SUPPLEMENTARY NOTE

Table of Contents

FORMULATION OF THE MULTISCALE STATISTICAL MECHANICAL FRAMEWORK

This section provides a formal treatment of the multiscale statistical mechanical framework. For each level of the framework (domain, protein, and perturbation), we will describe the statistical mechanical ensemble representing that level, the associated Hamilonian (energy function), if any, and the derivation of relevant probabilities used for predicting biologically important quantities for comparison with experiments.

Domain

Canonical Ensemble

We begin with the domain-level model, which describes an interaction between a single SH2 domain and a single tyrosine phosphosite. We consider the equilibrium state of the reaction representing the binding of two domains A and B:

$$A + B \leftrightarrow AB$$

Schematically we will represent such a reaction using the diagram shown in Supplementary Fig. 1 (domain ensemble). The two states comprise a canonical ensemble representing the unbound and bound macrostates of the two domains.

Hamiltonian

Let A, B be two domains, one containing an SH2 domain and the other a tyrosine phosphosite, without distinction. Let $\mathcal{H}(A+B)$ be the Hamiltonian representing the energy of the two domains in their unbound state, and $\mathcal{H}(AB)$ be the Hamiltonian of the bound state. We assume that the energies of the unbound domains are additive, i.e.:

$$\mathcal{H}(A+B) = \mathcal{H}(A) + \mathcal{H}(B)$$

Let A_i (and similarly for B) denote the identity of the ith amino acid in domain A, and let $|A|$ be the amino acid sequence length of domain A. We will assume the existence of a function $u_i^{(1)}\colon \{\text{Amino Acids}\} \to \mathbb{R}$ which assigns every amino acid at position i in a domain of type A an energy, and associate the following first-order Hamiltonian with the unbound state of a domain A:

$$\mathcal{H}(A) = \sum\nolimits_{i=1}^{|A|} u_i^{(1)}(A_i)$$

We note that $u_i^{(1)}$ is an overloaded function since its energies depend on the type of domain. As we are considering two types, SH2 domains and phosphodomains, one can treat $u_i^{(1)}$ as accepting a type argument such that:

$$u_i^{(1)}(A_i) = 1\{A_i \in \text{SH2}\} u_i^{(\text{SH2})}(A_i) + 1\{A_i \in \text{pY}\} u_i^{(\text{pY})}(A_i)$$

We also assume the existence of a family of functions $b_{i_1,\dots,i_k}^{(k)}\colon \{\text{Amino Acids}\}^k \to \mathbb{R}$ which assign to every combination of k amino acids at positions $i_1, \dots, i_k$ an energy, and associate the following second-order Hamiltonian with the bound state AB:

$$\mathcal{H}(AB) = \sum\nolimits_{i=1}^{|A|} b_i^{(1)}(A_i) + \sum\nolimits_{i=1}^{|B|} b_i^{(1)}(B_i) + \sum\nolimits_{i=1}^{|A|} \sum\nolimits_{j=1}^{|B|} b_{i,j}^{(2)}(A_i, A_j)$$

The functions $\left\{b_{i_1,\dots,i_k}^{(k)}\right\}$ are overloaded similarly to $u_i^{(1)}$. For both the first- and second-order Hamiltonians, we use a canonical set of 106 residue positions to represent SH2 domains, derived using a Hidden Markov Model (HMM) alignment of all known SH2 domains. For tyrosine phosphosites, we represent them using a 9-residue window centered around the phosphorylated tyrosine.

Canonical Ensemble Probabilities

The Boltzmann probability associated with the unbound state of the canonical ensemble is:

$$p^{(\mathrm{ce})}(A+B) = \frac{e^{-\frac{1}{kT}(\mathcal{H}(A)+\mathcal{H}(B)-TS_u)}}{e^{-\frac{1}{kT}(\mathcal{H}(A)+\mathcal{H}(B)-TS_u)} + e^{-\frac{1}{kT}(\mathcal{H}(AB)-TS_b)}}$$

The probability of the bound state is:

$$p^{(\mathrm{ce})}(AB) = \frac{e^{-\frac{1}{kT}(\mathcal{H}(AB)-TS_b)}}{e^{-\frac{1}{kT}(\mathcal{H}(A)+\mathcal{H}(B)-TS_u)} + e^{-\frac{1}{kT}(\mathcal{H}(AB)-TS_b)}}$$

T is temperature, k is Boltzmann's constant, and S_u and S_b are the entropies (non-sequence specific) associated with the unbound and bound states, respectively. The above probabilities reflect the relative concentrations of the domains A and B in their bound vs. unbound state at equilibrium.

Protein

Canonical Ensemble

We now turn our attention to the protein-level model, which describes an interaction between two proteins, each of which may be comprised of one or more SH2 domains, tyrosine phosphosites, or any combination thereof ("sites"). We consider the equilibrium state of the reaction representing the binding of two proteins $\mathcal{A}$ and $\mathcal{B}$:

$$\mathcal{A} + \mathcal{B} \leftrightarrow \mathcal{AB}$$

The set of permissible bound states of $\mathcal{AB}$ are ones in which each protein site is bound to at most one site on the other protein, and binding occurs only between SH2 domains and phosphosites. The schematic in Supplementary Fig. 1 (protein ensemble) illustrates this requirement for a pair of proteins, one comprised of two phosphosites and one SH2 domain, and the other comprised of two SH2 domains. The right macrostate represents the collection of states that correspond to the two proteins being bound.

Hamiltonian

Let $\mathcal{A}, \mathcal{B}$ be two proteins containing some combination of SH2 domains and phosphosites, indexed by $A \in \mathcal{A}$ and $B \in \mathcal{B}$, respectively. We associate the following Hamiltonian with the unbound state of $\mathcal{A}$ (similarly for $\mathcal{B}$):

$$\mathcal{H}(\mathcal{A}) = \sum\nolimits_{A \in \mathcal{A}} \mathcal{H}(A)$$

Let $\mathcal{E}(\mathcal{A}, \mathcal{B})$ be the set of permissible bound states of $\mathcal{A}$ and $\mathcal{B}$. For every $E \in \mathcal{E}(\mathcal{A}, \mathcal{B})$, let $E(\mathcal{A})$ and $E(\mathcal{B})$ be the set of sites in $\mathcal{A}$ and $\mathcal{B}$ that are bound in E, respectively. We associate the following Hamitonian with E:

$$\mathcal{H}(E) = \sum\nolimits_{(A,B)\in E} \mathcal{H}(AB) + \sum\nolimits_{A\in\mathcal{A}\setminus E(\mathcal{A})} \mathcal{H}(A) + \sum\nolimits_{B\in\mathcal{B}\setminus E(\mathcal{B})} \mathcal{H}(B)$$

$A \in \mathcal{A}\setminus E(\mathcal{A})$ represents the set of sites in $\mathcal{A}$ that are unbound in E, and similarly for B and $\mathcal{B}$. The values of the bound and unbound Hamiltonians are as defined in the previous section. We finally associate the following Hamiltonian with the ensemble representing the bound macrostate of $\mathcal{AB}$:

$$\mathcal{H}(\mathcal{AB}) = \sum\nolimits_{E\in\mathcal{E}(\mathcal{AB})} \mathcal{H}(E)$$

Canonical Ensemble Probabilities

We are now ready to derive expressions for the Boltzmann probabilities. We assume that each SH2 domain contributes one entropic term depending on whether it is bound (S_b) or unbound (S_u). The contribution per SH2 domain is fixed in magnitude. With that assumption, we have that the partition function over all permissible bound and unbound states is:

$$Z(\mathcal{A},\mathcal{B}) = e^{-\frac{1}{kT}(\mathcal{H}(\mathcal{A})+\mathcal{H}(\mathcal{B})-T|\mathcal{AB}|S_u)} + \sum\nolimits_{E\in\mathcal{E}(\mathcal{AB})} e^{-\frac{1}{kT}(\mathcal{H}(E)-T(|E|S_b+(|\mathcal{AB}|-|E|)S_u))}$$

$|\mathcal{AB}|$ is the total number of SH2 domains in the complex, and $|E|$ is the number of bound SH2 domains in the state E.

The Boltzmann probability corresponding to a state representing one bound configuration E is:

$$p^{(\mathrm{ce})}(E) = \frac{e^{-\frac{1}{kT}(\mathcal{H}(E)-T(|E|S_b+(|\mathcal{AB}|-|E|)S_u))}}{Z(\mathcal{A},\mathcal{B})}$$

The probability for the ensemble of states representing the bound macrostate $\mathcal{AB}$ is:

$$p^{(\mathrm{ce})}(\mathcal{AB}) = \frac{\sum_{E\in\mathcal{E}(\mathcal{AB})} e^{-\frac{1}{kT}(\mathcal{H}(E)-T(|E|S_b+(|\mathcal{AB}|-|E|)S_u))}}{Z(\mathcal{A},\mathcal{B})} = \sum\nolimits_{E\in\mathcal{E}(\mathcal{AB})} p^{(\mathrm{ce})}(E)$$

Finally the probability for the unbound state $\mathcal{A}+\mathcal{B}$ is:

$$p^{(\mathrm{ce})}(\mathcal{A}+\mathcal{B}) = \frac{e^{-\frac{1}{kT}(\mathcal{H}(\mathcal{A})+\mathcal{H}(\mathcal{B})-|\mathcal{AB}|TS_u)}}{Z(\mathcal{A},\mathcal{B})}$$

Perturbation

Canonical Ensemble

We now consider ensembles comprised of partitioned states. In each state, one partition corresponds to a pre-perturbation condition and the other to a post-perturbation condition. A perturbation is a mutation in one or more sequence positions in the domains of the ensemble, i.e. changes in A_is or B_js for some set of is and js. The two partitions are thermodynamically insulated to represent the semantics of a somatic mutation. I.e. we are assuming that the pre- and post-perturbation versions of a protein are not kinetically accessible to one another. The full ensemble includes all possible combinations of unbound and bound states in the pre- and post-perturbation conditions and will be denoted $\mathcal{F}$. The state space for a pair of proteins, one comprised of a single phosphosite and the other of two SH2 domains, is shown in Supplementary Fig. 2a, with the perturbation occurring in one of the SH2 domains.

We define two subsets of this ensemble. The *causal change* ensemble refers to the subset of states in which a perturbed domain changes from being bound to unbound or vice-versa, and nothing else changes. The *qualitative change* ensemble can be a *gain of function*, in which case it refers to the subset of states in which proteins $\mathcal{A}$ and $\mathcal{B}$ are unbound in the pre-perturbation condition and bound in the perturbation case, or alternatively a *loss of function*, in which case it refers to the subset of states in which proteins $\mathcal{A}$ and $\mathcal{B}$ are bound in the pre-perturbation condition and unbound in the perturbation case. The diagram in Supplementary Fig. 2b illustrates this using the same set of proteins as in Supplementary Fig. 2a, where the intersection of the two ensembles are states that are both *causal* and *qualitative*.

Let $\mathcal{A}^*$ and $\mathcal{B}^*$ denote the proteins $\mathcal{A}$ and $\mathcal{B}$ in the post-perturbation partitions. Let $\mathcal{G}^*$ be the subset of $\mathcal{E}(\mathcal{A}^*,\mathcal{B}^*)$ in which one or more of the perturbed domains are bound and nothing else is bound, and let $\mathcal{G}$ be the subset of $\mathcal{E}(\mathcal{A},\mathcal{B})$ in which the perturbed domains, in their pre-perturbation condition, are bound and nothing else is bound. The space of partitioned states is the Cartesian product $\mathcal{F} = \Big((\{\mathcal{A}+\mathcal{B}\} \cup \mathcal{E}(\mathcal{A},\mathcal{B})) \times (\{\mathcal{A}^*+\mathcal{B}^*\} \cup \mathcal{E}(\mathcal{A}^*,\mathcal{B}^*))\Big)$. We are interested in two equilibria. The first is between $\{(\mathcal{A}+\mathcal{B}, G^*)\}_{G^* \in \mathcal{G}^*}$, which is the subset of states that form the intersection of the *causal change* and *qualitative gain of function* ensembles, and $\mathcal{F} \backslash \{(\mathcal{A}+\mathcal{B}, G^*)\}_{G^* \in \mathcal{G}^*}$. This equilibrium represents the causal and qualitative gain of function perturbation. The second equilibrium is between $\{(G, \mathcal{A}^*+\mathcal{B}^*)\}_{G \in \mathcal{G}}$, which is the subset of states that form the intersection of the *causal change* and *qualitative loss of function* ensembles, and $\mathcal{F} \backslash \{(G, \mathcal{A}^*+\mathcal{B}^*)\}_{G \in \mathcal{G}}$. This equilibrium represents the causal and qualitative loss of function perturbation. Supplementary Fig. 1 (network-mutation ensemble) illustrates the equilibrium for the gain of function case. Note that the ensemble of perturbed states may contain more than one state if there are multiple potential partners for the perturbed domain or if there are multiple mutations.

Canonical Ensemble Probabilities

We are now ready to derive quantities corresponding to the perturbation probabilities of protein-protein interactions. We begin by noting that the probability of a partitioned state in the ensemble factorizes into the product of the partitions in the state. This is because, by construction, the two partitions are thermodynamically insulated and are thus independent. We will denote Boltzmann probabilities of the pre-perturbation partition by $p^{(\mathrm{ce})}_{\mathcal{A},\mathcal{B}}$ and the post-perturbation partition by $p^{(\mathrm{ce})}_{\mathcal{A}^*,\mathcal{B}^*}$. Note that these represent two distinct probability distributions, as the partition function for each is different. Even probabilities of equivalent states, for example the unbound states $\mathcal{A}+\mathcal{B}$ and $\mathcal{A}^* + \mathcal{B}^*$, can yield different values because their underlying ensembles are different.

The probability of being in the causal and qualitative gain of function ensemble, which we denote by $\mathcal{A}+\mathcal{B} \rightarrow \mathcal{A}^*\mathcal{B}^*$, is:

$$p^{(\mathrm{ce})}(\mathcal{A}+\mathcal{B} \rightarrow \mathcal{A}^*\mathcal{B}^*) = p^{(\mathrm{ce})}_{\mathcal{A},\mathcal{B}}(\mathcal{A}+\mathcal{B}) \left(\sum\nolimits_{G^* \in \mathcal{G}^*} p^{(\mathrm{ce})}_{\mathcal{A}^*,\mathcal{B}^*}(G^*) \right)$$

The probability of being in the causal and qualitative loss of function ensemble, which we denote by $\mathcal{AB} \rightarrow \mathcal{A}^* + \mathcal{B}^*$, is:

$$p^{(\text{ce})}(\mathcal{AB} \rightarrow \mathcal{A}^* + \mathcal{B}^*) = \left(\sum\nolimits_{G \in \mathcal{G}} p_{\mathcal{A},\mathcal{B}}^{(\text{ce})}(G)\right) p_{\mathcal{A}^*,\mathcal{B}^*}^{(\text{ce})}(\mathcal{A}^* + \mathcal{B}^*)$$

Since these two ensembles are mutually exclusive, the probability of being in either is their sum.

Perturbation Probabilities

Let $\mathcal{M}$ be the set of all possible perturbations, i.e. all sets of sequence mutations that transform $(\mathcal{A},\mathcal{B})$ into $(\mathcal{A}^*,\mathcal{B}^*)$. This is a large but finitely countable set. Let a *mutational process* be a probability distribution over $\mathcal{M}$, and denote it $\mathcal{P}$. Mutational processes can represent disease states such as cancer. The probability of a perturbation transforming $(\mathcal{A},\mathcal{B})$ into $(\mathcal{A}^*,\mathcal{B}^*)$ under $\mathcal{P}$ will be denoted by $p_{\mathcal{P}}^{(\text{genomic})}\big((\mathcal{A},\mathcal{B}) \rightarrow (\mathcal{A}^*,\mathcal{B}^*)\big)$.

We are now ready to derive the probability that a gain of function perturbation will occur in $\mathcal{P}$ for the two proteins $\mathcal{A}$ and $\mathcal{B}$, which we denote by $p_{\mathcal{P}}^{(\text{gain})}(\mathcal{A},\mathcal{B})$:

$$p_{\mathcal{P}}^{(\text{gain})}(\mathcal{A},\mathcal{B}) = \sum\nolimits_{\left((\mathcal{A},\mathcal{B}) \rightarrow (\mathcal{A}^*,\mathcal{B}^*)\right) \in \mathcal{M}} p_{\mathcal{P}}^{(\text{genomic})}\big((\mathcal{A},\mathcal{B}) \rightarrow (\mathcal{A}^*,\mathcal{B}^*)\big) p^{(\text{ce})}(\mathcal{A} + \mathcal{B} \rightarrow \mathcal{A}^*\mathcal{B}^*)$$

Similarly the probability that a loss of function perturbation will occur is:

$$p_{\mathcal{P}}^{(\text{loss})}(\mathcal{A},\mathcal{B}) = \sum\nolimits_{\left((\mathcal{A},\mathcal{B}) \rightarrow (\mathcal{A}^*,\mathcal{B}^*)\right) \in \mathcal{M}} p_{\mathcal{P}}^{(\text{genomic})}\big((\mathcal{A},\mathcal{B}) \rightarrow (\mathcal{A}^*,\mathcal{B}^*)\big) p^{(\text{ce})}(\mathcal{AB} \rightarrow \mathcal{A}^* + \mathcal{B}^*)$$

Finally the probability that a perturbation will occur, in either form, is:

$$p_{\mathcal{P}}^{(\text{perturb})}(\mathcal{A},\mathcal{B}) = p_{\mathcal{P}}^{(\text{gain})}(\mathcal{A},\mathcal{B}) + p_{\mathcal{P}}^{(\text{loss})}(\mathcal{A},\mathcal{B})$$

These quantities can be used to compute the expected effects of mutational processes such as cancer on the SH2 phosphosignaling network.

TRAINING OF THE MULTISCALE STATISTICAL MECHANICAL FRAMEWORK

The multiscale statistical mechanical framework relies on many quantities whose values are *a priori* unknown. In this section we describe the approach we have taken to infer values for these quantities. In principle, any source of experimental data that corresponds to measurable quantities at the domain, protein, or perturbation levels can be used to constrain the model at the appropriate level, yielding an inverse problem that can be solved using standard techniques if sufficient data exists. Moreover, the probabilistic quantities described in the previous section all have formal experimental analogues. In practice, the vast majority of data exists in the form of quantitative (e.g. K_a or fold enrichment) and qualitative (binary bound/unbound) measurements on domain-level binding. As a result, our approach primarily relies on using sparse reconstruction techniques to infer from such measurements the underlying residue-residue energies, and then use these energies to compute higher-level quantities. We will describe this process at each level of the framework.

Domain

We use experimental measurements to constrain the value of the domain-level binding probability $p^{(\text{ce})}(AB)$ for a broad range of SH2 domains and tyrosine phosphodomains. By constraining this value for many interactions we are then able to use optimization techniques to infer approximate values for quantities comprised of the functions $\left\{u_i^{(1)}\right\}$ and $\left\{b_{i_1,\dots,i_k}^{(k)}\right\}$. Due to an indeterminancy in our formulation it is not possible to recover exactly those values, nor is it necessary, as we shall discuss shortly. We begin by describing the experimental sources available and their processing, then provide a treatment of the indeterminancy of the model, and finally present our sparse reconstruction and cross-validation approaches.

Experimental Sources

We use a combination of qualitative and quantitative binding data at the level of single SH2 domain / phosphosite binding (Hause et al., 2012; Koytiger et al., 2013; Liu et al., 2010; Tinti et al., 2013). For a pair of domains A, B, we denote its corresponding empirical probability by $p^{(\mathrm{emp})}(AB)$. For qualitative binary data, we set $p^{(\mathrm{emp})}(AB)$ to 0 for non-interacting domains and to 1 for interacting domains. For quantitative data such as association constants, we note that $p^{(\mathrm{emp})}(AB) \propto K_a$, where the constant of proportionality depends on the choice of standard state. Due to differences in experimental assays however, we found that the incorporation of raw quantitative data results in lower performance. As a result we binarized quantitative data so that a K_d value less than 1μM yields a positive interaction $\left(p^{(\mathrm{emp})}(AB) = 1\right)$ and a K_d value greater than 1mM yields a negative interaction $\left(p^{(\mathrm{emp})}(AB) = 0\right)$. Data points with values between these thresholds were discarded. The table below gives a break down of positive and negative data points from each source:

Source	Positive Data Points	Negative Data Points
MacBeath	2,451	65,600
Jones	356	6,012
Nash	497	6,515
Cesareni	13,166	273,359
LT	390	0

To align SH2 sequences with SH2-peptide structural complexes in the Protein Data Bank we use the SCOP Hidden Markov Model (Murzin et al., 1995).

Formulation Indeterminancy

The domain-level Boltzmann probabilities we have derived so far contain an intrinsic indeterminacy. This can be seen through the following set of elementary algebraic manipulations:

$$p^{(\mathrm{ce})}(AB) = \frac{e^{-\frac{1}{kT}(\mathcal{H}(AB) - TS_b)}}{e^{-\frac{1}{kT}(\mathcal{H}(A)+\mathcal{H}(B)-TS_u)} + e^{-\frac{1}{kT}(\mathcal{H}(AB)-TS_b)}}$$

$$p^{(\mathrm{ce})}(AB) = \frac{e^{-\frac{1}{kT}\left(\mathcal{H}(AB)-\left(\mathcal{H}(A)+\mathcal{H}(B)\right)-T(S_b-S_u)\right)}}{1 + e^{-\frac{1}{kT}\left(\mathcal{H}(AB)-\left(\mathcal{H}(A)+\mathcal{H}(B)\right)-T(S_b-S_u)\right)}}$$

$$p^{(\mathrm{ce})}(AB) = \frac{e^{-\frac{1}{kT}\left(\sum_{i=1}^{|A|} b_i^{(1)}(A_i)+\sum_{i=1}^{|B|} b_i^{(1)}(B_i)+\sum_{i=1}^{|A|}\sum_{j=1}^{|B|} b_{i,j}^{(2)}(A_i,A_j)-\left(\sum_{i=1}^{|A|} u_i^{(1)}(A_i)+\sum_{i=1}^{|B|} u_i^{(1)}(B_i)\right)-T(S_b-S_u)\right)}}{1 + e^{-\frac{1}{kT}\left(\sum_{i=1}^{|A|} b_i^{(1)}(A_i)+\sum_{i=1}^{|B|} b_i^{(1)}(B_i)+\sum_{i=1}^{|A|}\sum_{j=1}^{|B|} b_{i,j}^{(2)}(A_i,A_j)-\left(\sum_{i=1}^{|A|} u_i^{(1)}(A_i)+\sum_{i=1}^{|B|} u_i^{(1)}(B_i)\right)-T(S_b-S_u)\right)}}$$

$$p^{(\mathrm{ce})}(AB) = \frac{e^{-\frac{1}{kT}\left(\sum_{i=1}^{|A|}\left(b_i^{(1)}(A_i)-u_i^{(1)}(A_i)\right)+\sum_{i=1}^{|B|}\left(b_i^{(1)}(B_i)-u_i^{(1)}(B_i)\right)+\sum_{i=1}^{|A|}\sum_{j=1}^{|B|}b_{i,j}^{(2)}(A_i,A_j)-T(S_b-S_u)\right)}}{1+e^{-\frac{1}{kT}\left(\sum_{i=1}^{|A|}\left(b_i^{(1)}(A_i)-u_i^{(1)}(A_i)\right)+\sum_{i=1}^{|B|}\left(b_i^{(1)}(B_i)-u_i^{(1)}(B_i)\right)+\sum_{i=1}^{|A|}\sum_{j=1}^{|B|}b_{i,j}^{(2)}(A_i,A_j)-T(S_b-S_u)\right)}}$$

The key issue is that the terms $b_i^{(1)}$ and $u_i^{(1)}$ always co-occur as differences, and so do S_b and S_u, which makes it impossible to uniquely identify the individual values of each quantity. Consequently we will define $\Delta H_i^{(1)}(A_i) \equiv b_i^{(1)}(A_i) - u_i^{(1)}(A_i)$ to denote the first-order change in enthalpy resultant from binding. For consistency we will also define $\Delta H_{i,j}^{(2)}\left(A_i, A_j\right) \equiv b_{i,j}^{(2)}\left(A_i, A_j\right) - 0$ to be the second-order change in enthalpy. Finally we define $\Delta S \equiv S_b - S_u$ to be the change in entropy due to binding.

Using the new notation we can rewrite the previous expression in terms of changes in enthalpy and entropy:

$$p^{(\mathrm{ce})}(AB) = \frac{e^{-\frac{1}{kT}\left(\sum_{i=1}^{|A|}\Delta H_i^{(1)}(A_i)+\sum_{i=1}^{|B|}\Delta H_i^{(1)}(B_i)+\sum_{i=1}^{|A|}\sum_{j=1}^{|B|}\Delta H_{i,j}^{(2)}(A_i,A_j)-T\Delta S\right)}}{1+e^{-\frac{1}{kT}\left(\sum_{i=1}^{|A|}\Delta H_i^{(1)}(A_i)+\sum_{i=1}^{|B|}\Delta H_i^{(1)}(B_i)+\sum_{i=1}^{|A|}\sum_{j=1}^{|B|}\Delta H_{i,j}^{(2)}(A_i,A_j)-T\Delta S\right)}}$$

The above expression is comprised exclusively of terms representing changes between the unbound and bound states. The quantities $\left\{\Delta H_i^{(1)}(a)\right\}_{a\in\mathrm{AA}}^{i\in\mathrm{SH2}}, \left\{\Delta H_i^{(1)}(a)\right\}_{a\in\mathrm{AA}\setminus\{C\}}^{i\in\mathrm{pY}}, \left\{\Delta H_{i,j}^{(2)}(a,b)\right\}_{(a,b)\in\mathrm{AA}\times(\mathrm{AA}\setminus\{C\})}^{(i,j)\in\mathrm{SH2}\times\mathrm{pY}}$ are sets defined over the amino acids (denoted by AA; C is cysteine), the residue positions of SH2 domains (denoted by SH2), the residue positions of the phosphosite window (denoted by pY), and pairs of such positions. These quantities, along with ΔS, are the only uniquely determinable terms in the model, and it is these terms that we will approximate from experimental data sources using sparse reconstruction techniques. We will collectively refer to them as $\{\Delta\}$.

Sparse Reconstruction

As discussed in the previous section, the $\{\Delta\}$ quantities are uniquely identifiable (given sufficient data) free parameters in our domain-level model. We seek to determine their values by solving the following optimization problem:

$$\{\Delta\} = \underset{\{\Delta\}}{\operatorname{argmin}}\left[\operatorname{reg}_{\lambda,\gamma}(\{\Delta\}, D) + \sum\nolimits_{(A,B)\in\mathrm{data}} \operatorname{loss}(A, B|\{\Delta\})\right]$$

We define the loss term to be:

$$\operatorname{loss}(A, B|\{\Delta\}) \equiv -\left(p^{(\mathrm{emp})}(AB)\ln\left(p_{\{\Delta\}}^{(\mathrm{ce})}(AB)\right) + \left(1 - p^{(\mathrm{emp})}(AB)\right)\ln\left(1 - p_{\{\Delta\}}^{(\mathrm{ce})}(AB)\right)\right)$$

This definition is the negative of the conditional log-likelihood of the data using the Boltzmann probabilities as the model. We subscript $p_{\{\Delta\}}^{(\mathrm{ce})}(AB)$ with $\{\Delta\}$ to indicate that these probabilities depend on the energy terms defined by $\{\Delta\}$. Minimizing this loss function will find the values of $\{\Delta\}$ that maximize the conditional likelihood of the data, and is the standard formulation used in logistic regression.

Unfortunately we do not have sufficient data to solve this problem uniquely in a reliable fashion and must rely on regularization techniques that favor sparse models. This is consistent with the physicochemistry of protein-protein interactions, which rely on a small number of residue-residue interactions for most of the enthalpic contribution of binding. We define the regularization term to be:

$$\mathrm{reg}_{\lambda,\gamma}(\{\Delta\},\mathrm{D}) \equiv \lambda\left(\sum\nolimits_{i\in\mathrm{SH2}}\sum\nolimits_{j\in\mathrm{pY}}\sum\nolimits_{a\in\mathrm{AA}}\sum\nolimits_{b\in\mathrm{AA}\{C\}}\left|\gamma\left(\Delta H_{i,j}^{(2)}(a,b)\right)D_{i,j}^{(2)}(a,b)+(1-\gamma)\right|\right)$$

The above function acts as a weighted ℓ_1 regularizer. The weights $\left\{\Delta D_{i,j}^{(2)}\right\}$ correspond to the spatial distance between the two closest atoms of the residues at position i and j in the SH2 domain and phosphosite, respectively, as derived by averaging distances from 25 structural complexes. Longer distances result in larger weights, which favor assigning smaller values to the energies, consistent with the physics of energy potentials. Interactions between residues that are very far apart (>15Å) were entirely removed from consideration for computational efficiency reasons.

λ and γ are two metaparameters in the model. λ sets the scale for the entire regularization term, while γ controls the degree to which the model penalizes energy terms in a distance-specific manner. A value of 1 corresponds to exclusively distance-dependent penalization, while a value of 0 corresponds to no distance-dependent penalization, which is the standard lasso behavior. Due to the computational complexity of simultaneously fitting λ and γ using a grid search, we instead carried out a preliminary analysis of model sensitivity varying γ values alone and choosing the best performing λ for each value of γ. We observed that higher values of γ resulted in better performance, but that at γ = 1 (the maximum value possible) our model experienced numerical instability which prevented it from running. Consequently we choose the largest value of γ that converged, which was γ = 0.9. Setting $\gamma = 0$ will give the standard lasso behavior. We found that larger γs perform better, but setting $\gamma = 1$ results in numerical instabilities. As a result we set $\gamma = 0.9$. The performance results resultant from different values of the metaparameters were determined using nested cross-validation (discussed next).

The final regularization weights are rescaled to sum to the total number of parameters, to remove scale as a free metaparameter. In addition, we experimented with raising $D_{i,j}^{(2)}$ to higher powers but found that unity gave the best performance. The first-order terms are not regularized and are thus always included in the model.

Using the above formulation we were able to recover robust values for $\{\Delta\}$, which enables us to perform all higher-level calculations. Note that the values of k and T are set to unity for convenience, as the inference procedure automatically inferes the energy scale.

Cross-Validation

We used a nested cross-validation approach to estimate model parameters and metaparameters (Supplementary Fig. 3a). We first divide the data set into n sets, which we term outer CV sets. Then, for each outer CV set, we divide the data outside the set into m sets, which we term inner CV sets. This results in a total of $n \times m$ inner CV sets and n outer CV sets.

For every inner CV set we perform the optimization described in the previous section using the data that is outside both the inner CV set and its corresponding outer CV set. This allows us to estimate the model parameters $\{\Delta\}$. However, since we do not know *a priori* the optimal values of the metaparameters λ and γ, we perform multiple optimizations for different combinations of λ and γ and then choose the one that performs best on the inner CV set. Once the optimal value is chosen, we retrain the model on the entirety of the data set but leave out the outer CV set. Using the previously determined values of λ and γ, we then test the performance of the model on the outer CV set. This insures that all model parameters and metaparameters are optimized without access to the subset of the data on which the model is tested.

We used this approach when testing overall performance, for which we set $n = 8$ and $m = 7$. When we tested the domain transferability of the model, we used two approaches. We first chose the outer CV sets to be comprised exclusively of data for one SH2 domain, which resulted in $n = 108$ and we set $m = 7$. We also grouped SH2 domains in $n = 8$ groups which had no overlap in their SH2 domains, and set $m = 7$. Finally when we tested the data set transferability of the model, we chose the outer CV sets to be comprised exclusively from either one of the four high-throughput experimental data sets or the combination of low-throughput positives-only data and data with multiple sources of evidence. This resulted in $n = 5$, and we set $m = 8$.

Evaluation of Other Methods

We evaluated the performance of other SH2 domain methods using the default settings from their respective websites or codes (Kundu et al., 2013; Li et al., 2008). Since these methods are unable to make predictions for arbitrary domain sequences, we restricted the test sets to interactions predictable by all methods. When evaluating PrePPI (Zhang et al., 2012), we had to perform the comparison on the protein level as its predictions are protein-based. We treated two proteins as interacting if any of their domains or phophosites interacted in the experimental data set, and used the resulting data set as the basis for assessing PrePPI's performance. As PrePPI was not designed or optimized for making such predictions, we included it merely to provide a representative of the performance of general PPI approaches on predicting SH2-peptide interactions.

Protein

All the calculations for the protein-level model use the domain-level energies whose inference was described in the previous section. The indeterminancy of the domain-level model extends to the protein-level model, which we show here using the $p^{(\mathrm{ce})}(E)$ term. The result extends to the other terms trivially.

Starting with $p^{(\mathrm{ce})}(E)$, we have:

$$p^{(\mathrm{ce})}(E) = \frac{e^{-\frac{1}{kT}(\mathcal{H}(E)-T(|E|S_b+(|\mathcal{AB}|-|E|)S_u))}}{Z(\mathcal{A},\mathcal{B})}$$

$$= \frac{e^{-\frac{1}{kT}\left(\sum_{(A,B)\in E}\mathcal{H}(AB)+\sum_{A\in\mathcal{A}\setminus E(\mathcal{A})}\mathcal{H}(A)+\sum_{B\in\mathcal{B}\setminus E(\mathcal{B})}\mathcal{H}(B)-T(|E|S_b+(|\mathcal{AB}|-|E|)S_u)\right)}}{e^{-\frac{1}{kT}(\sum_{A\in\mathcal{A}}\mathcal{H}(A)+\sum_{B\in\mathcal{B}}\mathcal{H}(B)-T|\mathcal{AB}|S_u)}+\sum_{E\in\mathcal{E}(\mathcal{AB})}e^{-\frac{1}{kT}\left(\sum_{(A,B)\in E}\mathcal{H}(AB)+\sum_{A\in\mathcal{A}\setminus E(\mathcal{A})}\mathcal{H}(A)+\sum_{B\in\mathcal{B}\setminus E(\mathcal{B})}\mathcal{H}(B)-T(|E|S_b+(|\mathcal{AB}|-|E|)S_u)\right)}}$$

$$= \frac{e^{-\frac{1}{kT}(\sum_{(A,B)\in E}\mathcal{H}(AB)+\sum_{A\in\mathcal{A}\setminus E(\mathcal{A})}\mathcal{H}(A)+\sum_{B\in\mathcal{B}\setminus E(\mathcal{B})}\mathcal{H}(B)-T(|E|S_b+(|\mathcal{AB}|-|E|)S_u)-\sum_{A\in\mathcal{A}}\mathcal{H}(A)-\sum_{B\in\mathcal{B}}\mathcal{H}(B)+T|\mathcal{AB}|S_u)}}{1+\sum_{E\in\mathcal{E}(\mathcal{AB})}e^{-\frac{1}{kT}(\sum_{(A,B)\in E}\mathcal{H}(AB)+\sum_{A\in\mathcal{A}\setminus E(\mathcal{A})}\mathcal{H}(A)+\sum_{B\in\mathcal{B}\setminus E(\mathcal{B})}\mathcal{H}(B)-T(|E|S_b+(|\mathcal{AB}|-|E|)S_u)-\sum_{A\in\mathcal{A}}\mathcal{H}(A)-\sum_{B\in\mathcal{B}}\mathcal{H}(B)+T|\mathcal{AB}|S_u)}}$$

$$= \frac{e^{-\frac{1}{kT}\left(\sum_{(A,B)\in E}\mathcal{H}(AB)+\sum_{A\in\mathcal{A}\setminus E(\mathcal{A})}\mathcal{H}(A)+\sum_{B\in\mathcal{B}\setminus E(\mathcal{B})}\mathcal{H}(B)-\sum_{A\in\mathcal{A}\setminus E(\mathcal{A})}\mathcal{H}(A)-\sum_{A\in E(\mathcal{A})}\mathcal{H}(A)-\sum_{B\in\mathcal{B}\setminus E(\mathcal{B})}\mathcal{H}(B)-\sum_{B\in E(\mathcal{B})}\mathcal{H}(B)-|E|T(S_b-S_u)\right)}}{1+\sum_{E\in\mathcal{E}(\mathcal{AB})}e^{-\frac{1}{kT}\left(\sum_{(A,B)\in E}\mathcal{H}(AB)+\sum_{A\in\mathcal{A}\setminus E(\mathcal{A})}\mathcal{H}(A)+\sum_{B\in\mathcal{B}\setminus E(\mathcal{B})}\mathcal{H}(B)-\sum_{A\in\mathcal{A}\setminus E(\mathcal{A})}\mathcal{H}(A)-\sum_{A\in E(\mathcal{A})}\mathcal{H}(A)-\sum_{B\in\mathcal{B}\setminus E(\mathcal{B})}\mathcal{H}(B)-\sum_{B\in E(\mathcal{B})}\mathcal{H}(B)-|E|T(S_b-S_u)\right)}}$$

$$= \frac{e^{-\frac{1}{kT}\left(\sum_{(A,B)\in E}\mathcal{H}(AB)-\sum_{A\in E(\mathcal{A})}\mathcal{H}(A)-\sum_{B\in E(\mathcal{B})}\mathcal{H}(B)-|E|T(S_b-S_u)\right)}}{1+\sum_{E\in\mathcal{E}(\mathcal{AB})} e^{-\frac{1}{kT}\left(\sum_{(A,B)\in E}\mathcal{H}(AB)-\sum_{A\in E(\mathcal{A})}\mathcal{H}(A)-\sum_{B\in E(\mathcal{B})}\mathcal{H}(B)-|E|T(S_b-S_u)\right)}}$$

Note that in the second to last step above we split the summation over $\mathcal{A}$ into two summations, one over $\mathcal{A}\backslash E(\mathcal{A})$ and the other over $E(\mathcal{A})$, and similarly for $\mathcal{B}$. This allowed us to cancel some terms. We now take advantage of the fact that for every $(A,B)\in E$, there is one and only one pair of $A\in E(\mathcal{A}), B\in E(\mathcal{B})$, which allows us to collapse the sums further:

$$p^{(\mathrm{ce})}(E) = \frac{e^{-\frac{1}{kT}\left(\sum_{(A,B)\in E}(\mathcal{H}(AB)-(\mathcal{H}(A)+\mathcal{H}(B)))-|E|T(S_b-S_u)\right)}}{1+\sum_{E\in\mathcal{E}(\mathcal{AB})} e^{-\frac{1}{kT}\left(\sum_{(A,B)\in E}(\mathcal{H}(AB)-(\mathcal{H}(A)+\mathcal{H}(B)))-|E|T(S_b-S_u)\right)}}$$

This yields an expression in terms of the differences of domain-level Hamiltonians, which are identical to the terms in the previous section, and $S_b - S_u$, which is equal to ΔS. Hence the indeterminancy extends to the protein model and we are able to use the quantities derived in the domain model to calculate all protein-level probabilities.

Perturbation

All perturbation probabilities are defined in terms of the protein-level probabilities, which we have shown can be directly calculated. The only term for which this is not true is $p_{\mathcal{P}}^{(\mathrm{genomic})}\left((\mathcal{A},\mathcal{B})\rightarrow(\mathcal{A}^*,\mathcal{B}^*)\right)$, which is the probability that the $(\mathcal{A},\mathcal{B})\rightarrow(\mathcal{A}^*,\mathcal{B}^*)$ perturbation will be observed in $\mathcal{P}$. This quantity can be empirically estimated from unbiased sources of data, e.g. whole genome sequencing data that does not target specific regions of the genome. In computing these probabilities for tumors, we used whole genome TCGA data.

POSITION ENERGY MATRIX (PEM)

Protein binding affinity is conventionally represented using PSSMs, a mathematical construct that specifies a probability distribution over amino acids at multiple sequence positions. The basic intuition behind PSSMs is that proteins prefer certain amino acids at certain positions over others, and PSSMs represent that preference by a sequence of probability distributions. An important advantage of PSSMs is that they can be derived directly from experimental data, such as an oriented peptide array library or phage display, without the need for model construction. Despite their utility, PSSMs have a number of drawbacks. First, because the probabilities at each position are by definition normalized to add to 1, all positions contribute the same amount to the total affinity. This is in contradiction with what has been long known about the sequence specificity of PIDs, which tend to depend on making energetically favorable contacts at a handful of positions. Second, interactions that are actively disfavored, i.e. which are repulsive in nature, are not possible to capture using PSSMs. This makes them unable to represent negative selection. Finally, because of their relative nature (probabilities are normalized at each position), it is impossible to use PSSMs to compare the difference in affinity of two proteins to the same peptide, only the affinity of one protein to two or more peptides.

We introduce a new representation for the binding affinity of proteins which we term the Position Energy Matrix or PEM to address these shortcomings. This representation captures, for a single protein, its attraction and repulsion to every possible amino acid at every residue position in its ligand. In the context of our multiscale statistical mechanical framework, we will describe how PEMs are constructed for SH2 domains.

PEMs depend on the underlying energy model representing domain-peptide interactions. Since we are only interested in the contributions of the peptide against a fixed protein background, we derive a special domain-level energy model in which all the first-order functions, i.e. $\left\{u_i^{(1)}\right\}$ and $\left\{b_i^{(1)}\right\}$, are set to 0. This results in a second-order only energy model comprised of the $\left\{b_{i,j}^{(2)}\right\}$ functions. Using this energy model, we have that for an SH2 domain A the energetic contribution of an amino acid a at peptide position k, denoted by $e_k(A,a)$, is:

$$e_k(A,a) = \sum\nolimits_{i\in|A|} b_{i,k}^{(2)}(A_i,a)$$

In the above definition, the variable i indexes the residue positions of A, and as before A_i denotes the amino acid at position i in domain A. Hence the PEM sums all the energetic contributions from the SH2 domain toward a given position in the peptide. The position energy matrix for domain A, denoted by $\mathrm{PEM}(A)$, is then defined as:

$$\mathrm{PEM}(A) \equiv \{e_k(A,a)\}_{a\in\mathrm{AA}\setminus\{C\},k\in 1,\ldots,|\mathrm{pY}|}$$

I.e. the rows of PEM(A) index the amino acids and its columns index the residue positions of the peptide. Although each PEM is specific to a single domain, the fact that the underlying energy model is universal across all SH2 domains implies that the energy terms $\{e_k(A,a)\}$ can be compared across domains. Furthermore since the energies are not normalized, it is possible for one peptide position to contribute considerably more energy than another position.

PEMs admit a natural visual representation in which the energetic contribution of every residue at every position is reflected by the height and opacity of the residue, as shown in Figure 7. A horizontal line divides the representation so that residues that increase affinity lie above the line and those that decrease it lie below the line. Vertically the residues are ordered according to the magnitude of their contribution, with the largest contributor shown closest to the dividing line.