Accounting for Conformational Entropy in Predicting Binding Free Energies of Protein-protein Interactions

Hetunandan Kamisetty[*] Arvind Ramanathan[†]
Chris Bailey-Kellogg[‡] Christopher James Langmead[*†‡]

Short Title: Entropy and Protein-protein Interactions
Keywords: probabilistic graphical models, variational inference, thermodynamics, mutation, prediction, protein complex

Abstract

Protein-protein interactions are governed by the change in free energy upon binding, ΔG=$\Delta H - T\Delta S$. These interactions are often marginally stable, so one must examine the balance between the change in enthalpy, ΔH, and the change in entropy, ΔS, when investigating known complexes, characterizing the effects of mutations, or designing optimized variants. In order to perform a large-scale study into the contribution of conformational entropy to binding free energy, we developed a technique called GOBLIN (*Graphical mOdel for BiomoLecular INteractions*) that performs physics-based free energy calculations for protein-protein complexes under both side-chain and backbone flexibility. GOBLIN uses a probabilistic graphical model that exploits conditional independencies in the Boltzmann distribution and employs variational inference techniques that approximate the free energy of binding in only a few minutes. We examined the role of conformational entropy on a benchmark set of more than 700 mutants in eight large, well-studied complexes. Our findings suggest that conformational entropy is important in protein-protein interactions—the root mean square error (RMSE) between calculated and experimentally measured $\Delta\Delta G$s decreases by 12% when explicit entropic contributions were incorporated. GOBLIN models all atoms of the protein complex and detects changes to the binding entropy along the interface as well as positions distal to the binding interface. Our results also suggest that a variational approach to entropy calculations may be quantitatively more accurate than the knowledge-based approaches used by the well-known programs FOLDX and ROSETTA—GOBLIN's RMSEs are 10% and 36% lower than these programs, respectively.

[*]Computer Science Department, Carnegie Mellon University, Pittsburgh, PA
[†]Lane Center for Computational Biology, Carnegie Mellon University, Pittsburgh, PA
[‡]Department of Computer Science, Dartmouth College, Hanover, NH
[‡]Corresponding author. 5000 Forbes Ave., Pittsburgh, PA 15213. Phone: (412) 268 7571. Fax: (412) 268 5576. E-mail: cjl@cs.cmu.edu

1 Introduction

Cellular processes, including regulation, signaling, transport, and catalysis, are mediated by the physical interactions between molecules. These interactions, in turn, are largely governed by changes in free energy upon binding, $\Delta G = \Delta H - T\Delta S$. Here, T is the absolute temperature, while ΔH and ΔS are the changes in enthalpy and entropy due to binding, respectively. The magnitude of the change in free energy is generally small, meaning that relatively small changes in either ΔH or ΔS (e.g., due to mutation) may change the sign of ΔG, thus turning a thermodynamically favorable interaction into an unfavorable one, or vice-versa. Consequently, it is important to examine both enthalpy and entropy when studying known protein-protein interactions, their sensitivity to various mutations (i.e., $\Delta\Delta G$), or in the design of variants with improved or novel properties.

There are numerous challenges encountered when attempting to study free energy changes using all-atom models of protein-protein interactions. Among these is the ability to accurately account for changes in entropy. To see why, consider the definitions of enthalpy and entropy. Mathematically, enthalpy is the expected internal energy: $H = \sum_{c \in \mathcal{C}} P(c)E(c)$. Entropy, on the other hand, is the expected log probability: $S = -k_B \sum_{c \in \mathcal{C}} P(c) \log P(c)$. Here, $\mathcal{C}$ is discrete set of conformations, $P(c)$ is the probability mass associated with conformation c, at equilibrium, and $E(c)$ is the internal energy of conformation c. In this paper we assume that the elements of $\mathcal{C}$ correspond to the conformations that can be realized after discretizing each internal degree of freedom in the system (e.g., torsion angles). Thus, the cardinality of $\mathcal{C}$ is exponential in the number of internal degrees of freedom.

Enthalpy and entropy are both functionals over $\mathcal{C}$ and are therefore intractable to compute, exactly, because they involve evaluating a function for each element of $\mathcal{C}$. Since H and S cannot be computed exactly, in general, we must consider approximations. The most common approx-

imation for H is to use $E(c^*)$, where c^* is a global minimum energy conformation (GMEC) in $\mathcal{C}$ (assuming, of course, c^* is known or can itself be estimated). The use of $E(c^*)$ instead of H can be justified by appealing to Boltzmann's law, $P(c) = \exp\left(\frac{-E(c)}{k_B T}\right)/Z$, which characterizes the probability distribution over conformations at equilibrium (the *Boltzmann distribution*) in terms of internal energies. The law implies that any conformation with higher internal energy than c^* will be exponentially down-weighted in H, and so it is reasonable to assume that the expected internal energy is approximately equal to $E(c^*)$.

Unfortunately, the GMEC alone is not sufficient for approximating the entropic term because we must know its probability, not just its internal energy. The conversion of energies into probabilities is done through the partition function, Z, whose calculation involves a summation over all conformations because $Z = \sum_{c \in \mathcal{C}} \exp\left(\frac{-E(c)}{k_B T}\right)$. While it may be tempting to assume that $P(c^*) \approx 1.0$ (or some other large probability), this assumption is only valid at temperatures at (or near) zero degrees Kelvin. Moreover, as $P(c^*)$ approaches 1, S approaches 0, which is equivalent to approximating G with $E(c^*)$. We will show that such approximations yield sub-optimal results, when predicting binding free energies.

In practice, the entropic contribution to the free energy is usually approximated in one of two ways. The first approach involves extensive sampling of conformations (e.g., via Molecular Dynamics simulations [1, 2, 3, 4]), from which approximate entropies and/or enthalpies are derived. The second approach applies a statistical correction factor to the internal energy of the GMEC to account for such things as a change in overall volume (e.g., [5, 6, 7, 8]). Sampling-based strategies are generally too expensive for a number of important applications, such as protein design, where many unique amino acid sequences must be considered. Statistical methods, on the other hand, while computationally expedient, typically ignore subtle yet important details of the interactions between atoms, and are thus subject to both over- and underestimating entropic effects. Indeed, it has been argued [9, 10] that statistical approaches have difficulties

estimating the change in entropy upon binding.

In order to perform a large-scale investigation into the effects of entropy on protein-protein interactions, we developed a novel approach, called GOBLIN (*Graphical mOdel for BiomoLecular INteractions*). GOBLIN treats free energy calculations as a variational inference problem on a *graphical model* of the Boltzmann distribution over conformations. A graphical model is a factored encoding of a multivariate probability distribution (in our case, the Boltzmann distribution). Our approach has both representational and computational advantages in that it encodes an exponentially large number of conformations in a linear amount of space, and doesn't require sampling in order to estimate the free energy of binding. In particular, GOBLIN performs free energy calculations in a few minutes. GOBLIN is a physics-based method because internal energies are computed using a standard molecular-mechanics force field, and the probability of any given conformation satisfies Boltzmann's law.

We use GOBLIN to study the role of conformational entropy in protein-protein interactions on a benchmark set of more than 700 mutants. We find that conformational entropy plays an important role in such interactions. Specifically, we show that the root mean square error (RMSE) between GOBLIN-calculated and experimentally measured $\Delta\Delta G$s is approximately 12% (resp. 9%) smaller than the RMSE obtained when using GOBLIN-calculated $\Delta\Delta H$ (resp. $\Delta\Delta E$) as a surrogate for $\Delta\Delta G$. GOBLIN provides mechanistic insights into protein-protein interactions by quantifying the change in enthalpy and entropy upon binding for each residue in the complex. That is, GOBLIN makes it possible to examine how a loss in entropy in the binding interface can be compensated for by the change in enthalpy and, significantly, by free energy changes elsewhere in the protein. Finally, we show that GOBLIN's approach to performing free energy calculations outperforms the well-known programs FOLDX [11] and ROSETTA [12]. In particular, GOBLIN achieves an RMSE of 1.6 kcal/mol, which is 10% lower than FOLDX and 36% lower than ROSETTA. This is significant because FOLDX and ROSETTA employ

knowledge-based entropy calculation, suggesting that a graphical model based approach to free energy calculations may capture effects that are missed by statistical approaches.

2 Materials and Methods

Data Preparation

The atomic coordinates for each complex were obtained from the PDB. Hydrogen atoms were then added using the REDUCE software program [13]. In order to compute $\Delta\Delta G$, we also need the structures of the individual partners. As is common in high-throughput approaches (e.g., [12, 11]), we assumed that the native backbone in the complex is also a good approximation for the apo and holo backbones of the engineered proteins. Thus, at the end of this process, we have generated plausible structures for the apo and holo forms of the engineered structures.

Backbone ensembles for the complexes were generated using the `-backrub` [14] option of Rosetta. The method performs independent Monte Carlo simulations with "generalized-backrub" moves and selects the lowest energy structure found in each simulation. When using the backrub option, we allowed all residues whose C_α atoms were within 6 Å of the mutated position (the distance suggested by [14]) and ran 10^4 Monte Carlo steps within each simulation.

Markov Random Field Models of Protein Complexes

GOBLIN uses a Markov Random Field (MRF) to encode a Boltzmann distribution over a set $\mathcal{C}$ of possible complex structures. Figure 1 illustrates an ensemble of structures of a protein and a protein complex and fragments of MRFs encoding the corresponding ensemble.

Markov Random Fields. A Markov Random Field $\mathcal{M}$ is a pair $(\mathcal{G}, \Phi)$, where $\mathcal{G} = (\mathbf{X}, \mathcal{E})$ is an undirected graph over a set of random variables, $\mathbf{X} = \{X_1, \ldots, X_n\}$, with edges $\mathcal{E}$, and

$\Phi = \{\phi_1, \phi_2, ..., \phi_m\}$ is a set of functions (popularly called factors) over the nodes and edges of the graph, such that $m = n + |\mathcal{E}|$. The semantics of each edge $e \in \mathcal{E} \subseteq \{\{u, v\} \mid u, v \in \mathbf{X}\}$ is that u and v are statistically dependent random variables.

Given $\mathcal{G}$ and Φ, the Hammersley-Clifford theorem [15] states that, provided each ϕ_i is a positive function, the probability of a specific assignment to the random variables, $\mathbf{X} = \mathbf{x}$, can be written as:

$$P(\mathbf{X} = \mathbf{x}) = \frac{1}{Z} \prod_{\phi_i \in \Phi} \phi_i(\mathbf{x}) \tag{1}$$

where Z is the partition function:

$$Z = \sum_{\mathbf{X}} \prod_{\phi_i \in \Phi} \phi_i(\mathbf{x}). \tag{2}$$

Thus the probability of a given state is simply the product of the functions, suitably normalized.

We have previously described the construction of MRFs for monomers [16, 17]. We extend that approach here to represent protein complexes and to account for backbone flexibility.

Side-chain variables. The first set of variables in our MRF, $\mathbf{X}_s$, represents side-chain flexibility. We employ the rotamer library of Dunbrack and co-workers [18] to define a discrete set of possible side-chain conformations, which we then model with multinomial random variables. Let $\mathbf{X}_s = \{X_{s1}, \ldots, X_{sn}\}$ denote such a set of n multinomials, one for each residue, indicating the chosen rotameric state. Then $\mathbf{X}_s$ can be thought of as a random variable over the side-chain portion of $\mathcal{C}$, and the assignment $\mathbf{X} = \mathbf{x}_c \in \mathcal{C}$ corresponds to a specific conformational state of all side-chains for a fixed backbone.

Backbone variable. An additional variable X_b represents backbone flexibility. As with side-chains, we consider a discrete sampling of backbone conformations, and model them with a

multinomial random variable. While any sampling method could be used, the results present those generated by BACKRUB motions [19]. For one of our outliers, we also generated backbones by molecular dynamics, as described in a later section.

Boltzmann distributions. Boltzmann's law describes the probability distribution over $\mathcal{C}$ at equilibrium. Let us first consider a fixed backbone b. The probability that side-chains $\mathbf{X_s}$ occupy a state $\mathbf{x_s}$ with internal energy $E_b(\mathbf{s})$ is:

$$P(\mathbf{X_s} = \mathbf{x_s}|X_b = x_b) = \frac{1}{Z_b} \exp\left(\frac{-E_b(\mathbf{s})}{k_B T}\right) \tag{3}$$

where

$$Z_b = \sum_{\mathbf{x_s}} \exp\left(-\frac{E_b(\mathbf{s})}{k_B T}\right) \tag{4}$$

with k_B for Boltzmann's constant, T for the absolute temperature in Kelvin. Here $E_b(\mathbf{s})$ is the internal energy of $\mathbf{x}_s$ in backbone b.

To account for the backbone variable, we express the total energy $E_\mathbf{c} = E_b + E_b(\mathbf{s})$, and then define the joint distribution and the partition function via the chain rule:

$$P(\mathbf{X} = \mathbf{x_c}) = \frac{1}{Z_{conf}} \exp\left(-\frac{E_\mathbf{c}}{k_B T}\right) = P(X_b = x_b) P(\mathbf{X_s} = \mathbf{x_s}|X_b = x_b) \tag{5}$$

with

$$Z_{conf} = \sum_{\mathbf{x_c}} \exp\left(-\frac{E_\mathbf{c}}{k_B T}\right) = \sum_{x_b} \sum_{\mathbf{x_s}} \exp\left(-\frac{E_b + E_b(\mathbf{s})}{k_B T}\right) = \sum_{x_b} \exp\left(-\frac{E_b}{k_B T}\right) Z_b \tag{6}$$

where Z_b is calculated according to Eq. 4 and Z_{conf} is the partition function over the conformational degrees of the protein alone (i.e., without solvent).

Graph structure. We place an edge between nodes for side-chain variables if their C_α atoms are within a cutoff distance d. In our experiments, we set d to 10 Å, the value used by ROSETTA. Consequently, the resulting graph for n residues has $O(n)$ edges (due to packing arguments), and if there are at most k rotamers per residues, this portion of the MRF encodes $O(k^n)$ unique conformations in $O(kn)$ space.

The node for the backbone variable is connected to all the side-chain nodes.

Graph potentials. By choosing the potentials Φ in terms of *Boltzmann factors*, we can directly model the Boltzmann distribution. That is,

$$\phi_i(x_{\phi_i}) = \exp\left(-\frac{E(x_{\phi_i})}{k_B T}\right) \tag{7}$$

where x_{ϕ_i} is the set of atoms that serve as arguments to ϕ_i, and $E(x_{\phi_i})$ is the potential energy of those atoms as defined by a molecular mechanics force field. In principle, any force field can be used; we use an implementation of the one ROSETTA uses in computing $\Delta\Delta G$ [12], which is composed of the following terms:

- E_{ljatr} and E_{ljrep}, the attractive and repulsive parts of a 6–12 Lennard-Jones potential, used to model van der Waals interactions as computed by [20];
- E_{hb}, the hydrogen bond energy, as computed by ROSETTA [21].

The parameters that define each individual term (atomic radii, etc.) were obtained from the `soft-rep` setting of ROSETTA since previous studies [22] have indicated that it is better suited for computations with discrete conformations. When combining multiple force-field terms, it is common [12] to parameterize the individual contribution of each term to the total energy using a weight for each individual term. Following this, we define parameters $w_{ljatr}, w_{ljrep}, w_{hb}$ that will be learned from training data, as described in a later section.

MRFs of apo and holo forms. The binding free energy of a protein complex is the difference between the free energies of the apo (unbound) and the holo (bound) forms. To compute this, it is therefore necessary to model both the apo and the holo forms. Thus, for a complex involving molecule A and B, we construct three separate MRFs: (i) one for the holo form, (ii) one for A in isolation, using the backbone of A from the holo form, and (iii) one for B in isolation, using the backbone of B from the holo form.

MRFs of mutants. Separate MRFs are constructed for each mutation considered. This is done by performing an *in silico* mutation to the PDB structure, and constructing a new MRF accordingly.

Free Energy Calculations by Probabilistic Inference

The GOBLIN MRF provides a compact encoding of the Boltzmann distribution over the conformation space $\mathcal{C}$. In general, the free energy of a physical system is related to the Boltzmann distribution by way of the partition function: $G = -k_B T \log Z$, where Z is the sum of the Boltzmann factor in all states of the system. For GOBLIN, Z (Eq. 6) is the sum of the Boltzmann factor over each conformation in our discretized representation of backbone and side-chain flexibility.

The task of probabilistic inference is to compute the probability of an event of interest. Since the unnormalized probability (i.e., the Boltzmann factor) is easy to compute in an MRF, the main task of probabilistic inference in such models is computing the normalizing constant—the partition function Z. This is computationally intractable in the general case [23]. However, the machine learning community has developed a number of efficient algorithms for performing probabilistic inference in MRFs, and shown their equivalence to specific free-energy approximations introduced by statistical physicists [24, 25, 26, 27]. We use such an approach here.

Belief propagation. We employ a variational inference technique called *loopy belief propagation* in order to compute an approximation to the partition function, and hence the free energy. Loopy belief propagation [28] is a variant of Pearl's *belief propagation* (BP) algorithm [29], which has been shown to be equivalent to the Bethe approximation [24] of the free energy. Since the Bethe approximation of the free energy can be written as a function of single variable and pairwise marginal probability distributions, BP approximates these marginals (called "beliefs", leading to its name) thereby approximating the global partition function.

The working of BP and its connections to the Bethe Free energy have been explored in great detail elsewhere (cf. [30]). The interested reader is directed to the appendix for a detailed example of how the algorithm works. Briefly, each node in the graph keeps track of its own marginal probability distribution (i.e., belief), starting from the prior (in our case, this is the statistical prior provided by [18]). Message passing is performed between nodes, with each node updating its own beliefs based on the beliefs of its neighbors in the graph and the value of the potential function relating them. When the algorithm converges, the final beliefs can be used to obtain approximations for various quantities of interest, including the enthalpy (H), entropy (S), free energy $G = H - TS$, and partition function $Z = \exp\left(\frac{G}{-k_B T}\right)$. In particular, if $b_i(x_{\phi_i})$ is the BP-computed belief (i.e., approximate marginal) that atoms associated with the ith factor are in conformation $x_{\phi_i} \in X_{\phi_i}$, then the Bethe approximation to the free energy can be computed as $G = H - TS$ where

$$H = -\sum_i \sum_{x_{\phi_i} \in X_{\phi_i}} b_i(x_{\phi_i}) \log \phi(x_{\phi_i}) \tag{8}$$

$$S = -\sum_i \sum_{x_{\phi_i} \in X_{\phi_i}} c_i b_i(x_{\phi_i}) \log b_i(x_{\phi_i}) \tag{9}$$

where $c_i = 1$ if ϕ_i corresponds to an edge and $1 - |Nbrs(i)|$ if ϕ_i corresponds to a vertex and $|Nbrs(i)|$ is the number of its neighbors in the graph $\mathcal{G}$[16]. The bethe approximation to the

partition function can then be computed as $Z = \exp(-G)$.

BP is exact and efficient (i.e., runs in polynomial time) in the case of trees. While there are no guarantees for loopy BP in general graphs (like those in GOBLIN), we have always found it to converge in practice. When estimating binding free energies, BP is invoked to compute residue-specific marginals in the bound and unbound forms (e.g., Fig. 3). A change in marginals, naturally, corresponds to a chance in enthalpy and entropy, and thus free energy.

Accounting for discretization. There is a subtle, yet important, issue associated with discretizing the conformational space $\mathcal{C}$. The use of discrete rotameric states is well-founded (cf. [18, 31, 32]), and poses no particular challenge when performing tasks such as side-chain placement, i.e., finding the single most energetically favorable side-chain conformation [33, 34, 35, 18]. The issue arises when using these libraries to compute free energies.

To understand the problem, let us consider an imaginary protein with exactly one residue whose side-chain atoms occupy some finite space but are otherwise unconstrained; i.e., the energy is E, regardless of conformation. This protein has a physically measurable (and finite) amount of conformational entropy. Now suppose we discretize the conformation space into n rotamers, each representing an equal fraction of the space; i.e., with probability $1/n$. We then compute the free energy in this discrete model as:

$$\begin{aligned} G_{discrete} &= H - TS_{discrete} \\ &= \langle E \rangle - k_B T \sum_{i=1}^{n} -\frac{1}{n} \log\left(\frac{1}{n}\right) \\ &= E - k_B T \log(n) \end{aligned}$$

where we used the relation between the thermodynamic entropy S and the Shannon Information entropy [36, 37].

Thus as the granularity of the discretization increases, the discrete entropy increases; as $n \longrightarrow \infty$, we see that $S_{discrete} \longrightarrow \infty$, completely unconnected to the finite physically measurable value.

We note that this problem is an artifact of the discretization of the probability distribution and does not occur in other, continuous, treatments [38]. This artifact of discretization arises in many scenarios, most notably for our purposes in information-theoretic treatments of statistical physics [37, 39]. Fortunately, a solution to this problem is available, which to the best of our knowledge is due to E.T. Jaynes [37]. By using a measure (i.e., a possibly unnormalized probability distribution) m over the space and replacing the discrete entropy by the relative entropy we obtain a quantity that behaves correctly in the limit.

To correctly handle the discretization of side-chains, we use the statistical prior provided by our rotamer library [18] as the measure. In the case of a fixed backbone, we then compute the relative entropy R_b as:

$$R_b = -k_B \sum_{\mathbf{X_s}} P(\mathbf{X_s}|X_b) \log \frac{P(\mathbf{X_s}|X_b)}{m(\mathbf{s})} = -\sum_{\mathbf{X_s}} (P(\mathbf{X_s}|X_b) \log P(\mathbf{X_s}|X_b) - P(\mathbf{X_s}|X_b) \log m(\mathbf{X_s}))$$

And by replacing the entropy with the relative entropy, we compute free energy as:

$$G_b = \sum_{\mathbf{X_s}} P(\mathbf{X_s}|X_b) E_b(\mathbf{s}) + k_B \sum_{\mathbf{X_s}} (P(\mathbf{X_s}|X_b) \log P(\mathbf{X_s}|X_b) - P(\mathbf{X_s}|X_b) \log P(m(\mathbf{X_s}))) \tag{10}$$

$$= \sum_{\mathbf{X_s}} P(\mathbf{X_s}|X_b)(E_b(\mathbf{s}) - k_B \log m(\mathbf{X_s})) - k_B \sum_{\mathbf{X_s}} P(\mathbf{X_s}|X_b) \log P(\mathbf{X_s}|X_b) \tag{11}$$

In other words, the move from the discrete entropy to the discrete relative entropy can be made by adding to the energy function (Eq. 12) a term $w_{rot} E_{rot}$, where the energy $E_{rot} = -k_B \log m(\mathbf{X_s})$. The total energy of a conformation according to GOBLIN's force-field is thus:

$$E_{goblin} = w_{ljatr}E_{ljatr} + w_{ljrep}E_{ljrep} + w_{hb}E_{hb} + w_{rot}E_{rot} \tag{12}$$

A similar problem arises when summing over multiple backbone samples according to Eq. 6: by increasing the number of backbone samples, the value of Z monotonically increases. For our experiments we assumed that the conformational space is uniformly sampled, i.e., each backbone represents an equal volume of the conformational space. This is equivalent to assuming a uniform prior measure $m(X_b)$, and leads to an equation using this measure analogously to $m(\mathbf{X_s})$ in Eq. 11.

Computing conformational binding free energy. We can now describe how binding free energies are computed. For a complex $C = AB$, if we assume that the free energy of the unfolded state is expressible as a sum of single-body terms [12, 11], then we can compute the *conformational* contribution to the binding free energy as:

$$\Delta G_{conf}(C) = G_{conf}(C) - (G_{conf}(A) + G_{conf}(B)) \tag{13}$$

$$= k_B T(\log Z_{conf}(A) + \log Z_{conf}(B) - \log Z_{conf}(C)) \tag{14}$$

We use belief propagation to compute conformational partition functions (and therefore, free energies) for each of the apo MRFs (A and B) and for the holo MRF (C). ΔH_{conf} can be computed in an analogous fashion by computing H_{conf} as the *expected energy* of the system using the probabilities computed by Belief Propagation. Furthermore, we note that ΔE_{conf} can also be computed using a closely related algorithm to Belief Propagation, known as max-product Belief Propagation that approximates the global minimum energy conformation (GMEC) [33].

Notice that $m(X_b)$ terms for a complex and its partners cancel out while computing ΔG. They therefore do not appear explicitly in our subsequent analysis.

Computing changes in binding free energy upon mutation. The change in binding free energy upon mutation, $\Delta\Delta G$, is the difference between the binding free energies of the wild-type and mutant forms. Thus for mutant protein C_{mut} and wild-type protein C_{wt}, we have

$$\Delta\Delta G(C_{mut}) = \Delta\Delta G_{conf}(C_{mut}) + \Delta\Delta G_{solvent,protein}(C_{mut}) + \Delta\Delta G_{coop}(C_{mut}) \tag{15}$$

where

- $\Delta\Delta G_{conf}(C_{mut}) = \Delta G_{conf}(C_{mut}) - \Delta G_{conf}(C_{wt})$ is the change in conformational binding free energy on mutation (from Eq. 14);
- $\Delta\Delta G_{solvent,protein} = w_{SASA}\Delta\Delta SASA$ accounts for interactions between the protein and solvent and is proportional to the change upon mutation of the loss of solvent-accessible surface area due to binding [11];
- $\Delta\Delta G_{coop} = w_{coop} ISA_{wt}$ captures the contribution of loss of cooperativity (i.e., multi-body interactions) due to mutations at the interface and is proportional to the area of the interface [40].

Learning force field parameters against free energies

A molecular mechanics force-field consists of: (i) a defined set of atom types; (ii) a function defining the internal energy of the system; and (iii) a set of parameters. It is common [11, 12, 40] to take the atom types and energy functions as fixed, but to adjust the parameters for a particular type of study.

A commonly used strategy for optimizing force field parameters is to minimize the sum of the squared errors between predicted and experimentally measured internal energies using fixed structures. In contrast, we consider the problem of minimizing the sum of of the squared

errors in free energies, as computed using MRFs. The two problems are fundamentally different. In particular, whereas minimizing differences in internal energies gives rise to a simple linear regression problem, minimizing differences in free energies is a complicated non-linear regression problem involving the minimization of a functional (i.e., the partition function). We developed a novel algorithm to solve this problem in an efficient albeit approximate manner.

Given a training set of experimentally measured $\Delta\Delta G$ values for N mutants of that complex, along with the wild-type ΔG, consider the problem of learning force-field parameters to minimize the mean square error (MSE) between predicted and observed $\Delta\Delta G$. We do so by adjusting the vector of weights $\mathbf{w} = [w_{ljatr}, w_{ljrep}, w_{hbond}, w_{rot}, w_{sasa}, w_{coop}]$ with which we linearly combine the corresponding force-field terms.

In referring to the different observations and predictions, let us use superscripts e for experimental and p for predicted, and subscripts $i \in \{1, \ldots, N\}$ for the various datapoints

This allows us to express the MSE as

$$mse = \frac{1}{N} \sum_{i=1}^{N} (\Delta\Delta G_i^p - \Delta\Delta G_i^e)^2 \tag{16}$$

To minimize MSE subject to $\mathbf{w} \succeq 0$ by gradient descent, we must compute the gradient ∇mse:

$$\nabla mse = \left[\frac{\partial mse}{\partial w_{ljatr}}, \frac{\partial mse}{\partial w_{ljrep}}, \ldots, \frac{\partial mse}{\partial w_{coop}} \right] \tag{17}$$

$$\forall w, \frac{\partial mse}{\partial w} = \frac{1}{N} \sum_{i=1}^{N} 2 \left(\Delta\Delta G_i^p - \Delta\Delta G_i^e \right) \left(\frac{\partial \Delta\Delta G_i^p}{\partial w} \right) \tag{18}$$

Using $\frac{\partial G_{conf}}{\partial w} = \langle E \rangle_{\mathbf{c}}$ where $\langle E \rangle_{\mathbf{c}}$ is the expected value of the corresponding force-field terms over all $\mathbf{x_c}$ using the current value of w, and the fact that the derivative of differences is just the difference of derivatives, we have:

$$\frac{\partial \Delta\Delta G^p}{\partial w} = \begin{cases} \frac{\partial \Delta\Delta G^p_{conf}}{\partial w} = \Delta\Delta\langle E_i \rangle_{\mathbf{c}}, & \text{for } w \in \{w_{ljatr}, w_{ljrep}, w_{hbond}, w_{rot}\} \\ \Delta\Delta SASA, & \text{for } w_{sasa} \\ I_{wt}, & \text{for } w_{coop} \end{cases} \tag{19}$$

respectively.

In the case of a fixed backbone, the expectation $\langle E \rangle$ is over just $\mathbf{x_s}$. In both cases the expectations, and thus the gradient, can be computed along with the free energy during inference. Given this method of computing gradients, we performed gradient descent, updating the weights $\mathbf{w}$ at iteration i using the following equation $\mathbf{w}^{(i+1)} = \mathbf{w}^{(i)} - \eta \nabla mse$ where the step size η was set to 0.05. To test for the sensitivity of our results to η, we experimented on one training set with various values between 0.01 and 0.15. We found that the step size affected the rate of convergence but had negligible effect on the final values indicating that the quality of results aren't sensitive to this parameter.

Given the enthalpies and free energies of each backbone trace, we can compute the free energy of the entire distribution and its derivatives.

Software Packages

For our comparison studies, we used FOLDX version 3.0 and ROSETTA version 2.3. For FOLDX, we prepared mutants using the BUILDMODEL option and computed binding free energies using the ANALYSECOMPLEX option as specified by the FOLDX manual. When using the ROSETTA software program to compute binding free energies, we used the `-analyze-interface` [12] flag according to the ROSETTA manual available on the ROSETTA commons webpage [41].

Molecular Dynamics Simulations

Molecular Dynamics simulations were performed for the wild-type and the R21P mutant of alpha-chymotrypsin (PDB ID: 1cho). The structures were prepared using the MAESTRO software (Schrödinger Inc.) [42] and solvated using SPC water model in a rectangular box with a buffer of 10 Å from the corners of the box. The system was first minimized until the root mean square (rms) of the gradients was less than 0.01 kcal/mol. Each system was then equilibrated using a standard protocol in MAESTRO. The protocol consisted of running a total of six successive steps of energy minimization and small MD simulations to remove bad contacts, allow solvent to equilibrate, and occupy vacuous regions in the solute. The MD simulations during the equilibration process gradually raised the temperature to 300 K. During the final stages of equilibration a small MD production run lasting 600 ps was carried out to ensure that the system retained its overall stability. Subsequent production runs were carried out using NVE ensemble using periodic boundary conditions using the DESMOND software package (version 2.2.7) [43]. For each system, a total of 6 ns sampling was carried out. Conformations were saved every 50 ps, accounting for a total of 3,000 conformers.

3 Results

We studied the importance of entropy on a database of 704 single-point mutants from eight large and well studied complexes. For each of these mutants, the database contains the $\Delta\Delta G^e$, the experimental change in binding free energy upon mutation. The details of the datasets, along with the Protein Data Bank (PDB) [44] ids of the wildtype complexes, are shown in Table 1. Of these, the three largest datasets (wildtype PDB ids: 1sgr, 1cho, 1ppf) are from the Kazal family of serine protease inhibitors [45] while the rest of the interactions are part of an Alanine-scanning database previously used in [12] and [21]. We note that the amount of

thermodynamic data available for protein-protein interactions is limited, and the database we considered is among the largest of its kind.

The atomic coordinates for each complex were obtained from the PDB and converted into probabilistic graphical models (PGM), which are used to perform free energy calculations. Briefly, if $\mathbf{X} = \{X_1, ..., X_n\}$ is a vector encoding the conformation of the protein (or protein complex), each PGM encodes the distribution $P(\mathbf{X})$ using a factored representation. By construction, the probability of any particular conformation is inversely proportional to the exponential of its internal energy, as computed using a molecular mechanics force field. Interaction energies fall off quickly with distance leading to conditional independencies in the Boltzmann distribution.[1] GOBLIN takes advantage of these conditional independencies in its factorization, which leads to a compact encoding of the joint distribution (Figure 1). This factorization also leads to an efficient means for performing free energy calculations, and for optimizing force field parameters against experimentally observed $\Delta\Delta G$ (see Methods).

For each complex $C = AB$ consisting of proteins A and B, we construct three separate PGMs (Figure 2). The first two PGMs model the Boltzmann distribution over the apo (unbound) conformations of A and B, and the third models the Boltzmann distribution over the holo (bound) conformation. Loopy Belief Propagation [28] is then performed on each PGM to compute the free energies G_A, G_B, and G_C (i.e., before and after binding). The free energy of binding is computed as: $\Delta G = G_C - (G_A + G_B)$. Similarly, the change in binding free energy upon mutation, $\Delta\Delta G$, is computed by performing an *in silico* mutation, repeating the binding free energy calculation, and computing the difference: $\Delta\Delta G = \Delta G_{mut} - \Delta G$. $\Delta\Delta E$ values were computed using max-product Belief Propagation that approximates the global minimum energy conformation (GMEC).

We considered two scenarios. In the first scenario, the PGM models the Boltzmann distri-

[1]Two random variables X and Y are said to be conditionally independent, given Z, iff $P(X, Y|Z) = P(X|Z)P(Y|Z)$.

bution over side-chain conformations, conditioned on a fixed backbone. Side-chains conformations are modeled using the backbone specific rotamer library described in [18]. The second scenario models the Boltzmann distribution over both side-chain and backbone conformations. Alternative backbone conformations were obtained using the BACKRUB method [14].

Computation of interaction energies employs a molecular mechanics force-field with weights on different terms; GOBLIN learns those weights from a training set of $\Delta\Delta G$ data. We thus formed 20 random partitions of the 704 mutants into 352 training structures and 352 testing structures. The training structures were used to optimize the parameters of GOBLIN's force-field. The optimized parameters are shown in Table 2. We note that the parameters are optimized to maximize predictive performance, and not necessarily for biophysical interpretability. However, our optimized parameters are at least in part comparable to those in existing force fields. Our weight for the Lennard-Jones term, for example, is comparable to the weight used by Rosetta when modeling discrete rotameric states. Similarly, the weights on the $SASA$ and the I_{wt} terms are comparable in magnitude to [40]. The differences between our weights and other force fields is likely due to both differences in training data and the fact that our method models entropy explicitly, which leads to a different objective function. Our weight for hydrogen bonding is lower than expected. We note, however, that our method models solvent using terms proportional to the SASA. The SASA term, in turn, partially accounts for hydrogen bonding between solvent atoms and solvent-protein atoms. Thus, the weight on the hydrogen bond term may be lower than might be expected if the solvent was modeled explicitly.

The trained model was used to predict the binding free energies for the test structures, and to identify the GMEC. In what follows, all errors are reported as averages over the 20 partitions. For comparison, we also used the programs FOLDX (version 3.0) and ROSETTA (version 2.3) to compute binding free energies.

Changes in Entropy Upon Binding

We first consider the nature of the changes in entropy caused by binding, and the effects of mutations on those changes. GOBLIN can compute detailed information on the changes in entropy (and enthalpy) for each residue. For example, Figure 3 illustrates the change in marginal over side chain configurations upon binding for residue Trp 304 in the HGH-HGHBP complex. Notice that the dominant rotamer has just under 50% of the probability mass before binding and that the rest of the mass is primarily distributed among four additional rotamers. After binding, there is a substantial shift in probability mass; the dominant rotamer now has almost 80% of the probability mass, and most of the rest of the mass is distributed among two additional rotamers. Figure 4-A visualizes GOBLIN's predicted change in entropy upon binding for the wild-type Human Leukocyte Elastase : Turkey Ovomucoid and the Human Growth Hormone : Human Growth Hormone Binding Partner complexes. In these figures, the surfaces of the two partners are shown in purple and yellow respectively. Spheres mark the C_α atoms of residues that show a non-trivial change (absolute change in entropy ≥ 0.1 k_B units) in their entropy, with red spheres for the largest change and blue for the smallest. Not surprisingly, all the interface residues showed large decrease in entropy. More interestingly, the decrease in flexibility in the interface affects the entropy of neighboring residues. In the HLE-OMTKY complex, Ser 214 of HLE showed lower entropy in the holo form than the apo form, despite being > 10 Å away from the interface. In the HGH-HGHBP complex, these distal effects were stronger: Trp 86 and Glu 373 of HGH and Glu 373 of HGHBP (distances to interface: 17.2 Å, 12.9 Å and 12.9 Å resp.) all showing a non-trivial change in binding despite being far away from the interface. These changes in entropy, which are unfavorable, are compensated by a corresponding decrease in enthalpy, to make the binding favorable.

Figure 4-B shows the difference in binding entropy upon mutation (i.e., $\Delta\Delta S$) for one mutation from each of these two complexes: L18H of OMTKY and D171A of HGH. Again,

the surface colors represent the two partners. All atoms of the residues showing a non-trivial binding entropy difference with respect to the wild-type ($\Delta\Delta S \geq 0.1$) are shown in spheres. Of the distal residues from panel A, Trp 86 of HGH and Ser 214 of HLE show a non-trivial change in entropy. In the former case, the change in entropy is actually positive (i.e., there is less entropic cost to binding in the alanine mutant than the wild-type) while in the latter case the change in entropy is negative. These results demonstrate that the distal entropic effects on binding of a mutation can be different from those in the wild-type, underlining the need to determine them accurately.

Quantitative Analysis

We next consider the quantitative accuracy of the free energy predictions, and the relative importance of side-chain and backbone conformational entropies.

Effects of Side-chain Entropy

The quantitative accuracies of GOBLIN under the fixed backbone scenario are presented in Table 3. The row labeled "GOBLIN" reports the root mean squared errors (RMSE) between prediction and observation for our method. The row labeled "GOBLIN-E" is the RMSE obtained when $\Delta\Delta G_{conf}$ is replaced with $\Delta\Delta E_{conf}$ — the change in internal energy for the GMECs in the apo and holo forms. The row labeled "GOBLIN-H" is the RMSE when $\Delta\Delta G_{conf}$ is replaced with $\Delta\Delta H_{conf}$, the expected energy averaged over the Boltzmann distribution, computed by neglecting the entropic component of the free energy computed by Belief Propagation. GOBLIN's RMSE is 1.6 kcal/mol, which is 9% lower than GOBLIN-E ($p < 0.05$), and about 12% lower than GOBLIN-H ($p < 0.01$). The drop in RMSEs persists when the 5th and 10th percentile of errors are removed (final two columns), suggesting that the difference in accuracies is robust to outliers. We conclude that entropic contributions play a significant role in protein-protein

interactions, because ignoring them results in a significant increase in RMSE. We discuss the errors in more detail below.

Table 3 also compares GOBLIN's RMSE with that for the programs FOLDX and ROSETTA. GOBLIN outperforms FOLDX by nearly 10% ($p < 0.03$), and ROSETTA by 36% ($p < 6.8 \times 10^{-7}$). Significantly, GOBLIN continues to have lower RMSEs after removing each approach's least-accurate predictions, suggesting that the difference in accuracies is robust to outliers. In particular, when the 5th (resp. 10th) percentile of errors are removed, GOBLIN outperforms FOLDX by 11% (resp. 9%) ($p < 0.01$; resp. $p < 0.08$), and outperforms ROSETTA by 34% (resp. 35%) ($p < 1.5 \times 10^{-5}$; resp. $p < 2.2 \times 10^{-4}$). Notice that the RMSE of FOLDX, which uses a knowledge-based approximation of the change in entropy, is approximately the same as GOBLIN-H, which ignores entropy altogether. These results suggest that GOBLIN's variational approach to free energy calculations is superior to the knowledge-based methods used by FOLDX and ROSETTA. The difference in accuracy is likely due to the fact that the variational approach considers not only the direct effects of each mutation on the free energies, but also the indirect effects on neighboring residues. Indeed, the very nature of Belief Propagation involves diffusing information throughout the graphical model.

Figure 5 shows a scatter plot comparing GOBLIN's predictions using the weights selected by cross-validation on the entire dataset. The correlation coefficient (R^2) across the entire dataset was 0.56. Outlier elimination improved this substantially, to 0.66 without the top 5% outliers and to 0.70 without the top 10% outliers. In comparison, the correlation coefficient for FoldX was 0.52 and increased to 0.62 (resp. 0.67) after removing the top 5% (resp. 10%) outliers. The corresponding values for Rosetta were 0.0, 0.06, and 0.15, respectively.

Errors for different residue types: Figure 6 (top and middle) show boxplots of the error in prediction according to mutant and wild-type amino acids respectively. In each box, the central red line is the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend

to the most extreme data points not considered outliers, and outliers are plotted individually with red '+' marks. Note that not all 20 amino acids were mutated in our dataset: some (Cys, Gly) weren't mutated at all and a mutation from Met occurred only once, while mutations from Lys and Glu each accounted for about 12% of all mutations. The median signed error for most mutant amino acid types is close to zero, with a few notable exceptions, namely Pro, Lys, Asp, and Glu. Of these, mutations to and from Pro produced the highest errors. Indeed the three largest RMSEs in the entire data set were all proline mutations. This is to be expected since proline has an atypical backbone and a mutation to it can cause significant structural changes. The median signed error for mutations to lysine was nearly 2.0 kcal/mol and mutations to aspartic acid and glutamic acid had a larger spread in errors than other amino acids. These errors were likely caused due to their charged nature, and the fact that our force field does not presently account for electrostatics. Across the entire data set, GOBLIN's error on charged mutants was larger than on neutral residues, indicating the greater difficulty in modeling their interactions.

Errors due to change in charge or volume: Tables 4 and 5 stratify the errors in terms of the change in residue charge and volume after mutation, respectively. Relative to a global RMSE of 1.6 kcal/mol, certain kinds of mutations yield larger than average RMSEs, including: mutations from one negatively charged residue to another; mutations from a neutral residue to a charged residue; and mutations from a positively charged residue to a neutral residue. These errors may reflect the fact that GOBLIN's force field, like ROSETTA's, does not explicitly account for electrostatic interactions other than hydrogen bonds. Mutations from small to large residues, and from large to either small or medium size residues are also associated with an increase in RMSE (Table 5). This is to be expected given that our backbones were held fixed for these experiments, and so no change is made to account for unfavorable packings. As explained subsequently, the error in such cases improves upon incorporating backbone flexibility.

Errors in different complexes: Figure 6-bottom shows the breakdown of GOBLIN's performance across the eight datasets listed in Table 1 arranged in decreasing order of number of mutants. In four of the eight complexes, GOBLIN's error is around 1.5 kcal/mol or smaller. The largest error is in 3hfm where the RMSE is nearly 3.0 kcal/mol, which is marginally better than FOLDX's RMSE, and previously published results using ROSETTA [12]. One possible reason for such behavior might be due to conformational changes with distal effects, as suggested by [46]. Additionally, the three programs (GOBLIN, FOLDX, and ROSETTA) assume that the apo backbones of the proteins are similar to their holo forms. When this assumption is violated, no program is expected to perform well, suggesting that it may be necessary to minimize the structure of the apo forms, or use apo forms deposited in the PDB.

Outliers: Table 6 lists GOBLIN's largest outliers (absolute error $\geq$ 5 kcal/mol). Four of the nine involve a mutation to proline from arginine in the serine protease inhibitor. Prolines have an atypical backbone, and often result in a substantial change in backbone configuration. Our means for sampling backbones does not, in general, handle such changes well. Of the remaining five outliers, four are mutants to the HyHEL-10 Fab-lysozyme complex (3hfm), a system that is known to undergo large scale re-arrangements upon binding. Here, our assumption that the apo and holo backbones are approximately the same is inappropriate. Moreover, one of these mutations, K96A, has been postulated to result in a loss of a salt bridge [46]. The force-field we currently use does not capture such interactions. The final outlier involves the loss of a strong electrostatic interaction, which, as previously mentioned, is not presently implemented in GOBLIN's force field.

Effects of Backbone and Side-Chain Entropy

We next consider PGMs modeling Boltzmann distributions over both side-chain and backbone conformations. Our expectation was that we would see a further reduction in RMSE, especially

for disruptive mutations (e.g., those involving a large increase in the size of the side chain). We generated a set of nine alternative backbone conformations using a Backrub-like method developed by Kortemme and others [14] that is implemented in Rosetta. The method runs independent Monte-carlo simulations with "generalized backrub" moves and selects the lowest energy structure from each simulation. Along with the native backbone, this gave us an ensemble of ten backbones which is the size of the ensemble used by [14, 47] in their backrub studies. We then re-optimized the parameters and re-computed RMSEs in a cross-validated fashion.

Surprisingly, as shown in Figure 7, incorporating backbone flexibility did not change the prediction significantly in most cases. As expected, the error on incorporating disruptive mutations (small $\rightarrow$ large) decreases. However, on some other mutations, incorporating backbone flexibility tends to increase RMSE slightly rather than decrease it (Tables 4 and 5). The overall test error using backbone flexibility increased slightly from the rigid backbone case, albeit not significantly ($p = 0.11$), to 1.69 kcal/mol. This is still less than the RMSEs of FOLDX and ROSETTA. We note that our data is dominated by mutations to alanines (a small residue). GOBLIN tends to underestimate binding free energy in mutations from a large residue to a small one, and does so to a greater degree when accounting for backbone flexibility. This partially explains the increase in RMSE.

There are three cases where backbone flexibility does tend to decrease RMSE (Table 5): when the wild-type residue is small, and the mutant is either medium or large, or when the wild-type has neutral charge, and the mutant is positively charged. These reductions in RMSE are due to more favorable enthalpies made possible through alternative backbones, as opposed to an entropic contribution. There are some notable exceptions to this trend, as seen in Figure 7. For example, the incorporation of backbone flexibility in mutations L18G of the SGP B : OMTKY complex and D101A of the HYHEL : HEL complex lead to more than 2 kcal/mol reduction in error. In both cases, the improvement was caused by accounting for the increase in

backbone entropy due to the mutation.

Further analysis from Molecular Dynamics simulations: To understand the cause of the outliers in our predictions, we investigated one of our largest outliers: R21P mutant (outlier) of the chymotrypsin : inhibitor complex (pdb id 1cho) by performing Molecular Dynamics simulations on the wild-type and the mutant in explicit solvent. The molecular dynamics simulations on the wild-type complex revealed a potential salt bridge between Arg 21 and Asp 35, and a hydrogen bond between the backbone of Arg 21 and Phe 41. Upon mutation to proline, the simulation revealed that both of these bonds were lost, resulting in a loop displacing further away from the serine protease inhibitor's hydrophobic pocket. This movement resulted in additional waters entering the binding site, further destabilizing the complex. This particular combination of changes (backbone conformational changes due to proline mutation, the loss of a salt-bridge, and solvent effects) explains why GOBLIN underestimates the change in free energy. In particular, the BACKRUB-generated backbone ensemble did not provide adequate sampling, GOBLIN's force-field doesn't account for electrostatics and doesn't explicitly model solvent.

To demonstrate the importance of backbone sampling, we replaced the BACKRUB-generated backbones in the graphical model with the 3,000 backbones generated via MD. The modified graphical model thus encoded a Boltzmann distribution over the side chains and this larger backbone ensemble. We then re-computed the free energies. We note that the complexity of performing this calculation scales linearly with the number of backbones, and is trivially distributed across a computer cluster. More importantly, there was a dramatic drop in error — from ≈ 8 kcal/mol to ≈ 1 kcal/mol. We note, however, that only 10 of the MD-generated backbones had substantial probability mass. This suggest that it may be necessary to generate a significantly larger backbone ensemble for such outliers. As a control experiment, we performed additional Molecular Dynamics simulations for L18GI, a mutation where GOBLIN was fairly accurate in its prediction when using the BACKRUB ensemble ($\Delta\Delta G^e$: 6 kcal/mol, $\Delta\Delta G^p$: 5.2

kcal/mol). On this mutant, the prediction did not change appreciably ($\Delta\Delta G^p$: 5.4 kcal/mol). This highlights the importance of obtaining a well-sampled ensemble of backbones when performing free energy calculations of protein-protein complexes and suggests that when given such a set of backbones, GOBLIN predictions can be substantially more accurate.

4 Discussion and Conclusions

The gain of inter-atomic interactions (hydrogen bonds, hydrophobic interactions, etc.) upon binding generally promotes the interactions between proteins, due to an overall decrease in enthalpy. These same interactions, however, also tend to decrease the entropy of the system, which is unfavorable. The free energy of binding reflects the opposition of these two effects. When the loss in entropy is more than compensated by the decrease in enthalpy, the interaction is favorable. GOBLIN makes residue-specific predictions regarding the changes in enthalpy and entropy upon mutation and binding. The results of our experiments are consistent with the idea that entropic contributions are significant, and therefore should not be ignored when studying protein-protein interactions. In particular, accounting for entropy results in a 9% to 12% decrease in RMSE, relative to an enthalpic or GMEC approximation, across a large benchmark set of protein-protein interactions.

Free energy calculations are usually performed via Molecular Dynamics simulations or using a knowledge-based approach. Molecular Dynamics simulations are too expensive to employ in large-scale studies, such as those considered here. Moreover, our experiments suggest that even best-of-breed knowledge-based methods, like FOLDX, may not capture some of the subtle interactions that affect entropy. GOBLIN introduces an effective alternative, which is to perform binding free energy calculations via variational inference. In terms of speed, GOBLIN runs in a few minutes, facilitating large-scale studies. GOBLIN also achieves higher accuracy than

FOLDX and ROSETTA. We believe that our variational approach to free energy calculations is more accurate because it explicitly accounts for changes in conformational entropy at the point of mutation and in the binding interface and, critically, because the Belief Propagation algorithm propagates the effects of those changes to neighboring residues that aren't directly involved in the mutation/binding.

Our graphical model-based framework provides an effective means for optimizing force-field parameters in a Bayesian fashion against experimentally measured changes in free energies. No other approach to free energy calculations provides this level of computational efficiency without relying on knowledge-based approximations to the entropic factors. We note that we have not been able to determine which complexes were used to optimize the force-field used in the version of FOLDX we used in our experiments. Our data set is among the largest of its kind, and so it is likely that there is substantial overlap between it and that used to optimize FOLDX. That is, it is possible that our experiments may have applied FOLDX to some of its own training data. This would tend to decrease FOLDX's RMSEs, while GOBLIN's RMSE are cross-validated values.

Surprisingly, our experiments employing both backbone and side-chain entropy did not yield lower RMSE than those employing side-chain entropy alone. This result can be explained, in part, by the fact that our benchmark set contains relatively rigid complexes [45], but there are a number of specific issues that should be addressed going forward. The first is that an MD simulation of one of our outliers revealed the importance of electrostatic interactions that are not presently modeled in GOBLIN's force field. The MD analysis also revealed the importance of solvent effects. The graphical model used by GOBLIN can, in principle, be extended to model explicit solvent. This can be accomplished, for example, by adding random variables corresponding to the solvent molecules which then interact with the protein (forming hydrogen bonds, etc.) in much the same way that the graphical model of the protein complex models the

interaction between different molecules. Finally, the modified version of the graphical model incorporating the MD-generated backbone ensemble suggests that substantial reductions in error can be obtained when a suitable collection of backbones is available. Obtaining such ensembles remains a challenge, although recent developments in backbone sampling strategies (e.g., [48, 49]) are promising. Moreover, recent advances in MD techniques (e.g., [50, 43, 51, 52, 53]) have dramatically reduced the cost of performing such simulations. In some contexts, it may be feasible to perform short simulations (like those performed here) in order to obtain a suitable ensemble of backbones for GOBLIN, rather than perform the long timescale simulations required by MD-based free energy calculations.

GOBLIN represents an ensemble of conformations using an undirected probabilistic graphical model. The size of the underlying ensemble is exponential in the number of residues. Variational inference is then used to perform binding free energy calculations. Our experiments were performed on complexes containing over 900 residues, and yet still run in minutes, demonstrating the benefits of a variational approach to free energy calculations. GOBLIN builds on previous applications of graphical models of all-atom structures (e.g., [33, 17, 16, 54]), but also makes several significant contributions. GOBLIN is the first undirected graphical model for studying all-atom protein-protein interactions. Algorithmically, our Bayesian approach to parameter optimization is also the first technique capable of finding parameters that minimize the difference between free energies, as opposed to internal energies (e.g., [22]).

There are number of algorithms for performing free energy calculations on graphical models (e.g., [55, 56, 57, 30]). The interested reader is directed to [58] for a recent review. An alternative approach to inference, including the pioneering work of Lee and Levitt [59, 60], computes estimates using a sampling scheme. Sampling is expensive, however, and so message-passing algorithms on graphical models, like those used by GOBLIN present an attractive alternative. We employed an algorithm called Belief Propagation [29], which computes the so-called Bethe

approximation of the free energy [30] via a message-passing scheme. Recent work [61] has shown that most message passing algorithms can be viewed as minimizing the divergence between the actual probability distribution and a family of suitably parametrized distributions. In previous work, we have investigated the use of algorithms for computing rigorous upper and lower bounds on the folding free energy using message-passing algorithms [62]. An interesting direction for future work would be the use of these algorithms in the context of binding free energies.

Finally, we note that one advantage of graphical models, versus tools like FOLDX and ROSETTA, is that they can be used to explicitly model the joint distribution over protein sequence and structure (e.g., [63, 64]). Integrating GOBLIN with models of protein families [65] and their interactions [66] would yield a framework for designing novel protein-protein interactions by performing inference over both sequence and structure. This is an exciting direction for future research.

Software GOBLIN is freely available to academic users in executable format and may be obtained by contacting the corresponding author. An open source version of the software will be released in the future.

Acknowledgments

This work is supported in part by US DOE Career Award (DE-FG02-05ER25696) (CJL), a grant from Microsoft Research (CJL), US NSF grant IIS-0905193 (CJL and CBK), and an Alfred P. Sloan Foundation Fellowship (CBK).

References

[1] C. Jarzynski. A nonequilibrium equality for free energy differences. *Physical Review Letters*, 78:2690, 1997.

[2] J. Srinivasan, T.E. Cheatham, P. Cieplak, P.A. Kollman, and D.A. Case. Continuum solvent studies of the stability of DNA, RNA, and phosphoramidate-DNA helices. *Journal of Americal Chemical Society*, 120(37):9401–9409, 1998.

[3] J. Åqvist, V.B. Luzhkov, and B.O. Brandsdal. Ligand binding affinities from MD simulations. *Accounts of Chemical Research*, 35(6):358–365, 2002.

[4] H. Gohlke, C. Kiel, and D.A. Case. Insights into protein-protein binding by binding free energy calculation and free energy decomposition for the Ras-Raf and Ras-RalGDS complexes. *Journal of Molecular Biology*, 330(4):891–913, 2003.

[5] H.-J. Böhm. The computer program LUDI: A new method for the de novo design of enzyne inhibitors. *Journal of Computer-Aided Molecular Design*, 6(1):61–78, 1992.

[6] I. Bahar, A. R. Atilgan, and B. Erman. Direct evaluation of thermal fluctuations in protein using a single parameter harmonic potential. *Folding and Design*, 2:173–181, 1997.

[7] M. L. Verdonk, J. C. Cole, M. J. Hartshorn, C. W. Murray, and R. D. Taylor. Improved protein-ligand docking using GOLD. *Proteins*, 52(4):609–623, 2003.

[8] I. Muegge. PMF scoring revisited. *Journal of Medicinal Chemistry.*, 49(20):5895–5902, 2006.

[9] A.R. Leach, B.K. Shoichet, and C.E. Peishoff. Prediction of protein-ligand interactions. Docking and scoring: successes and gaps. *Journal of Medicinal Chemistry.*, 49(20):5851–5855, 2006.

[10] G.L Warren, C.W. Andrews, A. Capelli, B. Clarke, J. LaLonde, M.H. Lambert, M. Lindvall, N. Nevins, S.F. Semus, S. Senger, G. Tedesco, I.D. Wall, J.M. Woolven, C.E. Peishoff, and M.S. Head. A critical assessment of docking programs and scoring functions. *Journal of Medicinal Chemistry.*, 49(20):5912–5931, 2006.

[11] R. Guerois, J. E. Nielsen, and L. Serrano. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *Journal of Molecular Biology*, 320:369–387, 2002.

[12] T. Kortemme and D. Baker. A simple physical model for binding energy hot spots in protein-protein complexes. *Proceedings of the National Academy of Sciences*, 99(22):14116–14121, 2002.

[13] J.M. Word, S.C. Lovell, J.S. Richardson, and D.C. Richardson. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *Journal of Molecular Biology*, 285(4):1735–1747, 1999.

[14] C. A. Smith and T. Kortemme. Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. *Journal of Molecular Biology*, 380(4):742–756, July 2008.

[15] P. Clifford. Markov random fields in statistics. In G. R. Grimmett and D. J. A. Welsh, editors, *Disorder in Physical Systems. A Volume in Honour of John M. Hammersley*, pages 19–32, Oxford, 1990. Clarendon Press.

[16] H. Kamisetty, E. P. Xing, and C. J. Langmead. Free energy estimates of all-atom protein structures using generalized belief propagation. *Journal of Computational Biology*, 15(7):755–766, September 2008.

[17] H. Kamisetty, E.P. Xing, and C.J. Langmead. Free energy estimates of all-atom protein structures using generalized belief propagation. In *Proceedings of the 7th Annual International Conference on Research in Computational Biology (RECOMB)*, pages 366–380, 2007.

[18] A. Canutescu, A. A. Shelenkov, and R. L. Dunbrack Jr. A graph theory algorithm for protein side-chain prediction. *Protein Science*, 12:2001–2014, 2003.

[19] I.W. Davis, W.B. Arendall, D.C. Richardson, and J.S. Richardson. The backrub motion: How protein backbone shrugs when a sidechain dances. *Structure*, 14(2):265–274, 2006.

[20] B. Kuhlman, G. Dantas, G. C. Ireton, G. Varani, B. L. Stoddard, and D. Baker. Design of a novel globular protein fold with atomic-level accuracy. *Science*, 302(5649):1364–1368, November 2003.

[21] T. Kortemme, A. V. Morozov, and D. Baker. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. *Journal of Molecular Biology*, 326(4):1239–1259, February 2003.

[22] C. Yanover, O. Schueler-Furman, and Y. Weiss. Minimizing and learning energy functions for side-chain prediction. In *Proceedings of the 7th Annual International Conference on Research in Computational Biology (RECOMB)*, pages 381–395, 2007.

[23] P. Dagum and R. M. Chavez. Approximating probabilistic inference in bayesian belief networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(3):246–255, 1993.

[24] H. A. Bethe. Statistical theory of superlattices. *Proceedings of the Royal Society London A*, 150:552–575, 1935.

[25] R. Kikuchi. A theory of cooperative phenomena. *Physical Review*, 81:988–1003, 1951.

[26] T. Morita. Cluster variation method for non-uniform Ising and Heisenberg models and spin-pair correlation function. *Progress of Theoretical Physics*, 85:243 – 255, 1991.

[27] T. Morita, T. M. Suzuki, K. Wada, and M. Kaburagi. Foundations and applications of cluster variation method and path probability method. *Progress of Theoretical Physics Supplement*, 115, 1994.

[28] J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, 1988.

[29] J. Pearl. Fusion, propagation, and structuring in belief networks. *Artificial Intelligence*, 29(3):241–288, 1986.

[30] J.S. Yedidia, W.T. Freeman, and Y. Weiss. Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Transactions on Information Theory*, 51:2282–2312, 2005.

[31] J. W. Ponder and F. M. Richards. Tertiary templates for proteins: Use of packing criteria in the enumeration of allowed sequences for different structural classes. *Journal of Molecular Biology*, 193(4):775–791, February 1987.

[32] M. J. McGregor, S. A. Islam, and M. J. Sternberg. Analysis of the relationship between side-chain conformation and secondary structure in globular proteins. *Journal of Molecular Biology*, 198(2):295–310, November 1987.

[33] C. Yanover and Y. Weiss. Approximate inference and protein folding. *Advances in Neural Information Processing Systems (NIPS)*, pages 84–86, 2002.

[34] J. Xu. Rapid protein side-chain packing via tree decomposition. In *Proceedings of the 9th Annual International Conference on Computational Biology (RECOMB)*, pages 423–439, 2005.

[35] C. L. Kingsford, B. Chazelle, and M. Singh. Solving and analyzing side-chain positioning problems using linear and integer programming. *Bioinformatics*, 21:1028–1036, 2005.

[36] M. Karplus and J.N. Kushick. Method for estimating the configurational entropy of macromolecules. *Macromolecules*, 14(2):325–332, 1981.

[37] E. T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106(4):620–630, May 1957.

[38] A. Doig. Thermodynamics of amino acid side-chain internal rotations. *Biophysical Chemistry*, 61(2-3):131–141, October 1996.

[39] E. T. Jaynes. Prior probabilities. *IEEE Transactions on Systems Science and Cybernetics*, 4(3):227–241, 1968.

[40] Alexander Benedix, Caroline M. Becker, Bert L. de Groot, Amedeo Caflisch, and Rainer A. Bockmann. Predicting free energy changes using structural ensembles. *Nature Methods*, 6(1):3–4, January 2009.

[41] Rosetta commons. `http://www.rosettacommons.org/`.

[42] Schrödinger, LLC, New York, NY. MAESTRO, version 9.0, 2009.

[43] K. J. Bowers, E. Chow, H. Xu, R. O. Dror, M. P. Eastwood, B. A. Gregersen, J. L. Klepeis, I. Kolossvary, M. A. Moraes, F. D. Sacerdoti, J. K. Salmon, Y. Shan, and D. E. Shaw. Scalable algorithms for molecular dynamics simulations on commodity clusters. *Supercomputing Conference*, page 43, 2006.

[44] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The protein data bank. *Nucleic Acids Research*, 28:235–242, 2000.

[45] S.M. Lu, W. Lu, M.A. Qasim, others, and M. Laskowski, Jr. Predicting the reactivity of proteins from their sequence alone: Kazal family of protein inhibitors of serine proteinases. *Proceedings of the National Academy of Sciences*, 98(4):1410–1415, February 2001.

[46] J. Pons, A. Rajpal, and J. F. Kirsch. Energetic analysis of an antigen/antibody interface: alanine scanning mutagenesis and double mutant cycles on the HyHEL-10/lysozyme interaction. *Protein Science*, 8(5):958–968, May 1999.

[47] G.D. Friedland, A.J. Linares, C.A. Smith, and T. Kortemme. A simple model of backbone flexibility improves modeling of side-chain conformational variability. *Journal of Molecular Biology*, 380(4):757–774, July 2008.

[48] T. Hamelryck, J.T. Kent, and A. Krogh. Sampling realistic protein conformations using local structural bias. *PLoS Computational Biology*, 2(9):e131, 2006.

[49] W. Boomsma, K.V. Mardia, C.C. Taylor, J. Ferkinghoff-Borg, A. Krogh, and T. Hamelryck. A generative, probabilistic model of local protein structure. *Proceedings of the National Academy of Sciences*, 105(26):8932–8937, 2008.

[50] J. C. Philips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. V. Kale, and K. Schulten. Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry.*, 26(16):1781–1801, 2005.

[51] V. S. Pande, I. Baker, J. Chapman, S. P. Elmer, S. Khaliq, S. M. Larson, Y. M. Rhee, M. R. Shirts, C.D. Snow, E. J. Sorin, and B. Zagrovic. Atomistic protein folding simulations

on the submillisecond time scale using worldwide distributed computing. *Biopolymers*, 68(1):91–109, 2003.

[52] J.E. Stone, J. C. Phillips, P. L. Freddolino, D. J. Hardy, L. G. Trabuco, and K. Schulten. Accelerating molecular modeling applications with graphics processors. *Journal of Computational Chemistry.*, 28:2618–2640, 2007.

[53] D. E. Shaw, M. M. Deneroff, R. O. Dror, J. S. Kuskin, R. H. Larson, J. K. Salmon, C. Young, B. Batson, K. J. Bowers, J. C. Chao, M. P. Eastwood, J. Gagliardo, J. P. Grossman, C. R. Ho, D. J. Ierardi, I. Kolossváry, J. L. Klepeis, T. Layman, C. McLeavey, M. A. Moraes, R. Mueller, E. C. Priest, Y. Shan, J. Spengler, M. Theobald, B. Towles, and S. C. Wang. Anton, a special-purpose machine for molecular dynamics simulation. In *ISCA '07: Proceedings of the 34th annual International Symposium on Computer Architecture*, pages 1–12, New York, NY, USA, 2007. ACM.

[54] F. DiMaio, A. Soni, G.N. Phillips Jr., and J.W. Shavlik. Creating all-atom protein models from electron-density maps using particle-filtering methods. *Bioinformatics*, 23:2851–2858, 2007.

[55] Lawrence Saul and Michael I. Jordan. Exploiting tractable substructures in intractable networks. In *Advances in Neural Information Processing Systems (NIPS)*, volume 8, pages 486–492, 1995.

[56] Martin J. Wainwright, Tommi S. Jaakkola, and Alan S. Willsky. A new class of upper bounds on the log partition function. In *Uncertainty in Artificial Intelligence*, volume 51, pages 536–543, 2002.

[57] E.P. Xing, M.I. Jordan, and S. Russell. A generalized mean field algorithm for variational inference in exponential families. In *Proceedings of the 19th Annual Conference on Uncertainty in Artificial Intelligence*, pages 583–591, 2003.

[58] M.J. Wainwright and M.I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.

[59] C. Lee. Predicting protein mutant energetics by self-consistent ensemble optimization. *Journal of Molecular Biology*, 236:918–939, 1994.

[60] C. Lee and M. Levitt. Accurate prediction of the stability and activity effects of site-directed mutagenesis on a protein core. *Nature*, 352:448–451, 1991.

[61] T. Minka. Divergence measures and message passing. Technical Report MSR-TR-2005-173, Microsoft Research, 2005.

[62] H. Kamisetty and C.J. Langmead. Conformational free energy of protein structures: Computing upper and lower bounds. In *Proceedings of the Structural Bioinformatics and Computational Biophysics (3DSIG)*, pages 23–24, 2008.

[63] M. Fromer and C. Yanover. A computational framework to empower probabilistic protein design. *Bioinformatics*, 24(13):i214–222, 2008.

[64] H. Kamisetty, B. Ghosh, C. Bailey-Kellogg, and C.J. Langmead. Modeling and inference of sequence-structure specificity. In *Proceedings of the of the 8th International Conference on Computational Systems Bioinformatics (CSB)*, pages 91–101, 2009.

[65] J. Thomas, N. Ramakrishnan, and C. Bailey-Kellogg. Graphical models of residue coupling in protein families. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 5(2):183–197, 2008.

[66] J. Thomas, N. Ramakrishnan, and C. Bailey-Kellogg. Graphical models of protein-protein interaction specificity from correlated mutations and interaction data. *Proteins*, 76(4):911–929, 2009.

[67] R. Wallis, K.Y. Leung, M.J. Osborne, R. James, G.R. Moore, and C. Kleanthous. Specificity in protein-protein recognition: conserved Im9 residues are the major determinants of stability in the colicin E9 DNase-Im9 complex. *Biochemistry*, 37(2):476–485, 1998.

[68] J.S. Yedidia, W.T. Freeman, and Y. Weiss. Generalized Belief Propagation. *Advances in Neural Information Processing Systems (NIPS)*, 13:689–695, 2000.

Tables

Table 1: **Datasets for mutant protein-protein complexes.**

Wt PDB id	Partner A	Partner B	# residues in A and B	# mutants in A, B
1sgr	OMTKY	SGP B	236	150,0
1cho	OMTKY	Chymotrypsin	291	170,0
1ppf	OMTKY	Human LE	274	170, 0
1a22 (AS)	HGH	HGHBP	429	34,29
1gc1 (AS)	CD4	GP120	920	49,0
1dan (AS)	BCF VII-A	TF	587	20,23
1bxi (AS)	E9 Dnase	IM 9	212	30,0
3hfm (AS)	HYHEL	HEL	558	12,13

"AS": alanine-scanning experiments

Table 2: **Learned Force-field Parameters**

Name	Learned Value
w_{ljatr}	0.46
w_{ljrep}	0.70
w_{hb}	0.11
w_{rot}	0.23 kcal mol^{-1}
w_{sasa}	0.027 kcal mol^{-1} A$^{\circ -2}$
w_{iwt}	0.0006 kcal mol^{-1} A$^{\circ -2}$

Parameters corresponding to terms from ROSETTA's force-field are dimensionless since the corresponding force-field terms already have units of energy.

Table 3: $\Delta\Delta G$ **(kcal/mol) root mean squared errors.**

Method	Overall RMSE (std err)	95% RMSE (std err)	90% RMSE (std err)
GOBLIN	1.63 (0.06)	1.27 (0.05)	1.10 (0.04)
GOBLIN-E	1.80 (0.06)	1.40 (0.05)	1.22 (0.04)
GOBLIN-H	1.85 (0.09)	1.42 (0.08)	1.22 (0.07)
FOLDX	1.82	1.42	1.20
ROSETTA	2.54	1.92	1.68

Root mean squared error (RMSE) for GOBLIN, FOLDX, and ROSETTA. GOBLIN-E and GOBLIN-H refer to the RMSE for the GMEC and enthalpy, respectively. The values for GOBLIN and its variants are cross-validated test errors with the standard errors for these estimates reported in parentheses. The final two columns are the RMSE after the 5% and 10% worst outliers have been removed, respectively.

Table 4: **RMSEs (kcal/mol) according to charge of amino acid: negative (D, E), neutral (A, N, C, Q, G, H, I, L, M, F, P, S, T, W, Y, V), and positive (R, K). Values in brackets indicate net change in RMSE upon incorporating backbone flexibility.**

Wild-Type \Mutant	Negative	Neutral	Positive
Negative	1.71 [0.02]	1.20 [0.05]	1.38 [0.03]
Neutral	2.40 [0.04]	1.56 [0.11]	1.85 [-0.21]
Positive	1.57 [0.22]	1.80 [0.05]	0.96 [0.3]

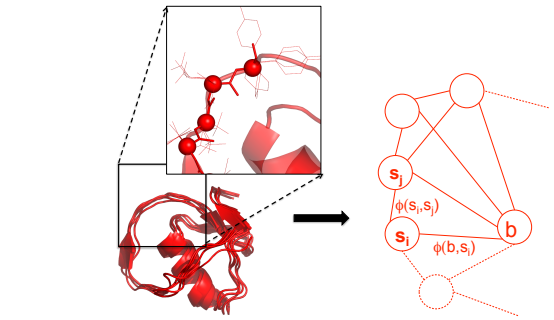
Table 5: **RMSEs (kcal/mol) according to volume of amino acid: small (A, G, S), medium (N, D, C, Q, E, H, I, L, K, M, P, T, V), and large (R, F, W, Y). Values in brackets indicate net change in RMSE upon incorporating backbone flexibility.**

Wild-Type\Mutant	Small	Medium	Large
Small	0.75 [0.05]	1.57 [-0.29]	2.04 [-0.24]
Medium	1.47 [0.04]	1.64 [0.01]	1.45 [0.18]
Large	1.81 [0.35]	2.22 [0.05]	1.16 [0.44]

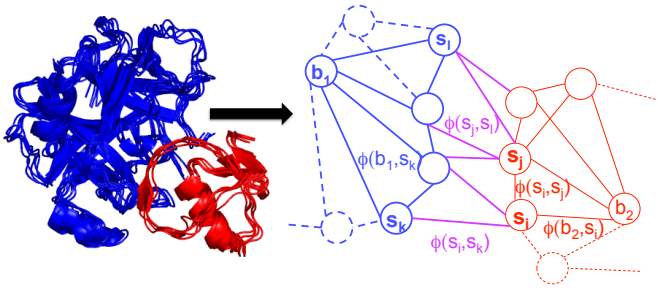
Table 6: **Outliers**

Mutant	Error (kcal/mol)	Complex	Possible Reasons
R21PI	-7.68	1sgr	Mutation to proline; solvent interactions
R21PI	-7.37	1cho	Mutation to proline; solvent interactions
Y50AH	-7.08	3hfm	Large-scale rearrangement [46]
R21PI	-6.88	1ppf	Mutation to proline; solvent interactions
K96AY	-6.29	3hfm	Large-scale rearrangement; loss of salt-bridge [46]
N32AL	-5.47	3hfm	Large-scale rearrangement [46]
Y33AH	-5.46	3hfm	Large-scale rearrangement [46]
D51AA	-5.21	1bxi	Loss of strong electrostatic interaction [67]
L18PI	-5.10	1cho	Mutation to proline; possible destabilization of complex

Figures



(A)



(B)

Figure 1: **Graphical models of protein complexes.** (A) Turkey ovomucoid third domain. (Left) Ensemble of backbones, with inset showing ensemble of side-chains for one backbone. (Right) Graphical model of the backbone and side-chain ensembles. For visual clarity, only the subscripts of the random variables are shown. The node labeled b corresponds to a random variable over the backbone ensemble, while the remaining nodes correspond to random variables over rotameric side-chain conformations. Edges capture intra-molecular interactions (vdW, hydrogen bonds, etc.) with ϕ functions according to a molecular mechanics force-field. The graphical model encodes a Boltzmann distribution over conformations in terms of the ϕ functions. Dashed nodes and edges represent a subset of the positions and interactions in the rest of the protein that GOBLIN models but have been omitted in this figure for simplicity. (B) Complex of chymotrypsin with turkey ovomucoid third domain. (Left) Ensemble of backbones. (Right) Graphical model. It combines the inhibitor model (red) with an analogous model for chymotrypsin (blue), and introduces inter-molecular edges (purple) with ϕ functions for inter-molecular interaction terms. This model encodes a Boltmann distribution over complex conformations.

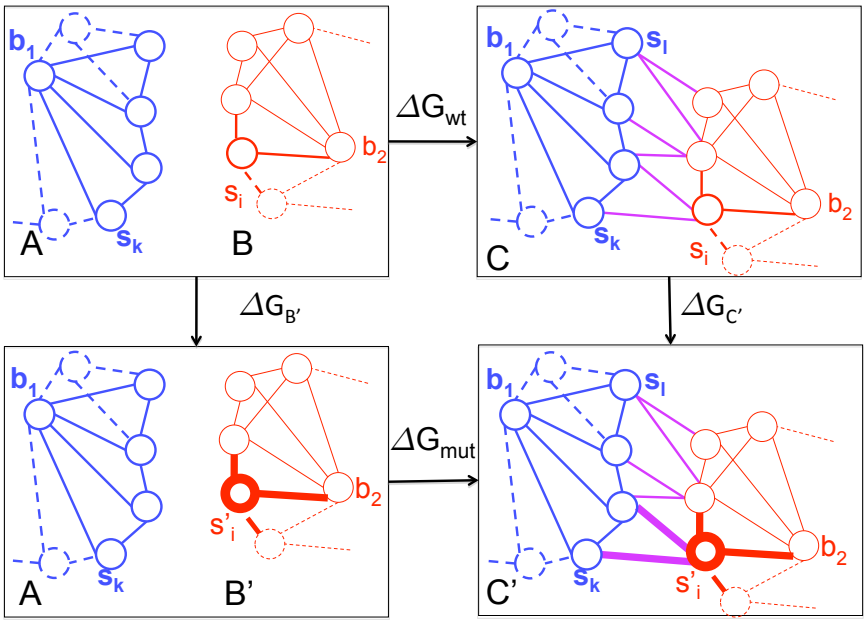


Figure 2: **Free energies in graphical models of protein complexes.** (Top-left) Two separate graphical models, A and B, encoding the wild-type apo forms of two proteins. (Top-right) A graphical model, C, encoding the wild-type complex. Binding free energies are obtained by computing the free energies of the three models, G_A, G_B, and G_C, and then calculating $\Delta G = G_C - (G_A + G_B)$. (Bottom-left, bottom-right) Graphical models of corresponding mutant forms. The mutated position and the interactions that are affected by it are shown in thick lines. While the energetic effect (via these interactions) is local, the entropic effect can be distal. GOBLIN accounts for both effects by performing variational inference to compute $\Delta G_{A'}$ and $\Delta G_{C'}$. $\Delta\Delta G$s are obtained by computing the binding free energy of the mutant, $\Delta G_{mut} = G_{C'} - (G_{A'} + G_B)$, and then calculating $\Delta\Delta G = \Delta G_{mut} - \Delta G$.

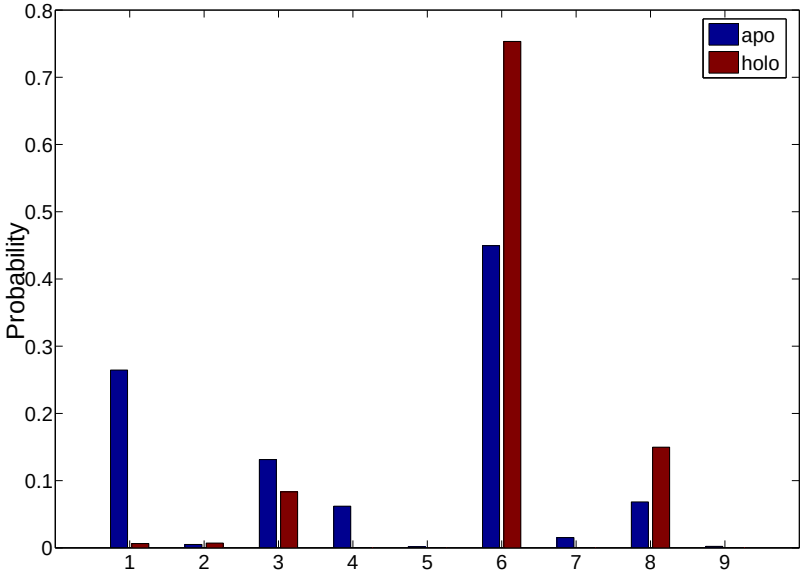
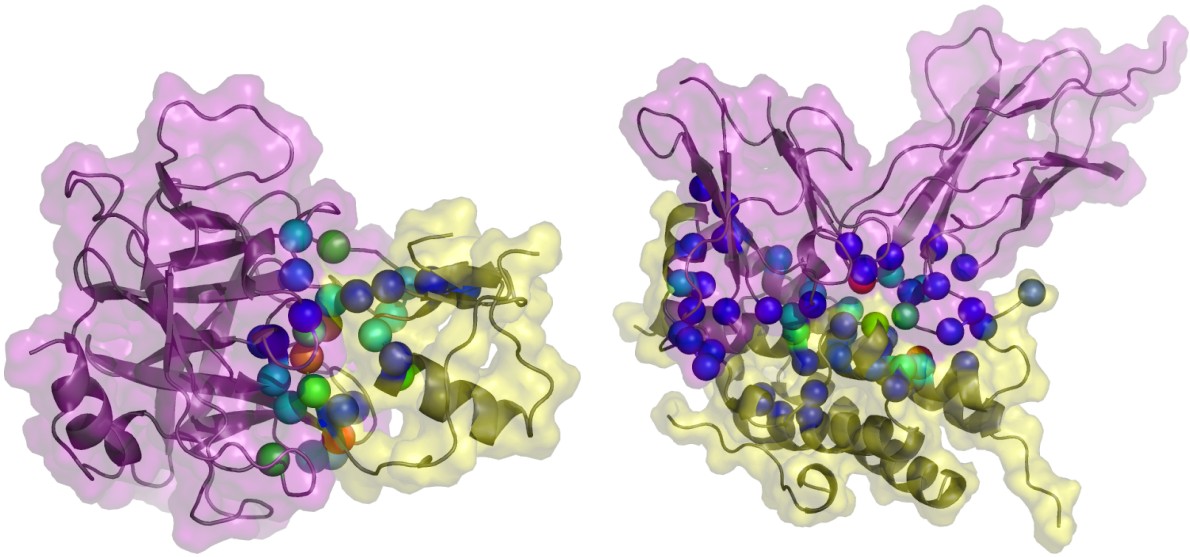
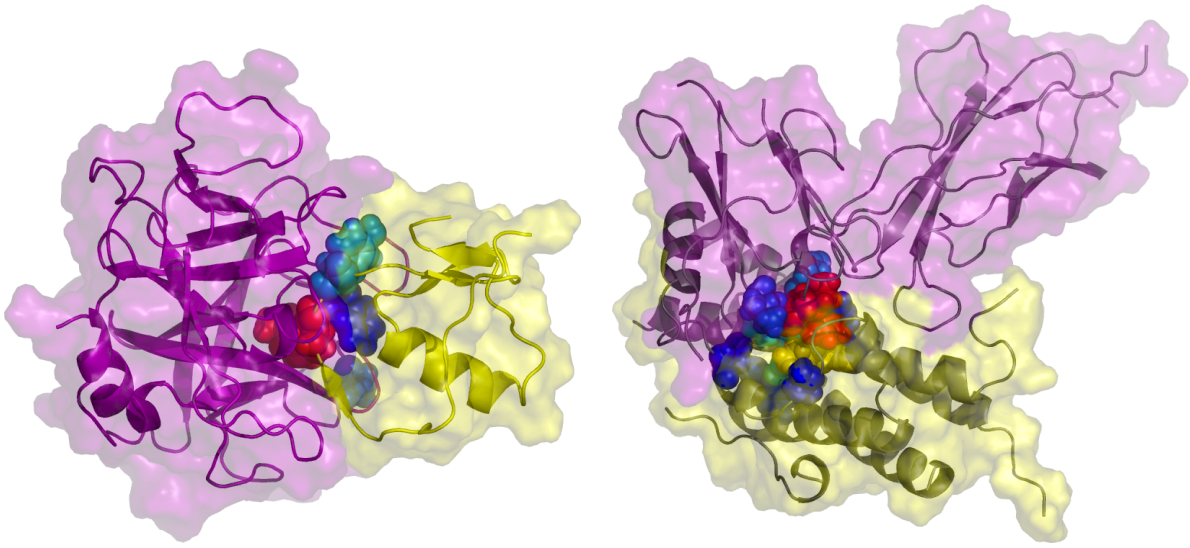


Figure 3: **Effect of rotameric-occupancy due to binding.** Change in rotameric probabilities for Trp 304 in the HGH-HGHBP complex. Trp has 9 rotamers in the rotamer library we use, 3 for each of $\chi_1 = \{60°, -180°, 60°\}$ respectively. The blue bars show the rotameric occupancies for the apo structure while the red bars show the occupancies in the holo structure. Upon binding, the probability mass redistributes.



HLE-OMTKY (wild-type) HGH-HGHBP (wild-type)



HLE-OMTKY (L18G mutant) HGH-HGHBP (D171A mutant)

Figure 4: **Localized evaluation of change in entropy.** (Top) Change in entropy upon binding; (Bottom) Change in entropy upon mutation. (Left) HLE-OMTKY, wild-type and with L18G mutation; (Right) HGH-HGHBP, wild-type and with D171A mutation. The surface color distinguishes the partners. Spheres mark C_α atoms of residues whose marginal entropy changes by more than $0.1k_B$ (yielding $< 10\%$ of the residues).

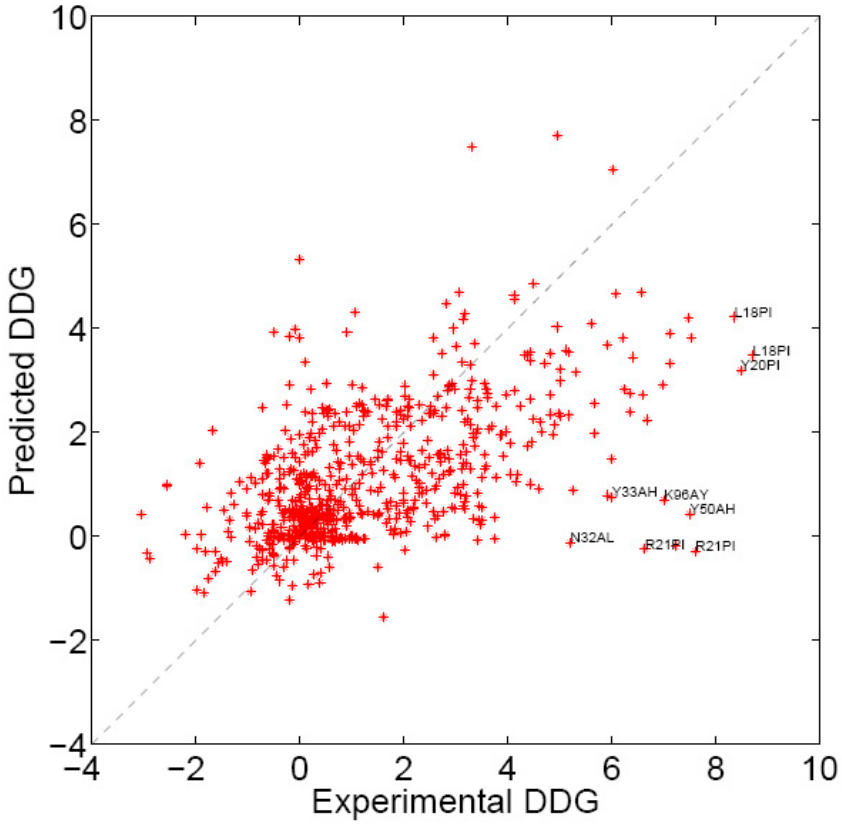


Figure 5: **Scatter plot comparing** GOBLIN**'s predictions with experimental values**. The correlation coefficient (R^2) was 0.56 across the entire dataset. The nine worst outliers are labeled with the mutation and chain id. When the worst 5% (resp. 10%) outliers are removed, $R^2 = 0.66$ (resp. $R^2 = 0.70$).

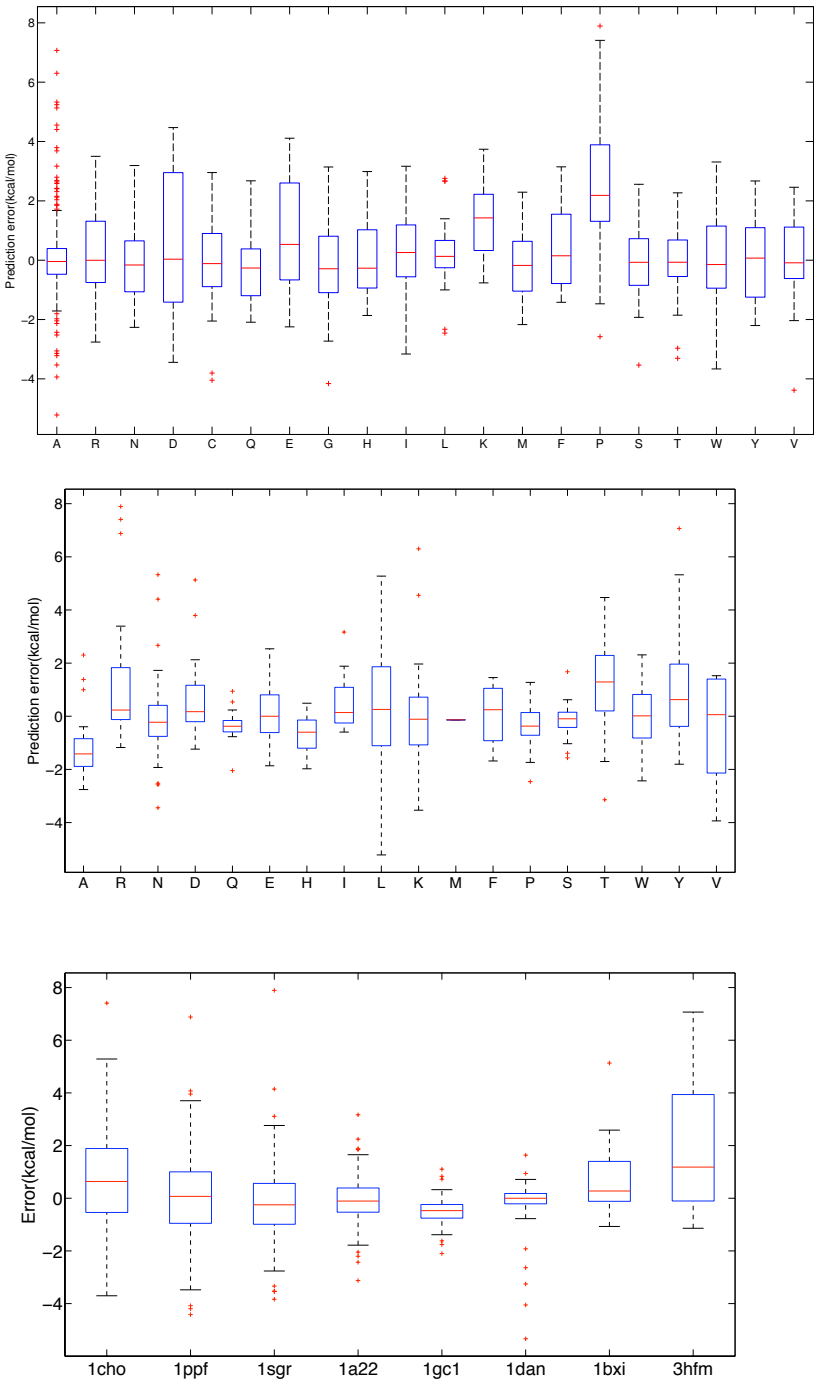


Figure 6: **Characterization of prediction error**. (Top) According to mutant residue type; (Middle) according to wild-type residue type; (Bottom) according to complex (in decreasing order by number of mutations). See text for an explanation of box plots. The error is actual minus predicted, so positive indicates an under-prediction, while negative means an over-prediction.

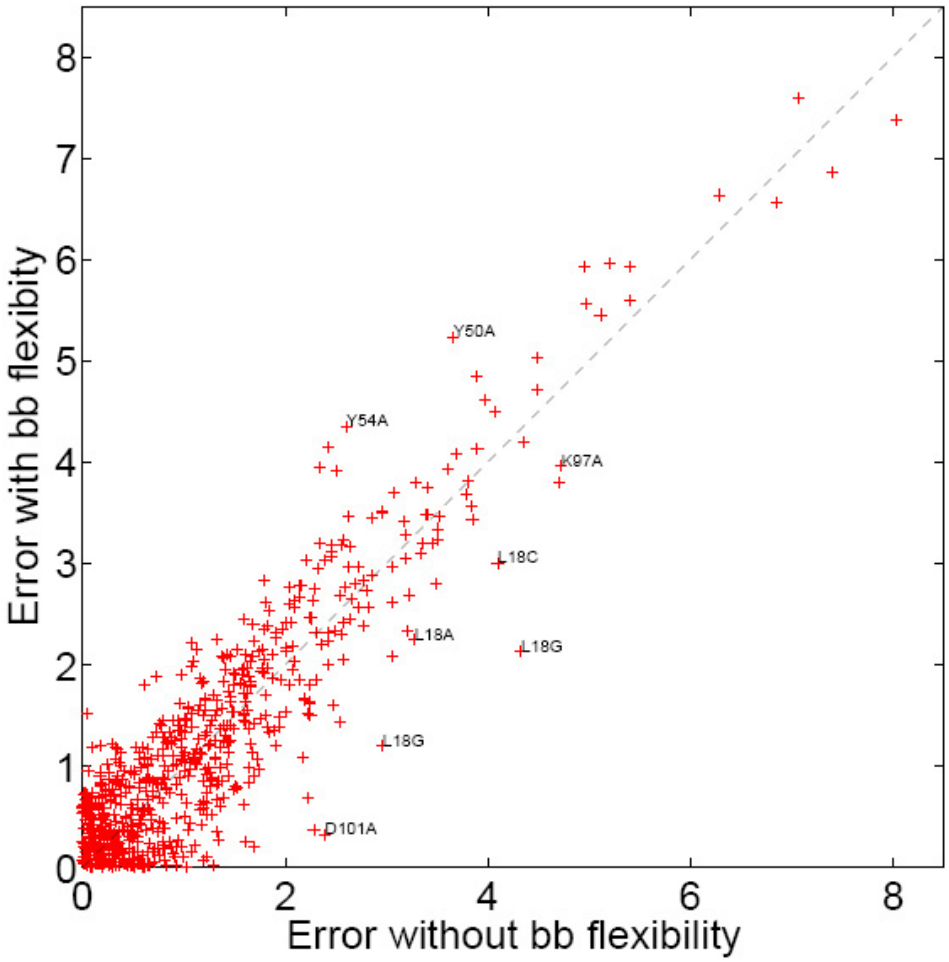


Figure 7: **Incorporation of backbone flexibility.** GOBLIN prediction error with (y-axis) and without (x-axis) backbone flexibility. Marked outliers indicate wild-type amino acid type, residue position and mutant amino acid type.

Appendix: Belief Propagation Example

Belief Propagation is a probabilistic inference algorithm that can be used to approximate the free energy of a distribution. In the form being presented here, the algorithm keeps track of marginal distributions (called "beliefs") over variables (indexed with i) and edges (i, j). It must be pointed out that the edges are undirected; (i, j) and (j, i) therefore refer to the same edge. The notation used here is one of many equivalent alternatives [30, 68].

Messages are passed between variables and edges to update beliefs until convergence. These messages at time step m^{t+1} are defined with respect to the beliefs at time step b^t as follows:

1. The message from variable i to edge (i, j): $m^{t+1}_{i\to(i,j)} = b^t_i / m^t_{(i,j)\to i}$
2. The message from edge (i, j) to variable i: $m^{t+1}_{(i,j)\to i} = \sum_j b^t_{(i,j)} / m^t_{i\to(i,j)}$

The beliefs themselves are a function of the *current* messages and are calculated as:

1. Belief of variable i, $b^t_i \propto \phi_i \prod_{j\in N(i)} m^t_{(i,j)\to i}$
2. Belief of edge (i, j) (joint belief over (i, j)): $b^t_{(i,j)} \propto \phi_{(i,j)} m^t_{i\to(i,j)} m^t_{j\to(i,j)}$

Consider the toy Markov Random Field shown in figure 8. As stated previously in equation 7, the potential functions $\phi_1, \ldots, \phi_{1,2}, \ldots$ are Boltzmann factors of the self and interaction energies.

At $t = 0$, let us initialize all messages. The beliefs will then be proportional to the corresponding potentials. After normalizing to ensure the beliefs sum to one, this results in

$$b^0_1 = [\ 0.5 \quad 0.5\ ], \qquad b^0_2 = [\ 0.375 \quad 0.625\ ], \qquad b^0_3 = [\ 0.5 \quad 0.5\ ]$$

$$b^0_{1,2} = \begin{bmatrix} 0.07 & 0.27 \\ 0.07 & 0.59 \end{bmatrix}, \qquad b^0_{2,3} = \begin{bmatrix} 0.05 & 0.67 \\ 0.24 & 0.04 \end{bmatrix}, \qquad b^0_{1,3} = \begin{bmatrix} 0.48 & 0.07 \\ 0.19 & 0.26 \end{bmatrix}$$

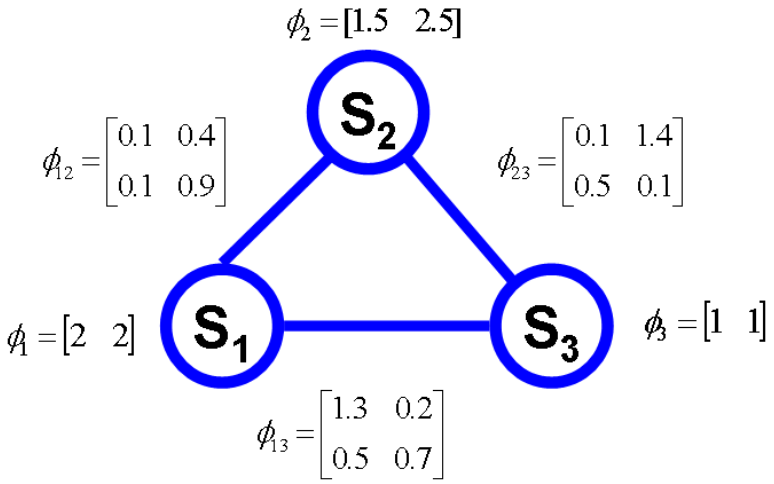


Figure 8: Example Markov Random Field

In the next iteration of Belief Propagation, the messages will then be

$$m^1_{1\to(1,2)} = [\ 0.5 \quad 0.5\ ]; m^1_{2\to(1,2)} = [\ 0.5 \quad 0.5\ ]; m^1_{(1,2)\to 1} = [\ 0.67 \quad 0.33\ ]; m^1_{(1,2)\to 2} = [\ 0.13 \quad 0.87\ ]$$

$$m^1_{1\to(1,3)} = [\ 0.5 \quad 0.5\ ]; m^1_{3\to(1,3)} = [\ 0.5 \quad 0.5\ ]; m^1_{(1,3)\to 1} = [\ 0.56 \quad 0.44\ ]; m^1_{(1,3)\to 3} = [\ 0.67 \quad 0.33\ ]$$

$$m^1_{2\to(2,3)} = [\ 0.38 \quad 0.62\ ]; m^1_{3\to(2,3)} = [\ 0.5 \quad 0.5\ ]; m^1_{(2,3)\to 2} = [\ 0.71 \quad 0.29\ ]; m^1_{(2,3)\to 3} = [\ 0.29 \quad 0.71\ ]$$

These can be used to compute the beliefs at iteration 1 and this process repeated until the beliefs converge. In this case, using Eq. 8 and 9 at convergence yields $G = -1.21$. On this model, exhaustive enumeration yields the exact value of the free energy: $G_{exact} = -1.12$. The corresponding values for the partition function are $Z = 3.3376, Z_{exact} = 3.068$.