventually decrease as the protein will be too stiff to deform (bottom red arrow, Fig. 5D). This can be tested by increasing CYT stiffness via directed evolution.[96] In contrast to CYT, the esterases have open active sites ($\Delta\theta_0 > 0$), and we know that promiscuity is correlated with the volume of the active site,[27,60] so we can infer that the proteins fall to the left of the optimal line (top purple arrow, Fig. 5D). In this case, further increasing esterase flexibility via directed evolution should reveal that there is an optimal range of flexibility where specificity is maximized (top red arrow, Fig. 5D).

Our model explains how affinity and specificity can be either positively[97,98] or negatively[99] correlated within a set of proteins, depending on how they differ in shape mismatch and flexibility (SI Fig 6). For example, affinity and specificity are positively correlated in two cases: when proteins differ along optimal line (black line, Fig. 5C-D), or orthogonal to the optimal line. In the former case, deformation energy decreases when affinity and specificity increase; in the latter case, deformation energy increases when affinity and specificity increase (SI Fig 6). This prediction may be tested by studying the transcription factor Pho4, where increased binding affinity of the cognate CACGTG nucleotide sequence was found to improve discrimination of the cognate over the non-cognate CACGTT sequence in 210 variants.[73] These variants can be studied using molecular dynamics simulations, where deformation energy in our model is analogous to the change in internal energy of a protein upon binding[100–102].

A challenge in testing these predictions is the vast amount of data that is needed, since one needs to measure multiple dimensions for a combination of proteins and ligands. However, multiple methods can characterise shape mismatch, flexibility, chemical bond energy, deformation energy, and entropy. We advocate combining molecular dynamics simulations (which can characterise flexibility and calculate deformation energy) with high-throughput experiments (which can measure binding kinetics), and to develop methods to control for orthogonal effects such as differences in protein stability or foldability.[72] Existing public data sets from previous experiments present an facile opportunity in this regard.[1,2,26,27,94]

Molecular discrimination by aminoacyl-tRNA synthetases. To evaluate the theory that protein size depends on task difficulty, we need to know both the relevant non-cognate ligands and necessary binding specificity. In the case of aminoacyl-tRNA synthetases (ARSs), we know the relevant ligands (the 20-30 proteogenic and non-proteogenic amino acids present in cells), and that they have similar *in vivo* concentrations and similar costs associated with missense mutations. This presents a natural control, such that task difficulty is reduced to a question of discriminability between cognate and non-cognate ligands. This is still difficult to evaluate, but we can start by using the available experimental data on pairwise binding specificity of ARSs. We can rationalise that one out of a pair is easier to recognize if it is smaller by a methyl group (steric exclusion), or has an extra hydroxyl group (can form more high energy bonds). Thus, it is difficult to discriminate: threonine from serine (minus one methyl), isoleucine from valine (minus one methyl), phenylalanine from tyrosine (plus one hydroxyl), and alanine from serine (plus one hydroxyl).[30]

Comparing ARSs of these pairs, we find that the ARS of the easier-to-recognize ligand has greater specificity and, with the exception of Val-Ile, they are also smaller (Fig. 6). This exception may be due to the difficulty in general in distinguishing between many aliphatic amino acids, a point illustrated by the fact that these ARSs all have post-transfer editing domains.[103] Furthermore, when we compare ARS enzymes to non-ARS enzymes that also act on amino acids (but with lower specificity requirements), the non-ARS enzymes tend to be considerably smaller than ARSs (SI Fig 7).[30] These findings support the theory that protein size is a product of both the difficulty of the discrimination task. We expect that ARSs are an exemplary class with which to further evaluate theories on protein specificity.

Limitations. By focusing on binding energy gap we have posed the molecular discrimination as a thermo-

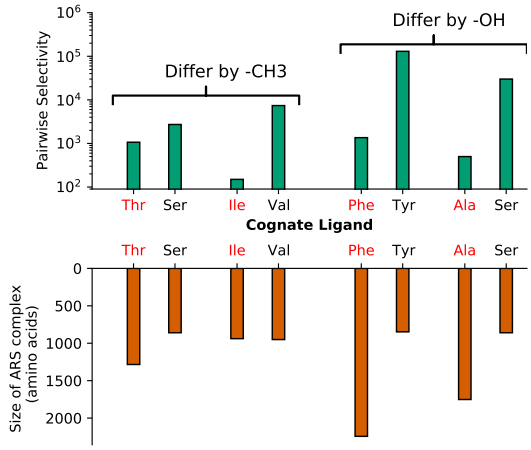

FIG. 6. Pairwise selectivity – ratio of cognate vs non-cognate K_M/k_{cat} – of aminoacyl-tRNA synthases (ARSs), and ARS complex size, of pairs of similar amino acid ligands, where in each pair the labels are coloured according to whether it should be easier (black) or more difficult (red) to discern. Ligands differ by one methyl group (left, ligand with additional $-CH_3$ in red) or one hydroxyl group (right, additional $-OH$ in black).

dynamic problem. It is not immediately clear whether a kinetics-focused approach would lead to the same conclusions. For example, we find that increasing rigidity can increase specificity, but from a kinetics point of view, one might expect the opposite. There is evidence to suggest that flexibility aids formation of initial encounter complexes, which reduces the rate of futile encounters.[9,13,104] Understanding the role of kinetics in molecular discrimination warrants a separate, focused study.

Our model treats deformations as elastic, so it may not generalize well to proteins that undergo plastic deformations upon binding. Furthermore, a key assumption in elastic network models is that bonds are at their equilibrium lengths in the ensemble-average protein structure. In reality, the restructuring of intramolecular bonds can result in a gain in entropy[22,105], or an increase in internal enthalpy.[106] Future work may begin to unravel the role of plastic deformations in biomolecular discrimination, but it will need a significantly more complex model.

Why are proteins so big? Proteins are large macromolecules. Large proteins require large genomes, and thus slow transcription and translation. Prokaryotic proteins, which have prioritize small genomes and fast replication, should benefit immensely from smaller proteins. We do see that proteins are on average smaller in prokaryotes (312 residues) than in eukaryotes (441 residues),[107] but they are still quite large. Thus, there is some indispensable functional reason for protein size. We have here proposed that the difficulty of evolving proteins that can discriminate necessitates proteins of a certain size. We now discuss some alternative factors affecting protein size.

Stability is typically a base requisite for a functional protein,[32] but most proteins are marginally stable – *i.e.*, stability is not maximised, but rather an acceptable level is reached.[63] Longer proteins can more easily fulfill these requirements,[108] but short sequences are often sufficiently stable.[109] Catalytic activity in enzymes is extremely important, but comparable activity levels can be found in much smaller organic catalysts.[110,111] Large surface area may be needed to make multiple interaction sites, whether for a single molecule,[112] or multiple molecules.[31,32,83] While we propose that size is determined by the difficulty of molecular discrimination, all of the above constraints may also be important factors.

Acknowledgements. We thank Jacques Rougemont for discussions and for providing code to calculate shear. This work was supported by the Institute for Basic Science, Project Code IBS-R020-D1. JPE is partially supported by the Fonds National Suisse Swissmap.

METHODS

A model of mechano-chemical binding of a protein to a ligand. Our simple model of a protein is designed in $d = 2$ dimensions, comprising $N_A \leq 22$ amino acids (a_i) and $N_B = 3$ chemical binding sites, where each amino acid represents a contiguous sub-unit wherein the dynamics are highly correlated, so they effectively move as a single unit. Likewise, the binding sites can be thought of as a subdivision of a binding site into the three most salient units (*i.e.*, amino acids, functional groups).[76] Opposite the protein is a set of Λ-shaped ligands with three binding sites, which are uniquely described by an angle, θ. For each protein-ligand pair we calculate the free energy of binding as the sum over the chemical energy E_{chem}, the deformation energy E_{def}, and change in entropy upon binding ΔS, $\Delta G = E_{chem} + E_{def} - T\Delta S$, where T is temperature.

We model the protein as an elastic network where nearest neighbours are connected by harmonic springs, so the deformation can be calculated as,

$$E_{def} = \sum_{i=1}^{N_A} \sum_{j=1}^{N_i} K_{ij} (\Delta r_{ij} - l_{ij})^2 ,$$

where K_{ij} is the spring constant for the bond between amino acids a_i and a_j, N_i is the number of neighbours for a_i, Δr_{ij} is the distance between a_i and a_j, and l_{ij} is the equilibrium distance between a_i and a_j; we use harmonic bond lengths of $l_{ij} = 1$ nm; in the initial configuration $E_{def} = 0$. The chemical binding energy is given by

$$E_{chem} = \sum_{i=1}^{N_B} \varepsilon_i e^{|r_i^p - r_i^b|^2/\sigma^2} ,$$

where ε_i is the energy scale of binding site i, r_i^p and r_i^b are the positions, respectively, of the amino acids a_i and ligand binding site i, and we set the length scale of the interaction at $\sigma = 0.3$ nm. We calculate the *global* minimum binding energy using a gradient descent algorithm with some simple physical constraints. Calculating this binding energy for multiple ligands gives us the binding energy gap between

two ligands, L and S, $\Delta\Delta G = \Delta G_{\mathrm{L}} - \Delta G_{\mathrm{S}}$, the most common quantity used to determine molecular discrimination.

A genetic, mechano-chemical model of molecular recognition. We define a sequence of length N (N_A amino acids, and $N_{\mathrm{B}} = 3$ binding sites) which determines the mechanical and chemical interactions. Each amino-acid can mutate to one of two choices ($\mathbf{w}$ and $\mathbf{s}$) and each chemical binding site can be either weak ($\mathbf{w}_c$) or strong ($\mathbf{s}_c$). Variations in protein sequence only affect protein mechanics and chemical binding. The amino-acid part of the sequence determines the mechanical forces: each neighboring pair of $\mathcal{A}$ is coupled by a harmonic spring with spring constant $K_s > K_m > K_w$; any pair which is linked by a spring has a constant K_w, K_m, K_m, K_s, respectively when the pair is **ww**, **ws**, **sw**, **ss**. Similarly, each of the 3 ligand positions make chemical bonds with one of the 3 sites. The strength of these bonds is given by $\varepsilon_s > \varepsilon_w$ depending on the sequence; the binding constant is ε_s or ε_w respectively for $\mathbf{w}_c$ or $\mathbf{s}_c$. The largest protein studied has $N_A = 22$, so the world of such proteins therefore has at most $2^{22} \cdot 2^3$ possibilities. This size restriction allows us to compute all quantities for the entire genome.

Calculating binding entropy. Binding will typically involve a loss in entropy. A protein with high entropy in the unbound state has more potential to lose entropy upon binding – *i.e.*, flexible proteins should incur a greater entropic cost than rigid ones. To calculate the change in conformational entropy upon binding, we calculate the entropy of a spring network, before and after binding. We assume that the entropy of an elastic network can be approximated by the sum of contributions from independent normal modes.

The elastic energy is given by $\mathcal{E} = \frac{1}{2}\langle u|\,\mathbf{H}\,|u\rangle$, where $|u\rangle$ is the displacement vector (of size $d \cdot N_A$) and $\mathbf{H}$ is the elasticity matrix (for details see Dutta *et al.*[113]). Spectral decomposition of $\mathbf{H}$ gives $\mathbf{H} = \sum_i \lambda_i\, |u_i\rangle\langle u_i|$, where $|u_i\rangle$ are the normal modes and λ_i are the eigenvalues. There are $d(d+1)/2 = 3$ zero energy modes of translation and rotation without deformation that we ignore. We can then express the energy in normal mode coordinates, $\xi_i = \langle u_i|u\rangle$, as $\mathcal{E} = \frac{1}{2}\sum_i \lambda_i \xi_i^2$.

To estimate the entropy, we consider the partition function of the elastic deformations $\mathcal{Z} = \int \mathrm{d}\,|u\rangle\, e^{-\beta\mathcal{E}}$, where β is the inverse temperature. Expressed in normal coordinates, we get

$$\mathcal{Z} = \prod_i \int \mathrm{d}\xi_i\, e^{-\frac{\beta}{2}\lambda_i \xi_i^2} = \prod_i \left(\frac{2\pi}{\beta\lambda_i}\right)^{\frac{1}{2}} ,$$

where the product is only taken over the normal modes for which $\lambda_i \neq 0$. We can calculate the entropy S from the relation $S = \ln\mathcal{Z} - \partial\ln\mathcal{Z}/\partial\ln\beta$, from which we find that $S = \frac{1}{2}\sum_i [\ln(2\pi e) - \ln(\beta\lambda_i)]$. It follows that the binding entropy is simply sum over the changes in the logarithms of the elastic mode energies before and after binding,

$$\Delta S = -\tfrac{1}{2}\sum_i \Delta \ln\lambda_i \; .$$

Typically, the mode energies λ_i increase or stay unchanged upon binding since the motion is more constrained, and therefore the entropy is reduced $\Delta S < 0$. To approximate ΔS, we make the binding site more rigid via the following procedure.[76] We add a new bond between the two a_i at the opening of the binding site, so that the three amino acids at the binding site form a triangle of bonds; we then increase the spring constants of these three bonds to $K_l = 10K_s$. This results in a binding entropy cost ranging $5.5\,\mathrm{kT} \leq -\Delta S \leq 10.2\,\mathrm{kT}$, from the least to the most flexible protein sequence.

The coarse-grained approach to entropy in this work relies on two simplifying assumptions: entropy decreases upon binding, and flexible proteins lose more conformational entropy than rigid proteins. Flexible proteins will of course have more entropy, and thus more to lose. Proteins are most often found to lose entropy upon binding, but many proteins gain entropy due to allosteric conformational change.[105] Moreover, there are other contributions to entropy, such as solvent entropy and ligand entropy,[114–116] that are not treated in our model (contributions of translational and rotational entropy do not depend on the internal degrees of freedom and can be included as a constant factor). Incorporating these details is beyond the scope of the model presented here.

Connecting genes to structure. We allow neighbouring $\mathbf{w}$ and $\mathbf{s}$ amino acids to interact such that they either attract, repel, or neither: attractions (repulsions) between two amino acids result in their equilibrium positions moving closer together (further apart) by $\delta\mathbf{r} = \alpha$ ($\delta\mathbf{r} = -\alpha$). We generate interaction tables $\alpha(a_i, a_j)$ such that $\mathbf{w}$–$\mathbf{w}$, $\mathbf{w}$–$\mathbf{s}$ / $\mathbf{s}$–$\mathbf{w}$, and $\mathbf{s}$–$\mathbf{s}$ bonds, respectively result in one of the three possible interactions. Thus, we can generate up to 15 unique sets of interactions (by accounting for symmetries). We show results for one set in Fig. 4; we verified that the results do not depend on a particular set, and show results for another set in SI Fig. 8. The new equilibrium position $\mathbf{r}_i$ of each amino acid is then determined by the original equilibrium position $\mathbf{r}_i^0$ plus the sum of all of the neighbouring vectors,

$$\mathbf{r}_i = \mathbf{r}_i^0 + \sum_j \alpha(a_i, a_j)\,\frac{\mathbf{r}_j^0 - \mathbf{r}_i^0}{|\mathbf{r}_j^0 - \mathbf{r}_i^0|} \; ,$$

where j is the index of a bonded neighbour.

The optimization algorithm. For fixed coupling constants, we find a configuration which optimizes energy. After moving the protein close to the ligand, we use the L-BFGS method[117] to find a minimum. To model steric repulsion, we restrict the positions of the amino acids so that they are outside the sector defined by the ligand position and angle θ. Two post-processing checks precede the analysis: We require that the orientation of each triangle in the protein maintains its orientation, *i.e.*, that the surface is not flipped, and the minimized energy must be at least as low as that obtained by a completely rigid protein. We discard any results that do not pass these tests.

Evolvability and robustness. *Evolvability* is the ability of a population of organisms to evolve new phenotypes. We measure this as the maximum number of mutations a fit sequence can accumulate without having fitness less than zero. *Robustness* is the ability of a sequence to mutate while retaining its original function. We measure robustness as the minimum number of mutations needed for a fit sequence to become non-fit. In practice, we calculated the edit distance between fit sequences, and clustered them using a density-based clustering algorithm (DBSCAN, implemented in the sklearn python module; eps = 1, min_samples = 1).[118] Evolvability of a sequence is then the maximum distance within a cluster, while robustness is the minimum distance within a cluster. For both measures we averaged over all fit sequences (according to Eq.(1)), and normalize by dividing by the genome size.

* tsvitlusty@gmail.com

1. H. Nam, N. E. Lewis, J. A. Lerman, D.-H. Lee, R. L. Chang, D. Kim, and B. O. Palsson, Network context and selection in the evolution to enzyme specificity, Science **337**, 1101 (2012).
2. I. Piazza, K. Kochanowski, V. Cappelletti, T. Fuhrer, E. Noor, U. Sauer, and P. Picotti, A map of protein-metabolite interactions reveals principles of chemical communication, Cell **172**, 358 (2018).
3. S. D. Copley, The physical basis and practical consequences of biological promiscuity, Phys. Biol. **17**, 051001 (2020).
4. D. W. Miller and K. A. Dill, Ligand binding to proteins: The binding landscape model, Protein Sci. **6**, 2166 (1997).
5. J. A. McCammon, Theory of biomolecular recognition, Curr. Opin. Struc. Biol. **8**, 245 (1998).
6. B. Ma, S. Kumar, C. J. Tsai, and R. Nussinov, Folding funnels and binding mechanisms, Protein Engineering, Design and Selection **12**, 713 (1999).
7. H. Gohlke and G. Klebe, Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors, Angewandte Chemie International Edition **41**, 2644 (2002).
8. D. L. Mobley and K. A. Dill, Binding of small-molecule ligands to proteins: "what you see" is not always "what you get", Structure **17**, 489 (2009).
9. Q. Wang, P. Zhang, L. Hoffman, S. Tripathi, D. Homouz, Y. Liu, M. N. Waxham, and M. S. Cheung, Protein recognition and selection through conformational and mutually induced fit, P. Natl. Acad. Sci. Usa. **110**, 20545 (2013).
10. P. L. Kastritis and A. M. J. J. Bonvin, On the binding affinity of macromolecular interactions: Daring to ask why proteins interact, J. Roy. Soc. . Interface **10**, 20120835 (2013).
11. T. R. Weikl and F. Paul, Conformational selection in protein binding and function, Protein Sci. **23**, 1508 (2014).
12. A. Stank, D. B. Kokh, J. C. Fuller, and R. C. Wade, Protein binding pocket dynamics, Accounts Chem. Res. **49**, 809 (2016).
13. N. Plattner, S. Doerr, G. De Fabritiis, and F. Noé, Complete protein–protein association kinetics in atomic detail revealed by molecular dynamics simulations and markov modelling, Nat. Chem. **9**, 1005 (2017).
14. A. Goldstein, The mechanism of enzyme-inhibitor-substrate reactions: Illustrated by the cholinesterase-physostigmine-acetylcholine system, The Journal of general physiology **27**, 529 (1944).
15. H.-C. Mahler, W. Friess, U. Grauschopf, and S. Kiese, Protein aggregation: Pathways, induction factors and analysis, J. Pharm. Sci-us. **98**, 2909 (2009).
16. M. F. Olson and R. Marais, Ras protein signalling, Semin. Immunol. **12**, 63 (2000).
17. R. Rohs, X. Jin, S. M. West, R. Joshi, B. Honig, and R. S. Mann, Origins of specificity in protein-dna recognition, Annu. Rev. Biochem. **79**, 233 (2010).
18. A. P. Sergeeva, P. S. Katsamba, F. Cosmanescu, J. J. Brewer, G. Ahlsen, S. Mannepalli, L. Shapiro, and B. Honig, Dip/dpr interactions and the evolutionary design of specificity in protein families, Nat. Commun. **11**, 2125 (2020).
19. Y. Savir and T. Tlusty, Conformational proofreading: The impact of conformational changes on the specificity of molecular recognition, Plos One **2**, 1 (2007).
20. O. Rivoire, Parsimonious evolutionary scenario for the origin of allostery and coevolution patterns in proteins, Phys. Rev. E **100**, 032411 (2019).
21. J. P. Eckmann, J. Rougemont, and T. Tlusty, Colloquium: Proteins: The physics of amorphous evolving matter, Rev. Mod. Phys. **91**, 031001 (2019).
22. S. A. Wankowicz, S. H. P. de Oliveira, D. W. Hogan, H. van den Bedem, and J. S. Fraser, Ligand binding remodels protein side chain conformational heterogeneity, bioRxiv 10.1101/2021.09.21.461269 (2021).
23. E. J. Sundberg and R. A. Mariuzza, Molecular recognition in antibody-antigen complexes, in *Protein Modules and Protein-Protein Interaction*, Vol. 61 (Academic Press, 2002) pp. 119–160.
24. M. Elias, A. Wellner, K. Goldin-Azulay, E. Chabriere, J. A. Vorholt, T. J. Erb, and D. S. Tawfik, The molecular basis of phosphate discrimination in arsenate-rich environments, Nature **491**, 134 (2012).
25. E. Jankowsky and M. E. Harris, Specificity and nonspecificity in rna–protein interactions, Nat. Rev. Mol. Cell Bio. **16**, 533 (2015).
26. H. Huang, C. Pandya, C. Liu, N. F. Al-Obaidi, M. Wang, L. Zheng, S. Toews Keating, M. Aono, J. D. Love, B. Evans, R. D. Seidel, B. S. Hillerich, S. J. Garforth, S. C. Almo, P. S. Mariano, D. Dunaway-Mariano, K. N. Allen, and J. D. Farelli, Panoramic view of a superfamily of phosphatases through substrate profiling, P. Natl. Acad. Sci. Usa. **112**, E1974 (2015).
27. M. Martínez-Martínez, C. Coscolín, G. Santiago, J. Chow, P. J. Stogios, R. Bargiela, C. Gertler, J. Navarro-Fernández, A. Bollinger, S. Thies, C. Méndez-García, A. Popovic, G. Brown, T. N. Chernikova, A. García-Moyano, G. E. K. Bjerga, P. Pérez-García, T. Hai, M. V. Del Pozo, R. Stokke, I. H. Steen, H. Cui, X. Xu, B. P. Nocek, M. Alcaide, M. Distaso, V. Mesa, A. I. Peláez, J. Sánchez, P. C. F. Buchholz, J. Pleiss, A. Fernández-Guerra, F. O. Glöckner, O. V. Golyshina, M. M. Yakimov, A. Savchenko, K. E. Jaeger, A. F. Yakunin, W. R. Streit, P. N. Golyshin, V. Guallar, M. Ferrer, and T. I. Consortium, Determinants and prediction of esterase substrate promiscuity patterns, Acs Chem. Biol. **13**, 225 (2018).
28. T. L. V. Lite, R. A. Grant, I. Nocedal, M. L. Littlehale, M. S. Guo, and M. T. Laub, Uncovering the basis of protein-protein interaction specificity with a combinatorially complete library, eLife **9**, e60924 (2020).
29. A. Biswas and M. Thattai, Promiscuity and specificity of eukaryotic glycosyltransferases, Biochem. Soc. T. **48**, 891 (2020).
30. D. S. Tawfik and I. Gruic-Sovulj, How evolution shapes enzyme selectivity - lessons from aminoacyl-trna synthetases and other amino acid utilizing enzymes, The FEBS Journal **287**, 1284 (2020).
31. T. A. J. Payens, Why are enzymes so large?, Trends Biochem. Sci. **8**, 46 (1983).
32. P. A. Srere, Why are enzymes so big?, Trends Biochem. Sci. **9**, 387 (1984).
33. M. M. Tirion, Large amplitude elastic motions in proteins from a single-parameter, atomic analysis, Phys. Rev. Lett. **77**, 1905 (1996).
34. C. Chennubhotla, A. J. Rader, L. W. Yang, and I. Bahar, Elastic network models for understanding biomolecular machinery: From enzymes to supramolecular assemblies, Phys. Biol. **2**, S173 (2005).
35. J. R. López-Blanco and P. Chacón, New generation of elastic network models, Curr. Opin. Struc. Biol. **37**, 46 (2016).
36. N. Halabi, O. Rivoire, S. Leibler, and R. Ranganathan, Protein sectors: Evolutionary units of three-dimensional structure, Cell **138**, 774 (2009).
37. A. R. Atilgan, S. R. Durell, R. L. Jernigan, M. C. Demirel, O. Keskin, and I. Bahar, Anisotropy of fluctuation dynamics of proteins with an elastic network model, Biophys. J. **80**, 505 (2001).
38. K. Hinsen, Structural flexibility in proteins: Impact of the crystal environment, Method. Biochem. Anal. **24**, 521 (2007).
39. D. Riccardi, Q. Cui, and G. N. Phillips, Application of elastic network models to proteins in the crystalline state, Biophys. J. **96**, 464 (2009).
40. K. K. Frederick, M. S. Marlow, K. G. Valentine, and A. J. Wand, Conformational entropy in molecular recognition by proteins, Nature **448**, 325 (2007).
41. S. R. Tzeng and C. G. Kalodimos, Protein activity regulation by conformational entropy, Nature **488**, 236 (2012).
42. M. Schapira, M. Totrov, and R. Abagyan, Prediction of the binding energy for small molecules, peptides and proteins, J. Mol. Recognit. **12**, 177 (1999).
43. E. Barratt, R. J. Bingham, D. J. Warner, C. A. Laughton, S. E. V. Phillips, and S. W. Homans, Van der waals interactions dominate ligand-protein association in a protein binding site occluded from solvent water, J. Am. Chem. Soc. **127**, 11827 (2005).

44. E. Gallicchio, M. Lapelosa, and R. M. Levy, Binding energy distribution analysis method (bedam) for estimation of protein-ligand binding affinities, J. Chem. Theory Comput. **6**, 2961 (2010).
45. I. H. Moal, R. Agius, and P. A. Bates, Protein-protein binding affinity prediction on a diverse set of structures, Method. Biochem. Anal. **27**, 3002 (2011).
46. Y. Huang and Z. Liu, Do intrinsically disordered proteins possess high specificity in protein-protein interactions?, Chemistry - A European Journal **19**, 4462 (2013).
47. R. Grünberg, M. Nilges, and J. Leckner, Flexibility and conformational entropy in protein-protein binding, Structure **14**, 683 (2006).
48. C. A. Chang, W. Chen, and M. K. Gilson, Ligand configurational entropy and protein binding, P. Natl. Acad. Sci. Usa. **104**, 1534 (2007).
49. Z. Sun, Y. N. Yan, M. Yang, and J. Z. H. Zhang, Interaction entropy for protein-protein binding, The Journal of Chemical Physics **146**, 124124 (2017).
50. R. T. Hietpas, J. D. Jensen, and D. N. A. Bolon, Experimental illumination of a fitness landscape, P. Natl. Acad. Sci. Usa. **108**, 7896 (2011).
51. J. A. G. M. de Visser and J. Krug, Empirical fitness landscapes and the predictability of evolution, Nat. Rev. Genet. **15**, 480 (2014).
52. K. B. Levin, O. Dym, S. Albeck, S. Magdassi, A. H. Keeble, C. Kleanthous, and D. S. Tawfik, Following evolutionary paths to protein-protein interactions with high affinity and selectivity, Nature Structural & Molecular Biology **16**, 1049 (2009).
53. O. K. Tawfik and D. S., Enzyme promiscuity: A mechanistic and evolutionary perspective, Annu. Rev. Biochem. **79**, 471 (2010).
54. A. Wagner, *Robustness and Evolvability in Living Systems* (Princeton university press, 2013).
55. J. F. Kribelbauer, C. Rastogi, H. J. Bussemaker, and R. S. Mann, Low-affinity binding sites and the transcription factor specificity paradox in eukaryotes, Annu. Rev. Cell Dev. Bi. **35**, 357 (2019).
56. F. X. Contreras, A. M. Ernst, P. Haberkant, P. Björkholm, E. Lindahl, B. Gönen, C. Tischer, A. Elofsson, G. von Heijne, C. Thiele, R. Pepperkok, F. Wieland, and B. Brügger, Molecular recognition of a single sphingolipid species by a protein's transmembrane domain, Nature **481**, 525 (2012).
57. O. Rivoire, Geometry and flexibility of optimal catalysts in a minimal elastic model, The Journal of Physical Chemistry B **124**, 807 (2020).
58. M. Gade, L. L. Tan, A. M. Damry, M. Sandhu, J. S. Brock, A. Delaney, A. Villar-Briones, C. J. Jackson, and P. Laurino, Substrate dynamics contribute to enzymatic specificity in human and bacterial methionine adenosyltransferases, bioRxiv 10.1101/2021.02.18.431797 (2021).
59. A. Peracchi, The limits of enzyme specificity and the evolution of metabolism, Trends Biochem. Sci. **43**, 984 (2018).
60. A. Barrozo, F. Duarte, P. Bauer, A. T. P. Carvalho, and S. C. L. Kamerlin, Cooperative electrostatic interactions drive functional evolution in the alkaline phosphatase superfamily, J. Am. Chem. Soc. **137**, 9061 (2015).
61. J. D. Stephenson and S. J. Freeland, Unearthing the root of amino acid similarity, J. Mol. Evol. **77**, 159 (2013).
62. A. Rácz, D. Bajusz, and K. Héberger, Life beyond the tanimoto coefficient: Similarity measures for interaction fingerprints, Journal of Cheminformatics **10**, 48 (2018).
63. T. Sikosek and H. S. Chan, Biophysics of protein evolution and evolutionary protein biophysics, J. Roy. Soc. . Interface **11**, 20140419 (2014).
64. A. Lopes, M. S. A. Busch, and T. Simonson, Computational design of protein-ligand binding: Modifying the specificity of asparaginyl-trna synthetase, J. Comput. Chem. **31**, 1273 (2010).
65. R. M. Adams, J. B. Kinney, A. M. Walczak, and T. Mora, Epistasis in a fitness landscape defined by antibody-antigen binding free energy, Cell Systems **8**, 86 (2019).
66. M. Wang, Z. Cang, and G. W. Wei, A topology-based network tree for the prediction of protein–protein binding affinity changes following mutation, Nature Machine Intelligence **2**, 116 (2020).
67. P. Koenig, C. V. Lee, B. T. Walters, V. Janakiraman, J. Stinson, T. W. Patapoff, and G. Fuh, Mutational landscape of antibody variable domains reveals a switch modulating the interdomain conformational dynamics and antigen binding, P. Natl. Acad. Sci. Usa. **114**, E486 (2017).
68. R. Otten, L. Liu, L. R. Kenner, M. W. Clarkson, D. Mavor, D. S. Tawfik, D. Kern, and J. S. Fraser, Rescue of conformational dynamics in enzyme catalysis by directed evolution, Nat. Commun. **9**, 1314 (2018).
69. T. Modi, V. A. Risso, S. Martinez-Rodriguez, J. A. Gavira, M. D. Mebrat, W. D. Van Horn, J. M. Sanchez-Ruiz, and S. Banu Ozkan, Hinge-shift mechanism as a protein design principle for the evolution of β-lactamases from substrate promiscuity to specificity, Nat. Commun. **12**, 1852 (2021).
70. N. J. Ose, B. M. Butler, A. Kumar, M. Sanderford, S. Kumar, and S. B. Ozkan, Dynamic coupling of residues within proteins as a mechanistic foundation of many enigmatic pathogenic missense variants, bioRxiv 10.1101/2021.09.24.461738 (2021).
71. D. Ray, L. Le, and I. Andricioaei, Distant residues modulate conformational opening in sars-cov-2 spike protein, P. Natl. Acad. Sci. Usa. **118**, 10.1073/pnas.2100943118 (2021).
72. C. J. Markin, D. A. Mokhtari, F. Sunden, M. J. Appel, E. Akiva, S. A. Longwell, C. Sabatti, D. Herschlag, and P. M. Fordyce, Revealing enzyme functional architecture via high-throughput microfluidic enzyme kinetics, Science **373**, eabf8761 (2021).
73. A. K. Aditham, C. J. Markin, D. A. Mokhtari, N. DelRosso, and P. M. Fordyce, High-throughput affinity measurements of transcription factor and dna mutations reveal affinity and specificity determinants, Cell Systems **12**, 112 (2021).
74. A. P. Popp, J. Hettich, and J. C. M. Gebhardt, Transcription factor residence time dominates over concentration in transcription activation, bioRxiv 10.1101/2020.11.26.400069 (2020).
75. H. Cao, Y. Huang, and Z. Liu, Interplay between binding affinity and kinetics in protein-protein interactions, Proteins. **84**, 920 (2016).
76. J. P. Richard, Protein flexibility and stiffness enable efficient enzymatic catalysis, J. Am. Chem. Soc. **141**, 3320 (2019).
77. Y. Wang, X. Chu, S. Longhi, P. Roche, W. Han, E. Wang, and J. Wang, Multiscaled exploration of coupled folding and binding of an intrinsically disordered molecular recognition element in measles virus nucleoprotein, P. Natl. Acad. Sci. Usa. **110**, E3743 (2013).
78. J. Yang, M. Gao, J. Xiong, Z. Su, and Y. Huang, Features of molecular recognition of intrinsically disordered proteins via coupled folding and binding, Protein Sci. **28**, 1952 (2019).
79. N. D. Keul, K. Oruganty, E. T. Schaper Bergman, N. R. Beattie, W. E. McDonald, R. Kadirvelraj, M. L. Gross, R. S. Phillips, S. C. Harvey, and Z. A. Wood, The entropic force generated by intrinsically disordered segments tunes protein function, Nature **563**, 584 (2018).
80. T. Scheidt, J. A. Carozza, C. C. Kolbe, F. A. Aprile, O. Tkachenko, M. M. J. Bellaiche, G. Meisl, Q. A. E. Peter, T. W. Herling, S. Ness, M. Castellana-Cruz, J. L. P. Benesch, M. Vendruscolo, C. M. Dobson, P. Arosio, and T. P. J. Knowles, The binding of the small heat-shock protein αb-crystallin to fibrils of α-synuclein is driven by entropic forces, P. Natl. Acad. Sci. Usa. **118**, 10.1073/pnas.2108790118 (2021).
81. F. A. Quiocho, D. K. Wilson, and N. K. Vyas, Substrate specificity and affinity of a protein modulated by bound water molecules, Nature **340**, 404 (1989).
82. J. Michel, J. Tirado-Rives, and W. L. Jorgensen, Energetics of displacing water molecules from protein binding sites: Consequences for ligand optimization, J. Am. Chem. Soc. **131**, 15403 (2009).
83. M. H. V. Van Regenmortel, Specificity, polyspecificity, and heterospecificity of antibody-antigen recognition, J. Mol. Recognit. **27**, 627 (2014).
84. A. K. Mishra and R. A. Mariuzza, Insights into the structural basis of antibody affinity maturation from next-generation sequencing, Frontiers in Immunology **9**, 117 (2018).
85. V. Ovchinnikov, J. E. Louveau, J. P. Barton, M. Karplus, and

A. K. Chakraborty, Role of framework mutations and antibody flexibility in the evolution of broadly neutralizing antibodies, eLife **7**, e33038 (2018).
86. R. Chen and Z. Weng, A novel shape complementarity scoring function for protein-protein docking, Proteins. **51**, 397 (2003).
87. B. Y. Chen and B. Honig, Vasp: A volumetric analysis of surface properties yields insights into protein-ligand binding specificity, Plos Comput. Biol. **6**, 1 (2010).
88. Y. Yan and S. Y. Huang, Pushing the accuracy limit of shape complementarity for protein-protein docking, Bmc Bioinformatics **20**, 696 (2019).
89. M. Karplus and J. A. McCammon, Molecular dynamics simulations of biomolecules, Nat. Struct. Biol. **9**, 646 (2002).
90. M. Münz, J. Hein, and P. C. Biggin, The role of flexibility and conformational selection in the binding promiscuity of pdz domains, Plos Comput. Biol. **8**, 1 (2012).
91. N. A. Pabon and C. J. Camacho, Probing protein flexibility reveals a mechanism for selective promiscuity, eLife **6**, e22889 (2017).
92. A. S. Kamenik, I. Singh, P. Lak, T. E. Balius, K. R. Liedl, and B. K. Shoichet, Energy penalties enhance flexible receptor docking in a model cavity, P. Natl. Acad. Sci. Usa. **118**, 10.1073/pnas.2106195118 (2021).
93. D. Petrović, V. A. Risso, S. C. L. Kamerlin, and J. M. Sanchez-Ruiz, Conformational dynamics and enzyme evolution, J. Roy. Soc. . Interface **15**, 20180330 (2018).
94. D. Becker, P. V. Bharatam, and H. Gohlke, F/g region rigidity is inversely correlated to substrate promiscuity of human cyp isoforms involved in metabolism, J. Chem. Inf. Model. **61**, 4023 (2021).
95. C. Nutschel, C. Coscolín, B. David, D. Mulnaes, M. Ferrer, K. E. Jaeger, and H. Gohlke, Promiscuous esterases counter-intuitively are less flexible than specific ones, J. Chem. Inf. Model. **61**, 2383 (2021).
96. M. E. Hill, D. J. MacPherson, P. Wu, O. Julien, J. A. Wells, and J. A. Hardy, Reprogramming caspase-7 specificity by regio-specific mutations and selection provides alternate solutions for substrate recognition, Acs Chem. Biol. **11**, 1603 (2016).
97. B. E. Eaton, L. Gold, and D. A. Zichi, Let's get specific: The relationship between specificity and affinity, Chem. Biol. **2**, 633 (1995).
98. F. Gao, J. Glaser, and S. C. Glotzer, The role of complementary shape in protein dimerization, Soft Matter **17**, 7376 (2021).
99. N. S. Greenspan, Cohen's conjecture, howard's hypothesis, and ptashne's ptruth: An exploration of the relationship between affinity and specificity, Trends Immunol. **31**, 138 (2010).
100. C. M. Reyes and P. A. Kollman, Structure and thermodynamics of rna-protein binding: Using molecular dynamics and free energy analyses to calculate the free energies of binding and conformational change, J. Mol. Biol. **297**, 1145 (2000).
101. M. Kankainen, T. Laitinen, and M. Peräkylä, Recognition of reactive high-energy conformations by shape complementarity and specific enzyme–substrate interactions in family 10 and 11 xylanases, Phys. Chem. Chem. Phys. **6**, 5074 (2004).
102. J. Park, J. J. McDonald, R. C. Petter, and K. N. Houk, Molecular dynamics analysis of binding of kinase inhibitors to wt egfr and the t790m mutant, J. Chem. Theory Comput. **12**, 2066 (2016).
103. J. J. Perona and A. Hadd, Structural diversity and protein engineering of the aminoacyl-trna synthetases, Biochemistry-us. **51**, 8705 (2012).
104. A. Bar-Even, R. Milo, E. Noor, and D. S. Tawfik, The moderately efficient enzyme: Futile encounters and enzyme floppiness, Biochemistry-us. **54**, 4969 (2015).
105. A. T. Fenley, H. S. Muddana, and M. K. Gilson, Entropy–enthalpy transduction caused by conformational shifts can obscure the forces driving protein–ligand binding, P. Natl. Acad. Sci. Usa. **109**, 20006 (2012).
106. D. U. Ferreiro, J. A. Hegler, E. A. Komives, and P. G. Wolynes, Localizing frustration in native proteins and protein assemblies, P. Natl. Acad. Sci. Usa. **104**, 19819 (2007).
107. T. U. Consortium, Uniprot: A worldwide hub of protein knowledge, Nucleic Acids Res. **47**, D506 (2018).
108. U. Bastolla and L. Demetrius, Stability constraints and protein evolution: The role of chain length, composition and disulfide bonds, Protein Engineering, Design and Selection **18**, 405 (2005).
109. A. R. Davidson, K. J. Lumb, and R. T. Sauer, Cooperatively folded proteins in random sequence libraries, Nat. Struct. Biol. **2**, 856 (1995).
110. T. H. Maugh, Catalysts that break nature's monopoly: Chiral complexes can approach the specificity of enzymes for synthesis of optically active compounds, and can act on a wider variety of substrates, Science **221**, 351 (1983).
111. D. W. C. MacMillan, The advent and development of organocatalysis, Nature **455**, 304 (2008).
112. M. J. Henley, B. M. Linhares, B. S. Morgan, T. Cierpicki, C. A. Fierke, and A. K. Mapp, Unexpected specificity within dynamic transcriptional protein–protein complexes, P. Natl. Acad. Sci. Usa. **117**, 27346 (2020).
113. S. Dutta, J. P. Eckmann, A. Libchaber, and T. Tlusty, Green function of correlated genes in a minimal mechanical model of protein evolution, P. Natl. Acad. Sci. Usa. **115**, E4559 (2018).
114. Y. Savir and T. Tlusty, Optimal design of a molecular recognizer: Molecular recognition as a bayesian signal detection problem, Ieee J-stsp. **2**, 390 (2008).
115. A. I. Dragan, C. M. Read, and C. Crane-Robinson, Enthalpy–entropy compensation: The role of solvation, Eur. Biophys. J. **46**, 301 (2017).
116. F. Peccati and G. Jiménez-Osés, Enthalpy–entropy compensation in biomolecular recognition: A computational perspective, ACS Omega **6**, 11122 (2021).
117. R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu, A limited memory algorithm for bound constrained optimization, Siam J. Sci. Comput. **16**, 1190 (1995).
118. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. **12**, 2825 (2011).

Supporting information for "Theory of specific binding: insights from a genetic, mechano-chemical protein model"

John M. McBride[1], Jean-Pierre Eckmann[2], and Tsvi Tlusty[1,3,*]

[1]Center for Soft and Living Matter, Institute for Basic Science, Ulsan 44919, South Korea
[2]Département de Physique Théorique and Section de Mathématiques, University of Geneva, Geneva, Switzerland
[3]Departments of Physics and Chemistry, Ulsan National Institute of Science and Technology, Ulsan 44919, South Korea

[*]tsvitlusty@gmail.com

February 23, 2022

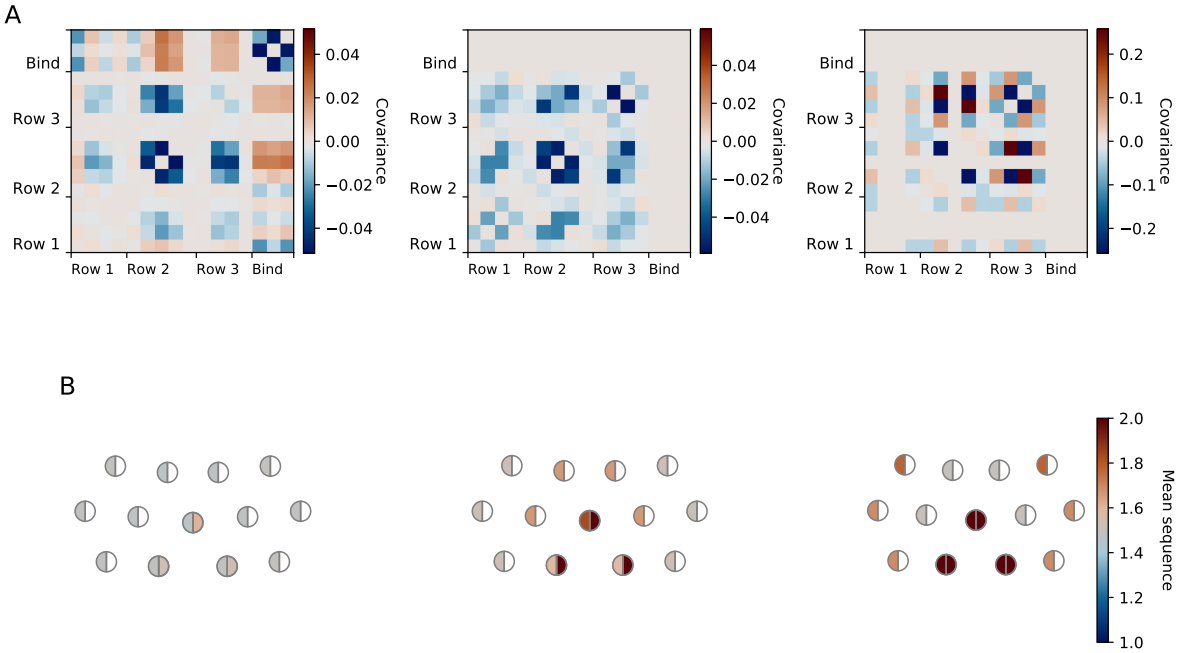

Figure 1: Sequence covariance (A) and average sequence (B) for all loose-recognition solutions (left), for solutions with $|\Delta\Delta G| > 3.6$ kT (middle), and for solutions with $|\Delta\Delta G| > 3.8$ kT (right). Left-filled circles indicate the sequence element that determines the spring constant; the three right-filled circles at the binding site indicate the sequence element that determines binding strength. For this example (same as in the main text), the parameters are: $K_s = 1000$, $K_m = 400$, $K_w = 100$ kT / nm^2; $\varepsilon_s = 6$, $\varepsilon_w = 1.75$ kT; $\theta_A = 80°$, $\theta_B = 90°$.
Out of all the loose-recognition solutions, there is a slight tendency for weaker springs, and stronger chemical binding, compared to a random set of sequences. For solutions with $|\Delta\Delta G| > 3.6$ kT, all of the sequences have maximal chemical interaction strength ($\mathbf{s}_c$), and they tend to have a rigid core. For solutions with $|\Delta\Delta G| > 3.8$ kT, all of the sequences have maximal chemical interaction strength ($\mathbf{s}_c$), and maximally rigid binding site; in this case, the exterior tends to be rigid rather than the core.

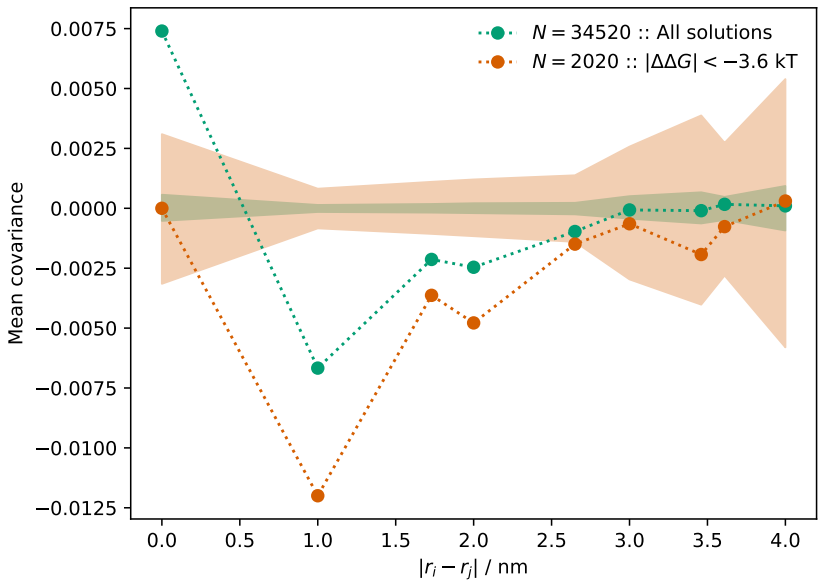

Figure 2: Sequence covariance between residues as a function of distance between residues, respectively, for all loose-recognition solutions, and for solutions with $|\Delta\Delta G| > 3.6$ kT. Shaded regions show the expected value and standard deviation for sets of N sequences chosen randomly without replacement.

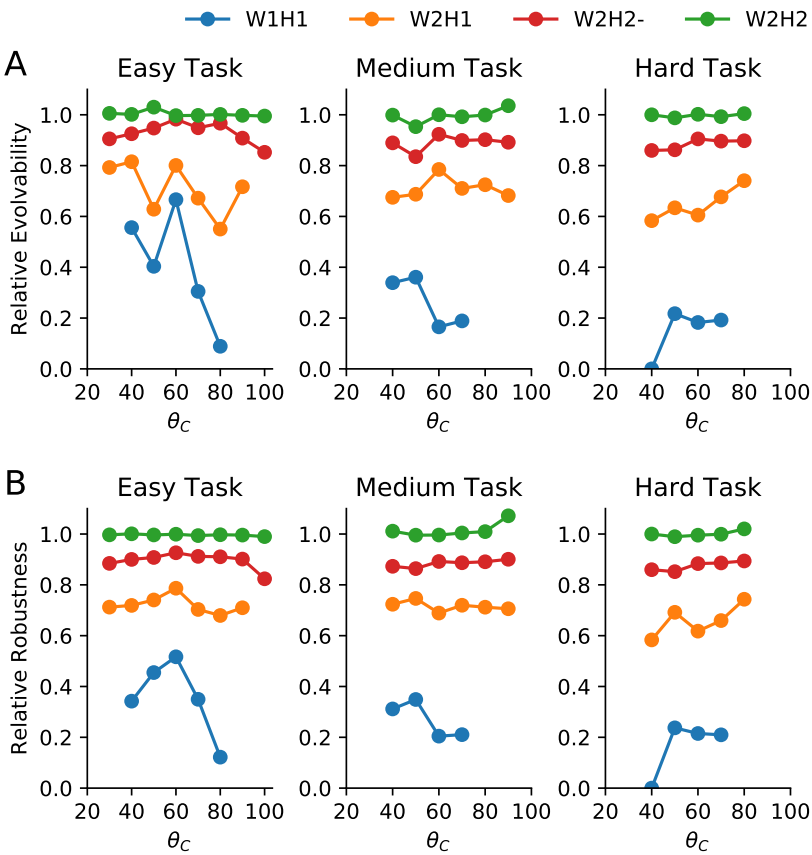

Figure 3: Evolvability (A) and robustness (B) relative to the largest protein (W2H3-), as a function of the cognate ligand shape, θ_C, for easy ($\phi = 10$), medium ($\phi = 20$) and hard ($\phi = 30$) tasks. Colours indicate models of different size.

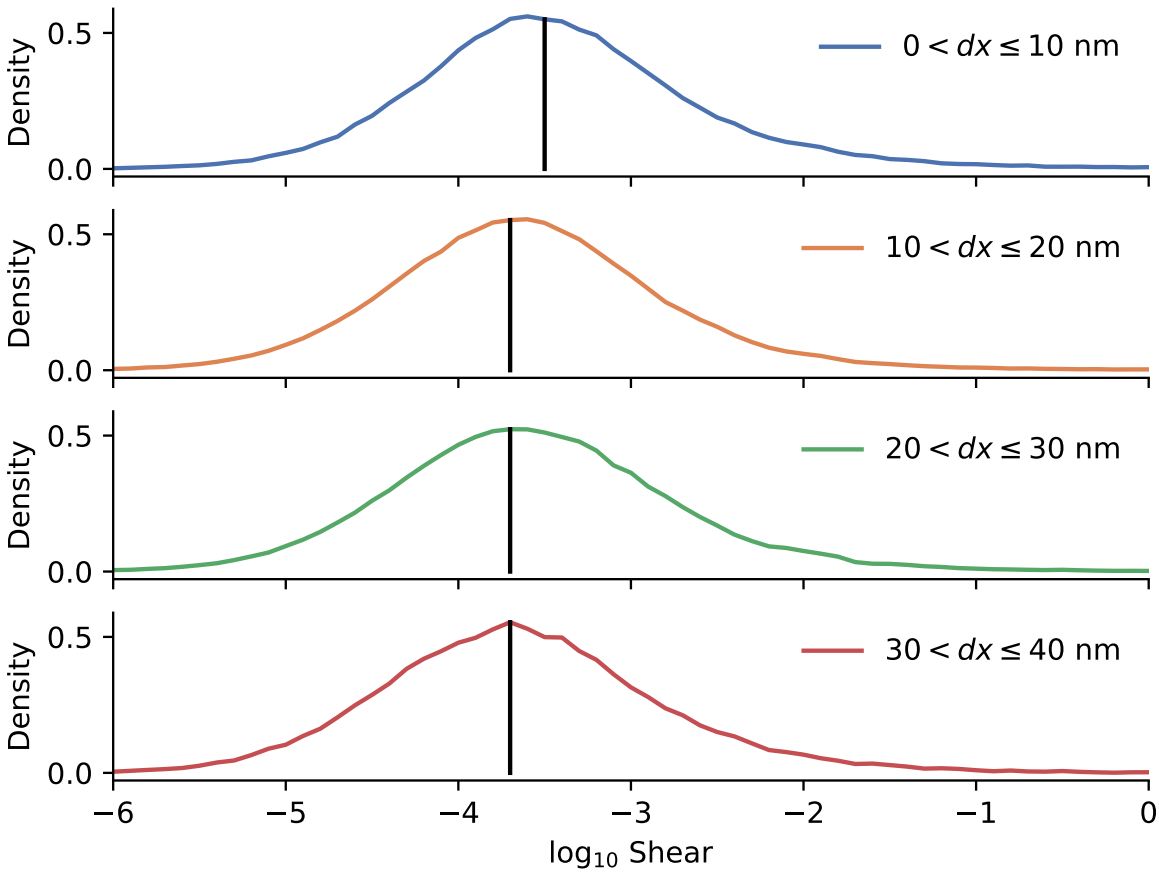

Figure 4: Effect of amino acid mutations on protein structure. We measure the shear deformation of C_α position that results from mutating a single amino acid, and plot the distribution of shear as a function of distance from the mutated residue. Black line indicates median. We calculated this for 6,739 pairs of proteins taken from the Protein Data Bank [1], where we excluded protein-protein and protein-nucleotide complexes, and controlled for presence of bound molecules. We calculated shear as detailed in Eckmann et al. [2].

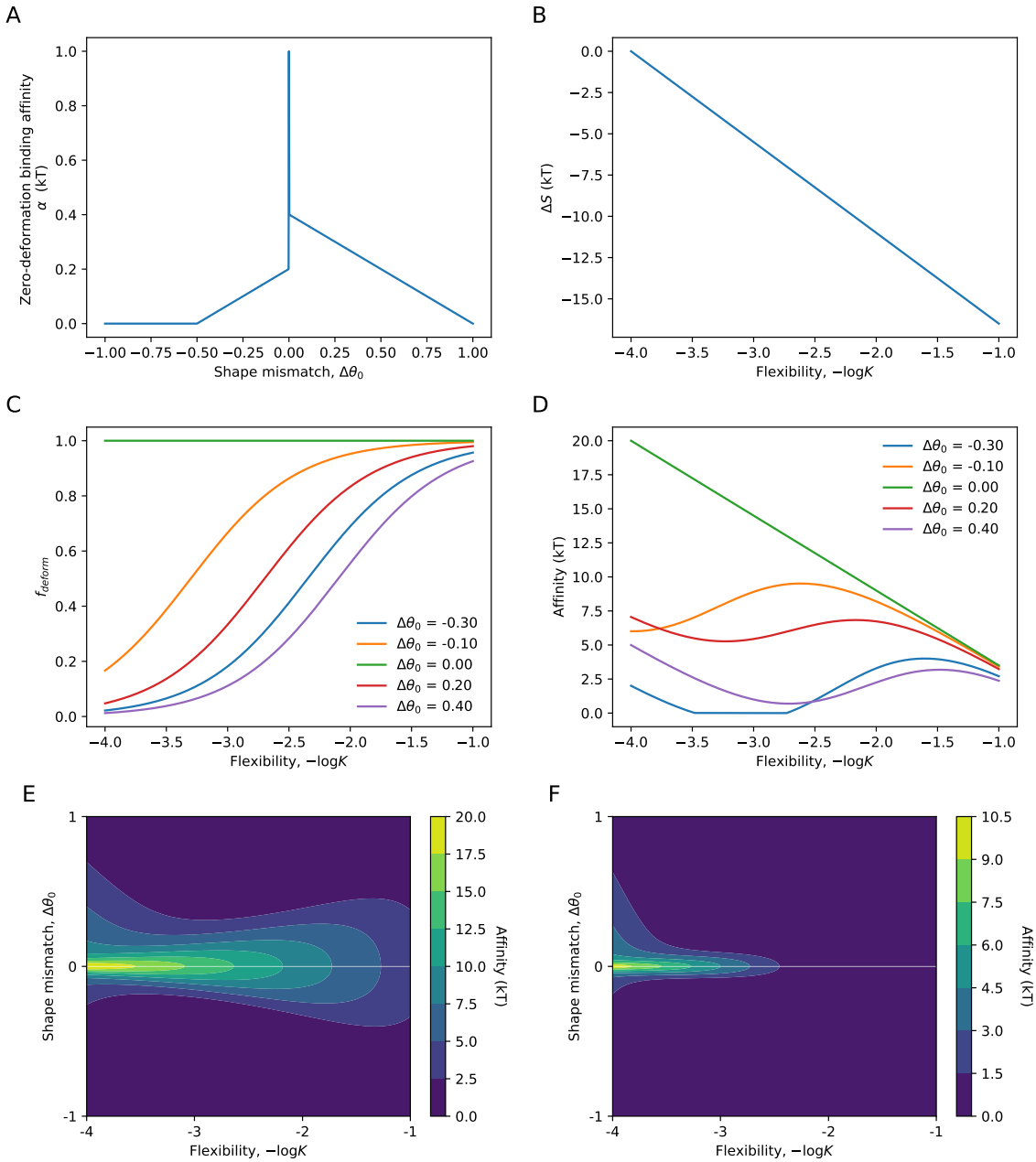

Figure 5: Here we show how the affinity phase diagram can be explained by simple relations between shape, flexibility, and chemical binding energy. We combine these terms using the equation, $\Delta G\left(\varepsilon, \Delta\theta_{LS}, K\right) = -\varepsilon_{\max}\left[\alpha + (1-\alpha) f_{\text{deform}}\right] - \Delta S$. A: α is the fraction of the total possible binding energy that can be achieved without deformation. $\alpha = 1$ when there is no shape mismatch, and is greater when the binding pocket is open ($\Delta\theta_0 > 0$) than closed ($\Delta\theta_0 < 0$). B: Binding entropy depends on flexibility of the protein, $\Delta S = -5.5\log(K_s/K)$, where we use $K_s = 10^4$ kT / nm^2. C: f_{deform} is a function that accounts for how much of the maximum chemical binding energy can be attained, depending on the cost of deformation energy. Deformation energy may be negligible if: the protein is highly flexible; there is little shape mismatch; the chemical binging energy is very high. We need a form that satisfies: $f_{\text{deform}} \to 1$ as $\Delta\theta_0 \to 0$; $f_{\text{deform}} \to 0$ as $\Delta\theta_0 \to \infty$; $f_{\text{deform}} \to 1$ as $K \to 0$; $f_{\text{deform}} \to 0$ as $K \to \infty$ if $\Delta\theta_0 \neq 0$; and $f_{\text{deform}} \to 1$ as $\varepsilon \to \infty$. We thus plot $f_{\text{deform}} = \left[\varepsilon/(K\Delta\theta_0)\right]/\left[\varepsilon/(K\Delta\theta_0)+1\right]$. D: We plot $|\Delta G|$ against K for different values of $\Delta\theta_0$ ($\varepsilon = 20$ kT), and plot $|\Delta G = 0$ if there is no binding. E-F: We plot $|\Delta G|$ against K and $\Delta\theta_0$ for $\varepsilon_{\max} = 20$ kT (E) and $\varepsilon_{\max} = 10$ kT.

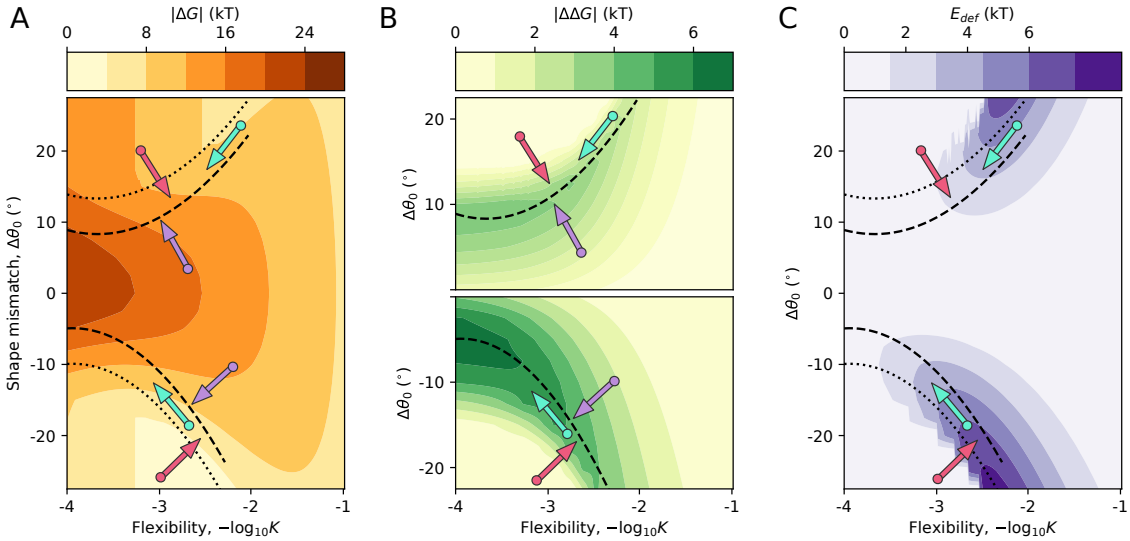

Figure 6: Binding affinity (A), specificity (B) and deformation energy (C) as a function of protein flexibility, and shape mismatch between cognate ligand and binding pocket ($\varepsilon = 8\,\mathrm{kT}$, $|\Delta\theta_{LS}| = 5^\circ$, $-\Delta S = 5.5\log_{10}(K_s/K)$, $K_s = 10^4\,\mathrm{kT/nm^2}$). In the top plot (B) the cognate ligand is smaller than the non-cognate, and the binding pocket is open to the ligand; in the bottom plot (B) the cognate ligand is larger than the non-cognate, and the binding pocket is narrow compared to the ligand. The optimal mismatch for a given flexibility is shown by the dashed line; the optimum for the non-cognate ligand is shown by the dotted line. Red arrows indicate regions where affinity is positively correlated with specificity and deformation energy increases with specificity. Cyan arrows indicate regions where affinity is positively correlated with specificity and deformation energy decreases with specificity. Purple arrows indicate regions where affinity is negatively correlated with specificity.

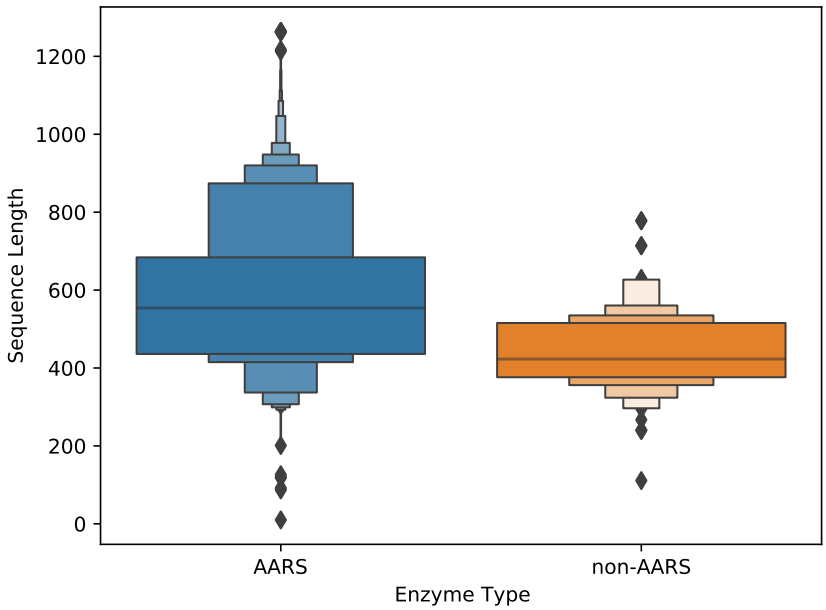

Figure 7: Sequence length distributions of enzymes whose substrates are amino acids. We show separate distributions depending on whether the enzyme is an aminoacyl-tRNA synthetase (AARS) or not.

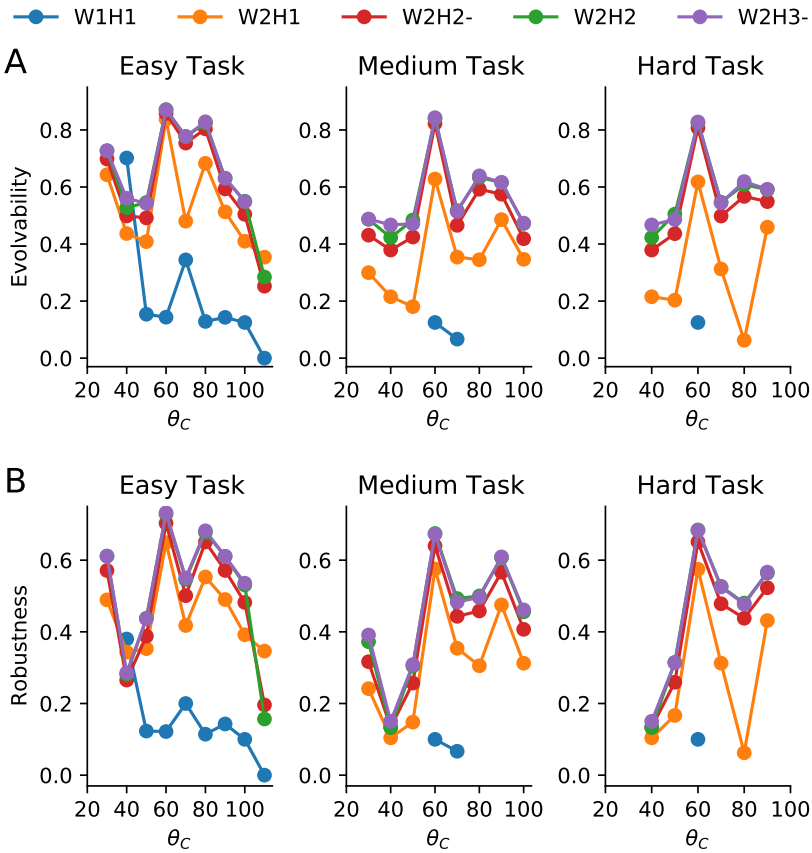

Figure 8: Evolvability (A) and robustness (B) as a function of the angle of the cognate ligand, θ_C, for easy, medium, and hard recognition tasks. Colours indicate protein models of different size. Purple and green lines often overlap. Symbols are only shown for ligands for which solutions were found. These results were obtained for structural perturbations resulting from $\delta r = 0.1$, according to sequence: **ww**, $-\delta r$; **ws**, or **sw**, δr; **ss**, δr.

References

[1] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 01 2000. ISSN 0305-1048. doi: 10.1093/nar/28.1.235. URL `https://doi.org/10.1093/nar/28.1.235`.

[2] Jean-Pierre Eckmann, Jacques Rougemont, and Tsvi Tlusty. Colloquium: Proteins: The physics of amorphous evolving matter. *Rev. Mod. Phys.*, 91: 031001, Jul 2019. doi: 10.1103/RevModPhys.91.031001. URL `https://link.aps.org/doi/10.1103/RevModPhys.91.031001`.