Computational design of PDZ domains: parameterization and performance of a simple model

David Mignon[1,3], Nicolas Panel[1,3], Xingyu Chen[1], Ernesto J. Fuentes[2] and Thomas Simonson[1,*]

[1]Laboratoire de Biochimie (UMR CNRS 7654), Ecole Polytechnique, Palaiseau, France
[2]Department of Biochemistry, Roy J. and Lucille A. Carver College of Medicine, and Holden Comprehensive Cancer Center, University of Iowa, Iowa City, Iowa 52242-1109, United States
[3]Joint first authors. *Corresponding author: thomas.simonson@polytechnique.fr

Short title: Computational design of PDZ domains

Abstract

PDZ domains help direct protein-protein interactions and have been extensively studied and engineered. Here, a protein design energy function was optimized for a small set of PDZ domains. It combines a molecular mechanics protein energy, Generalized Born solvent, and an empirical unfolded state model characterized by amino acid type-dependent chemical potentials. These were optimized using a maximum likelihood formalism and several model variants. Sequences designed with the best variants were almost all recognized by the Superfamily fold recognition tool. They had similarity scores relative to natural sequences comparable to similarities between natural PDZ domains and to sequences designed with the Rosetta software. Five Tiam1 designs proved stable in 200–600 ns molecular dynamics simulations. The model was then used to redesign the hydrophobic core of four of the PDZ domains, by gradually varying an energy term that alters the chemical potential of hydrophobic amino acid types. The tendency of each position to retain, lose, or gain a hydrophobic character represents a novel, structure-based hydrophobicity index, whose mean value differs by a factor of two between two of the proteins. In a second application, we redesigned four Tiam1 positions involved in peptide binding and specificity, three of which are part of the hydrophobic core. The calculations were done for the apo protein and two distinct protein-peptide complexes. The designed sequences are homologous to the wildtype sequence and an experimental quadruple mutant that has different affinities for the two peptides. The calculated affinity differences between designed protein variants reproduce experimental data qualitatively.

Author summary

Computational protein design is an emerging method that seeks to engineer new properties into proteins by mimicking the natural mechanisms of mutation and selection. Proteins that help organize networks of interactions in cells, like PDZ domains, are attractive targets. The space of possible sequences grows exponentially with protein size, making the problem tremendously complex. Thus, new models and parameterizations still need to be developed and explored. Here, we parameterize one such model, using statistical inference to choose optimal parameter values. We generate thousands of theoretical sequences using a powerful Monte Carlo exploration and we compare them to natural sequences, to determine the predictive power of the model. Results are encouraging, so that two applications are carried out. First, we do simulations that are increasingly biased towards hydrophobic residue types, which then invade the protein from the inside out; this gives a measure of the structure's susceptibility to, or tolerance of hydrophobic groups. Second, we study four positions in the Tiam1 PDZ domain that help establish specific binding of its partners, and we allow them to mutate. This provides another test of predictive power, and suggests mutations that might alter Tiam1 specificity, a step towards cellular network engineering.

1 Introduction

PDZ domains ("Postsynaptic density-95/Discs large/Zonula occluden-1") are small, globular protein domains that help establish protein-protein interaction networks in the cell [1–6]. They form specific interactions with other, target proteins, usually by recognizing a few amino acids at the target C-terminus. Due to their biological importance, PDZ domains and their ligands have been extensively studied and engineered, including many studies with computational methods. Peptide ligands have been designed that modulate the activity of PDZ domains involved in various pathologies [7–9]. Engineered PDZ domains and PDZ ligands have been used to elucidate principles of protein folding and evolution [10–13]. In addition, these small domains with their peptide ligands provide benchmarks to test the computational methods themselves [14–16].

One emerging method that has been applied to several PDZ domains is computational protein design (CPD) [17–22]. Starting from a 3D structural model, CPD explores a large space of amino acid sequences and conformations to identify protein variants that have certain predefined properties, such as stability or ligand binding. Conformational space is usually defined by a library of sidechain rotamers, which can be discrete or continuous, and by a finite set of backbone conformations or a specific repertoire of allowed backbone deformations. The energy function usually combines physical and empirical terms [23–25]. Both solvent and the unfolded protein state are described implicitly.

Here, we consider a simple but important class of CPD models; we optimize selected parameters of the energy function for a group of PDZ proteins; we test the quality of each model, and we use an optimized model for two applications. Our CPD models are implemented in the Proteus software [26–28]. They use an "MMGBSA" energy function, which combines a molecular mechanics protein energy with a Generalized Born + Surface Area implicit solvent treatment. The folded protein is represented by a single, fixed, backbone conformation and a discrete sidechain rotamer library. The unfolded state energy depends only on sequence composition, not an explicit structural model. The main adjustable model parameters are the protein dielectric constant ϵ_P, a small set of atomic surface energy coefficients σ_i, and a collection of amino acid chemical potentials, or "reference energies" E_t^r. Each surface coefficient measures the preference of a particular atom type to be solvent-exposed. Each reference energy represents the contribution of a single amino acid of type t to the unfolded state energy.

We optimize the reference energies E_t^r using a maximum likelihood formalism and a

set of eight PDZ test proteins. For two of the proteins, Tiam1 and Cask, we compare two sets of surface coefficients and two values of the dielectric constant. The resulting parameter sets are tested by generating designed sequences for all eight proteins and comparing them to natural sequences, as well as sequences generated with the Rosetta energy function and software [29]. The sequence design is done by performing long Monte Carlo simulations, where all protein positions except Gly and Pro are allowed to mutate freely, leading to thousands of designed protein variants. We also perform 100-600 nanosecond molecular dynamics simulations for a few the sequences designed with our model, to help assess their stability; five are stable over 200 ns and one shows stability over 600 ns similar to the wildtype.

We then apply the model to two problems, using optimized parameters. First, we do a series of Monte Carlo simulations of four of our PDZ domains where the chemical potential of the hydrophobic amino acid types is gradually increased, artificially biasing the protein composition. As we increase the bias, hydrophobic amino acids gradually invade the protein from the inside out, forming a hydrophobic core that is initially smaller, then becomes larger than the natural one. The propensity of each core position to become hydrophobic at a high or low level of bias can be seen as a structure-dependent hydrophobicity index, providing information on the designability of the protein core. The second application consists in designing four Tiam1 positions that are known to be involved in specific target recognition, and have been experimentally mutated so as to modify the preferred Tiam1 target [30], increasing its preference for the Caspr4 peptide, with respect to the syndecan-1 peptide (Sdc1). We mutate these positions through Monte Carlo simulations of either the apo-protein or the protein in complex with either peptide ligand. The simulations give encouraging agreement with experimental sequences and binding affinities, and suggest new variants that could have altered specificities.

2 The unfolded state model

2.1 Maximum likelihood reference energies

We use Monte Carlo to generate a Markov chain of states [31, 32], such that the states are populated according to a Boltzmann distribution. One possible elementary move is a "mutation": we modify the sidechain type $t \to t'$ at a chosen position i in the folded protein, assigning a particular rotamer r' to the new sidechain. At the same time, we

perform the reverse mutation in the unfolded protein, $t' \to t$. For a particular sequence S, the unfolded state energy has the form:

$$E^u = \sum_{i \in S} E^r(t_i). \tag{1}$$

The sum is over all amino acids; t_i represents the sidechain type at position i. The type-dependent quantities $E^r(t) \equiv E^r_t$ are referred to as "reference energies"; they can be thought of as effective chemical potentials of each amino acid type. The energy change due to a mutation has the form:

$$\Delta E = \Delta E^f - \Delta E^u = \left(E^f(...t'_i, r'_i...) - E^f(...t_i, r_i...)\right) - \left(E^r(t'_i) - E^r(t_i)\right) \tag{2}$$

where ΔE^f and ΔE^u are the energy changes in the folded and unfolded state, respectively. The reference energies are essential parameters in the simulation model. Our goal here is to choose them empirically so that the simulation produces amino acid frequencies that match a set of target values, for example experimental values in the Pfam database. Specifically, we will choose them so as to maximize the probability, or likelihood of the target sequences.

Let S be a particular sequence. Its Boltzmann probability is

$$p(S) = \frac{1}{Z} \exp(-\beta \Delta G_S), \tag{3}$$

where $\Delta G_S = G^f_S - E^u_S$ is the folding free energy of S, G^f_S is the free energy of the folded form, $\beta = 1/kT$ is the inverse temperature and Z is a normalizing constant (the partition function). We then have

$$kT \ln p(S) = \sum_{i \in S} E^r(t_i) - G^f_S - kT \ln Z = \sum_{t \in \mathrm{aa}} n_S(t) E^r_t - G^f_S - kT \ln Z, \tag{4}$$

where the sum on the right is over the amino acid types and $n_S(t)$ is the number of amino acids of type t within the sequence S.

We now consider a set $\mathcal{S}$ of N target sequences S; we denote $\mathcal{L}$ the probability of the entire set, which depends on the model parameters E^r_t; we refer to $\mathcal{L}$ as their likelihood [33]. We have

$$kT \ln \mathcal{L} = \sum_S \sum_{t \in \mathrm{aa}} n_S(t) E^r_t - \sum_S G^f_S - NkT \ln Z = \sum_{t \in \mathrm{aa}} N(t) E^r_t - \sum_S G^f_S - NkT \ln Z, \tag{5}$$

where $N(t)$ is the number of amino acids of type t in the whole dataset $\mathcal{S}$. The normalization factor or partition function Z is a sum over all possible sequences R:

$$Z = \sum_R \exp(-\beta \Delta G_R) = \sum_R \exp(-\beta \Delta G_R^f) \Pi_{t \in aa} \exp(\beta n_R(t) E_t^r) \tag{6}$$

In view of maximizing $\mathcal{L}$, we consider the derivative of Z with respect to one of the E_t^r:

$$\frac{\partial Z}{\partial E_t^r} = \sum_R \beta n_R(t) \exp(-\beta \Delta G_R^f) \prod_{s \in aa} \exp(\beta n_R(s) E_s^r) \tag{7}$$

We then have

$$\frac{kT}{Z} \frac{\partial Z}{\partial E_t^r} = \frac{\sum_R n_R(t) \exp(-\beta \Delta G_R)}{\sum_R \exp(-\beta \Delta G_R)} = \langle n(t) \rangle. \tag{8}$$

The quantity on the right is the Boltzmann average of the number $n(t)$ of amino acids t over all possible sequences. In practice, this is the average population of t we would obtain in a long MC simulation. We note that, as usual in statistical mechanics [34], the derivative of $\ln Z$ with respect to one quantity (E_t^r) is equal to the ensemble average of the conjugate quantity $(\beta n_S(t))$.

A necessary condition to maximize $\ln \mathcal{L}$ is that its derivatives with respect to the E_t^r should all be zero. We see that

$$\frac{1}{N} \frac{\partial}{\partial E_t^r} \ln \mathcal{L} = \frac{1}{N} \sum_S n_S(t) - \langle n(t) \rangle = \frac{N(t)}{N} - \langle n(t) \rangle \tag{9}$$

so that

$$\mathcal{L} \text{ maximum} \Longrightarrow \frac{N(t)}{N} = \langle n(t) \rangle, \ \ \forall t \in \text{aa} \tag{10}$$

Thus, to maximize $\mathcal{L}$, we should choose $\{E_t^r\}$ such that a long simulation gives the same amino acid frequencies as the target database.

2.2 Searching for the maximum likelihood

To approach the maximum likelihood $\{E_t^r\}$ values, starting from a current guess $\{E_t^r(n)\}$, we will use two methods. With the first method, we step along the gradient of $\ln \mathcal{L}$, using the update rule [33]:

$$E_t^r(n+1) = E_t^r(n) + \alpha \frac{\partial}{\partial E_t^r} \ln \mathcal{L} = E_t^r(n) + \delta E \left(n_t^{\text{exp}} - \langle n(t) \rangle_n \right) \tag{11}$$

Here, α is a constant; $n_t^{\text{exp}} = N(t)/N$ is the mean population of amino acid type t in the target database; $\langle \rangle_n$ indicates an average over a simulation done using the current

reference energies $\{E_t^r(n)\}$, and δE is an empirical constant with the dimension of an energy, referred to as the update amplitude. This update procedure is repeated until convergence. We refer to this method as the linear update method.

The second method, used previously [26, 27], employs a logarithmic update rule:

$$E_t^r(n+1) = E_t^r(n) + kT \ln \frac{\langle n(t) \rangle_n}{n_t^{\exp}} \tag{12}$$

where kT is a thermal energy, set empirically to 0.5 kcal/mol (1 cal = 4.184 J). We refer to this as the logarithmic update method. Both the linear and logarithmic update methods converge to the same optimum, specified by (Eq. 10).

In the later iterations, some E_t^r values tended to converge slowly, with an oscillatory behavior. Therefore, we sometimes used a modified update rule, where the $E_t^r(n+1) - E_t^r(n)$ value computed with the linear or logarithmic method for iteration n was mixed with the value computed at the previous iteration, with the $(n-1)$ value having a weight of 1/3 and the current value a weight of 2/3. At each iteration, we typically ran 500 million steps (per replica) of Replica Exchange Monte Carlo.

3 Computational methods

3.1 Effective energy function for the folded state

The energy matrix was computed with the following effective energy function for the folded state:

$$E = E_{\text{bonds}} + E_{\text{angles}} + E_{\text{dihe}} + E_{\text{impr}} + E_{\text{vdw}} + E_{\text{Coul}} + E_{\text{solv}} \tag{13}$$

The first six terms in (13) represent the protein internal energy. They were taken from the Amber ff99SB empirical energy function [35], slightly modified for CPD (see below). The last term on the right, E_{solv}, represents the contribution of solvent. We used a "Generalized Born + Surface Area", or GBSA implicit solvent model [36]:

$$E_{\text{solv}} = E_{\text{GB}} + E_{\text{surf}} = \frac{1}{2}\left(\frac{1}{\epsilon_W} - \frac{1}{\epsilon_P}\right) \sum_{ij} q_i q_j \left(r_{ij}^2 + b_i b_j \exp[-r_{ij}^2/4b_i b_j]\right)^{-1/2} + \sum_i \sigma_i A_i \tag{14}$$

Here, ϵ_W, ϵ_P are the solvent and protein dielectric constants; r_{ij} is the distance between atoms i, j and b_i is the "solvation radius" of atom i [36, 37]. A_i is the exposed solvent

accessible surface area of atom i; σ_i is a parameter that reflects each atom's preference to be exposed or hidden from solvent. The solute atoms were divided into 4 groups with specific σ_i values (see below): nonpolar, aromatic, polar, and ionic. Hydrogen atoms were assigned a surface coefficient of 0. Surface areas were computed by the Lee and Richards algorithm [38], implemented in the XPLOR program [39], using a 1.5 Å probe radius. Most of the MC simulations used a protein dielectric of $\epsilon_P = 4$ or 8 (see Results).

In the GB energy term, the atomic solvation radius b_i approximates the distance from i to the protein surface and is a function of the coordinates of all the protein atoms. The particular b_i form corresponds to a GB variant we call GB/HCT, after its original authors [36], with model parameters optimized for use with the Amber force field [37]. Since b_i depends on the coordinates of all the solute atoms [36], an additional approximation is needed to make the GB energy term pairwise additive and define the energy matrix. We use a "Native Environment Approximation", or NEA, where the solvation radius b_i of each particular group (backbone, sidechain or ligand) is computed ahead of time, with the rest of the system having its native sequence and conformation [27, 40].

The surface energy contribution $E_{\rm surf}$ is not pairwise additive either, because in a protein structure, surface area buried by one sidechain may also be buried by another. To make this energy pairwise, Street et al proposed a simple procedure [41]. The buried surface of a sidechain is computed by summing over the neighboring sidechain and backbone groups. For each neighboring group, the contact area with the sidechain of interest is computed, independently of other surrounding groups. The contact areas are then summed. To avoid overcounting of buried surface area, a scaling factor is applied to the contact areas involving buried sidechains. Previous work showed that a scaling factor of 0.65 works well [37, 40].

The Amber force field ff99SB is slightly modified for CPD, with the original backbone charges replaced by a unified set, obtained by averaging over all amino acid types and adjusting slightly to make the backbone portion of each amino acid neutral [42].

3.2 Reference energies in the unfolded state

In the unfolded state, the energy depends on the sequence composition through a set of reference energies E_t^r (Eq. 1). The values are assigned based on amino acid types t, taking into account also the position of each amino acid in the folded structure, through its buried or solvent-exposed character. Thus, for a given type (Ala, say), there are two distinct E_t^r values: a buried and an exposed value. This is done even though the reference energies are

used to represent the unfolded, not the folded state. There are three rationales for this. First, we assume residual structure is present in the unfolded state, so that amino acids partly retain their buried/exposed character. Second, we hypothesize that the unfolded state model compensates in a systematic way for errors in the folded state energy function, so that the folded structure matters. Third, this strategy makes the model less sensitive to variations in the length of surface loops, and to the proportion of surface vs. buried residues, which can vary widely among homologs (see below); as a result, the model should be more transferable within a protein family.

Distinguishing buried/exposed positions doubles the number of adjustable E_t^r parameters. Conversely, to reduce the number of adjustable parameters, we group amino acids into homologous classes (given in Results). Within each class c, and for each type of position (buried or exposed), the reference energies have the form

$$E_t^r = E_c^r + \delta E_t^r \tag{15}$$

Here, E_c^r is an adjustable parameter while δE_t^r is a constant, computed as the molecular mechanics energy difference between amino acid types within the class c, assuming an unfolded conformation where each amino acid interacts only with itself and with solvent. During likelihood maximization, E_c^r is optimized while δE_t^r is held fixed. To optimize the E_c^r values, we apply the linear or logarithmic method above; the target frequencies correspond to the experimental frequencies of the amino acid classes, $n_c^{\exp}$, rather than of the individual types ($n_t^{\exp}$, above).

3.3 Experimental sequences

We considered a set of eight PDZ domains, whose PDB codes are listed in Table 1. The top four belong to PDZ class I and the bottom four to class II. To define the target amino acid frequencies for likelihood maximization, we collected homologous sequences for each of the eight. We started by a Blast search of the Uniprot database with the PDB sequence as the query, using the Blosum62 matrix, and retained homologs with a sequence identity, relative to the query, above a certain threshold, around 60–80% depending on the test protein. Sequences that were over 95% or 85% identical to the query were removed, and homologs with mutual identities above 95% were pruned, keeping just one of the redundant variants. This led to about 40–120 homologs per test protein; see details in Table 1. For each set, amino acid frequencies were computed, and averaged over positions. The averages were computed separately for buried and exposed positions. Buried positions

were defined to have a solvent-accessible surface area below 20% of that obtained for the amino acid alone, which led to similar numbers of buried/exposed positions. The eight sets of mean frequencies were themselves averaged, giving the overall target amino acid frequencies (see below). Distinct target frequencies are thus obtained for buried and exposed positions.

Table 1: Test proteins and their homologs

protein name[a]	PDB code	residue numbers	# active positions[c]	number of homologs	[d]E-value threshold	[e]identity % range
NHERF(1)	1G9O	9-99	76	62	1e-32	67-95
syntenin(2)	1R6J	192-273	72	85	1e-43	85-95
DLG2(2)	2BYG	186-282	82	43	1e-41	78-95
PSD95(3)	3K82	305-402	80	50	1e-46	81-95
INAD(1)	[b]1IHJ	12-105	82	42	1e-10	38-95
GRIP(6)	[b]1N7E	667-761	79	48	1e-45	84-95
Cask	[b]1KWA	487-568	74	126	7e-28	60-85
Tiam1	[b]4GVD	837-930	84	50	2e-23	60-85

[a]In parentheses: number of the PDZ domain within the protein. [b]Holo structures. [c]The number of non-Gly, non-Pro positions, which can mutate during the design simulations. [d]E-values are for the Blast search for experimental homologs. [e]Identity % is between the homologs and the query protein.

3.4 Structural models

As test systems, we used a mixture of PDZ domains from the two main specificity classes, I and II. Model parameterization and testing were mostly done for the apo proteins. Consistent with this, we used PDB structures of the apo state for four of the test proteins, which all belong to the PDZ class I (NHERF, syntenin, DLG2, and PSD95). For the class II PDZ domain of Tiam1, we wanted to use the parameterized CPD model to design the protein:peptide complex, and so we decided to use a holo structure, then modelled the apo state by removing the peptide. For this PDZ domain, the backbone rms deviation between the apo and holo X-ray structures is just 0.5 Å (and similar apo/holo deviations are found for the class I domains above); therefore, we expect the design model to be

transferable between apo/holo Tiam1 states. For the class II domains GRIP and Cask, no apo X-ray structure was available at the beginning of this work, so holo structures were also used. For the class II INAD PDZ domain, the available apo structures have an unusual orientation of the α_1 helix [43] and so a holo structure was used.

For the design calculations, structures were prepared and energy matrices computed using procedures described previously [15, 44]. For Tiam1, two missing segments (residues 851-854 and 868-869) were built using the Modeller program [45]. For most of the design calculations, we removed the peptide ligand, when present in the PDB structure, before computing the energy matrix. In the energy matrix calculations, for each residue pair, interaction energies were computed after 15 steps of energy minimization, with the backbone fixed and only the interactions of the pair with each other and the backbone included. This short minimization of pairs alleviates the discrete rotamer approximation. Sidechain rotamers were described by a slightly expanded version of the library of Tuffery et al [46], which has a total of 254 rotamers (sum over all amino acid types). The expansion consists in allowing additional hydrogen orientations for OH and SH groups [40]. This rotamer library was chosen for its simplicity and because it gives very good performance in sidechain placement tests, comparable to the specialized Scwrl4 program (which uses a much larger library) [47, 48].

3.5 Monte Carlo simulations

With Proteus, sequence design is done by running long Monte Carlo (MC) simulations where selected amino acid positions can mutate freely. Here, the choice of mutating positions depends on the calculation. To optimize the reference energies, we do simulations where about half of the positions can mutate at a time. To test the optimized models, we mostly do simulations where all positions except Gly and Pro are free to mutate. To produce designed sequences that will be tested through molecular dynamics, we do MC simulations where Gly, Pro, and 11 positions closely involved in peptide binding are held fixed, while all other positions are allowed to mutate. Finally, in the Tiam1 quadruple mutant application, only four positions in the protein can mutate (and the peptide ligand is present). In all cases (with two exceptions), mutations occur randomly, subject only to the MMGBSA energy function that drives the simulation. In just two tests, we use an additional, "experimental" energy term to explicitly bias the simulation to stay close to the natural, Pfam sequences; see below.

The Monte Carlo simulations use one- and two-position moves, where either rotamers,

types, or both are changed. For two-position moves, the second position is selected among those that have a significant (unsigned) interaction energy with the first one, meaning that there is at least one rotamer conformation where their interaction is 10 kcal/mol or more. In addition, to enhance sampling, we mostly perform Replica Exchange Monte Carlo (REMC), where several MC simulations ("replicas" or "walkers") are run in parallel, at different temperatures; periodic swaps are attempted between the conformations of two walkers i, j (adjacent in temperature). The swap is accepted with probability

$$acc(\text{swap}_{ij}) = \text{Min}\left[1, e^{(\beta_i - \beta_j)(\Delta E_i - \Delta E_j)}\right] \tag{16}$$

where β_i, β_j are the inverse temperatures of the two walkers and ΔE_i, ΔE_j are the changes in their folding energies due to the conformation change [49, 50]. We use eight walkers, with thermal energies kT_i that range from 0.125 to 3 kcal/mol, and are spaced in a geometric progression: T_{i+1}/T_i = constant [49]. Simulations are done with the proteus program (which is part of the Proteus package) [27]; REMC uses an efficient, shared-memory, OpenMP parallelization [51].

For the Tiam1 and Cask proteins, one simulation (each) was also done that included an "experimental", biasing energy term, which penalizes sequences that have a low similarity to a reference, experimental set. The bias energy had the form

$$\delta E_{\text{bias}} = c \sum_i \left(S(t_i) - S_i^{\text{rand}} \right), \tag{17}$$

where the sum is over the amino acid positions i, t_i is the sidechain type at position i, $S(t_i)$ is the (dimensionless) Blosum40 similarity score versus the corresponding position in the Pfam RP55 sequence alignment; S_i^{rand} is the mean score (versus the same Pfam column) for a random type (where all types are equiprobable), and $c = 0.5$ kcal/mol.

3.6 Rosetta sequence generation

Monte Carlo simulations were also done using the Rosetta program and energy function [29]. The simulations were done using version 2015.38.58158 of Rosetta (freely available online), using the command

```
fixbb -s Tiam1.pdb -resfile Tiam1.res -nstruct 10000 -ex1 -ex2 -linmem_ig 10
```

where the last option corresponds to on-the-fly energy calculation, ex1 and ex2 activate an enhanced rotamer search for buried sidechains, and default parameters are used otherwise.

Simulations were run for each protein until 10000 unique low energy sequences were identified, corresponding to run times of about 5 minutes per sequence on a single core of a recent Intel processor, for a total of 10 hours (per protein) using 80 cores. This is comparable to the cost of the Proteus calculations (energy matrix plus Monte Carlo).

3.7 Sequence characterization

Designed sequences were compared to the Pfam alignment for the PDZ family, using the Blosum40 scoring matrix and a gap penalty of -6. Each Pfam sequence was also compared to the Pfam alignment. For these Pfam/Pfam comparisons, if a test protein T was part of the Pfam alignment, the T/T self comparison was left out, to be more consistent with the designed/Pfam comparisons. The Pfam alignment was the "RP55" alignment, with 12255 sequences. Similarities were computed for 14 core residues and 16 surface residues, defined by their near-complete burial or exposure (listed in Results) and for the entire protein.

Designed sequences were submitted to the Superfamily library of Hidden Markov Models [52, 53], which attempts to classify sequences according to the SCOP classification [54]. Classification was based on SCOP version 1.75 and version 3.5 of the Superfamily tools. Superfamily executes the hmmscan program, which implements a Hidden Markov model for each SCOP family and superfamily; here hmmscan was executed with an E-value threshold of 10^{-10}, using a total of 15438 models to represent the SCOP database.

To compare the diversity in the designed sequences with the diversity in natural sequences, we used a standard, position-dependent sequence entropy [55], computed as follows:

$$S_i = -\sum_{j=1}^{6} f_j(i) \ln f_j(i) \tag{18}$$

where $f_j(i)$ is the frequency of residue type j at position i, either in the designed sequences or in the natural sequences (organized into a multiple alignment). Instead of the usual, 20 amino acid types, we employ six residue classes, corresponding to the following groups: {LVIMC}, {FYW}, {G}, {ASTP}, {EDNQ}, and {KRH}. This classification was obtained by a cluster analysis of the BLOSUM62 matrix [56], and also by analyzing residue-residue contact energies in proteins [57]. To get a sense of how many amino acid types appear at a specific position i, we report the residue entropy in its exponentiated form, $\exp(S_i)$ (which ranges from 1 to 6), averaged over the protein chain (i.e., S_i is exponentiated first, then averaged over positions).

3.8 Molecular dynamics simulations

For wildtype Tiam1, a qudruple mutant, and ten sequences designed with Proteus, we ran MD simulations with explicit solvent. Starting structures were taken from the MC trajectory or the crystal structure (wildtype protein and quadruple mutant: PDB code 4NXQ) and slightly minimized with harmonic restraints to maintain the backbone geometry. The protein was immersed in a large box of water; waters overlapping protein were eliminated, and the solvated system was truncated to the shape of a truncated octahedral box using the Charmm graphical interface or GUI [58]; the minimum distance between protein atoms and the box was 15 Å; final models included about 11000 water molecules. A few sodium or chloride ions were included to ensure overall electroneutrality. Protonation states of histidines were assigned to be neutral, based on visual inspection. MD was done at room temperature and pressure, using a Nose-Hoover thermostat and barostat. Long-range electrostatic interactions were treated with a Particle Mesh Ewald approach [59]. The Amber ff99SB forcefield was used for the protein; the TIP3P model [60] was used for water. Simulations were run for 100–600 nanoseconds, depending on the sequence, using the Charmm and NAMD programs [61, 62].

4 Results

4.1 Experimental structures and sequences

3D structures of the test proteins are shown in Fig. 1A; for clarity, only four of the eight proteins are shown. 14 core residues, identified visually (and highlighted by their C_β atoms, shown as spheres) superimpose well between structures; loops and chain termini display large deviations, and the Tiam1 α_2 helix is rotated slightly outwards compared to the other three structures. Fig. 1B illustrates the similarity between all eight domains, measured by the rms deviation between structurally-aligned C_α atoms. Structure pairs with rms deviations of 1 Å or less and 60 aligned residues or more are linked; 2BYG (DLG2) forms the center of the group, with five links; Cask is not quite linked to 1IHJ (rms deviation of 1.3 Å over 63 residues) and Tiam1 is isolated. Sequence identities are also shown, including those between Tiam1 and its closest homologs. Overall, six proteins form a tight group, with deviations of 1 Å or less, while Cask and Tiam1 are further away. The Tiam1/Cask sequence identity is 33%; their structural deviation is 1.7 Å based on a 42-residue alignment.

⋆⋆⋆ insert Figure 1 near here ⋆⋆⋆

Sequence conservation within our eight proteins and a subset of the Pfam seed alignment (half) is shown in Fig. 2. The 14 positions we use to define the hydrophobic core are highly, though not totally conserved within the Pfam seed alignment. Arg, Lys and Gln appear at some of the positions, since in a small PDZ protein, the long hydrophobic portion of these sidechains can be buried in the core while still allowing the polar tip of the sidechain to be exposed to solvent. A few Asp and Glu residues also appear, in places where the sequence alignment may not reflect closely the 3D sidechain superposition.

⋆⋆⋆ insert Figure 2 near here ⋆⋆⋆

4.2 Optimizing the unfolded state model

We optimized the reference energies E_t^r for six of the eight proteins, with the other two (Tiam1, Cask) left for cross validation. The protein dielectric constant was $\epsilon_P = 8$, and a first set of atomic surface coefficients was used. We refer to the resulting model either as model A or as the (ϵ_P=8, S1, n=6) model (S1 for "set 1"; n=6 represents the number of proteins used to define the target amino acid frequencies). We repeated the optimization for Tiam1 and Cask alone, to see what improvement is obtained, if any, when a smaller set of proteins is specifically optimized. The resulting model is called A' or (ϵ_P=8, S1, n=2). Finally, we repeated the E_t^r optimization using a second set of surface coefficients and an ϵ_P of either 8 or 4; these calculations were done for Tiam1 and Cask alone, due to the cost of repeated optimizations with multiple proteins. The corresponding models are called B (ϵ_P=8, S2, n=2) and B' (ϵ_P=4, S2, n=2). Model names are listed in Table 2. The E_t^r optimizations all converged to within 0.05 kcal/mol after about 20 iterations for most amino acid types, and within 0.1 kcal/mol for the others (the weakly-populated types), using either the linear or the logarithmic method (Eq. 11 or 12). Table 3 indicates the final reference energies for models A and B, which give the best performance (see below), and for model A' (closely related to A). The E_t^r values are compared to, and agree qualitatively with the energies computed from an extended peptide structure, which provides a less empirical model of the unfolded state. Table 4 compares the amino acid frequencies from experiment and the simulations. The theoretical populations of the different amino acid classes agree well with experiment, with rms deviations of about 1% for models A, A', and B, for both exposed and buried classes. The agreement for the amino acid types is less good, with rms deviations of 2.0%/2.5% for

model A (buried/exposed positions, respectively) and 3.9%/2.4% for model B. The intra-class frequency distributions depend explicitly on the energy offsets δE_t^r defined within each class, which are computed with molecular mechanics (see Methods, Eq. 15). Notice also that since model A uses three times more target proteins than model B (n=6 vs. n=2), the target frequencies are presumably defined about $\sqrt{3} \approx 1.7$ times less precisely for model B, so a larger deviation is acceptable in principle.

Table 2: Model variants

model name	model symbol	dielectric constant	σ_i set	# target proteins	σ_i values (cal/mol/Å^2) [b](unpol,arom,polar,ionic)
A	(ϵ_P=8,S1,n=6[a])	8	set 1	6	-5, -12, -8, -9
A'	(ϵ_P=8,S1,n=2)	8	set 1	2	-5, -12, -8, -9
B	(ϵ_P=8,S2,n=2)	8	set 2	2	-5,-40,-80,-100
B'	(ϵ_P=4,S2,n=2)	4	set 2	2	-5,-40,-80,-100

[a]Number of proteins whose composition forms the target set; [b]atom surface energy coefficients.

Table 3: Unfolded state reference energies E_t^r (kcal/mol)

		Model A'		Model A		Model B	
Residues	Peptide[a]	Buried	Exposed	Buried	Exposed	Buried	Exposed
ALA	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CYS	-1.04	-1.04	-1.04	-1.04	-1.04	-0.85	-0.85
THR	-3.82	-3.82	-3.82	-3.82	-3.82	-5.44	-5.44
SER	-2.85	-2.71	-2.62	-3.73	-2.80	-3.71	-4.74
ASP	-9.17	-10.32	-10.24	-9.19	-9.80	-11.90	-15.88
GLU	-7.88	-9.03	-8.95	-7.90	-8.51	-11.97	-15.95
ASN	-5.64	-6.26	-5.69	-5.94	-6.00	-7.82	-10.22
GLN	-4.42	-5.04	-4.46	-4.72	-4.78	-7.07	-9.47
[b]HIS^+	15.72	14.82	15.18	14.53	14.96	12.53	9.73
[b]HIS_ϵ	12.62	11.72	12.08	11.43	11.85	10.49	7.69
[b]HIS_δ	13.16	12.25	12.62	11.96	12.39	10.86	8.06
ARG	-25.30	-25.97	-24.76	-28.29	-28.90	-32.00	-35.18
LYS	-4.21	-3.16	-2.96	-4.56	-4.47	-6.76	-10.17
ILE	1.63	3.12	2.05	4.72	2.11	4.63	3.63
VAL	-2.25	-0.76	-1.83	0.83	-1.77	0.26	-0.74
LEU	-1.92	-0.42	-1.49	1.17	-1.44	-0.12	-1.12
MET	-2.44	-2.85	-3.03	-2.78	-3.54	-2.05	-2.40
PHE	-1.42	-1.71	-3.25	-0.37	-2.55	-0.23	-4.17
TRP	-2.66	-2.95	-4.49	-1.61	-3.79	-2.21	-6.15
TYR	-4.56	-4.67	-5.58	-4.20	-6.10	-5.80	-9.82

[a]Energies within an extended peptide structure (averaged over positions). [b]His protonation states.

Table 4: Amino acid composition (%) of experimental and designed PDZ proteins

type	Design, model A				Experiment, n=6 set				Experiment, n=2 set				Design, model B			
	Buried		Exposed		Buried		Exposed		Buried		Exposed		Buried		Exposed	
A	11.1	17.0	4.4	12.0	10.9	16.9	4.6	13.4	5.9	11.2	4.6	13.4	4.1	12.7	7.2	13.6
C	0.0	[0.1]	0.3	[-1.4]	1.3		0.5		1.5		1.2		8.6	[1.5]	5.8	[0.2]
T	5.9		7.3		4.7		8.3		3.8		7.6		0.0		0.6	
S	4.3	4.3 [-1.0]	8.7	8.7 [1.1]	5.3	5.3	7.6	7.6	4.7	4.7	10.2	10.2	4.9	4.9 [0.2]	10.7	10.7 [0.5]
D	4.5	6.7	5.6	16.7	4.3	6.8	6.0	17.9	3.5	9.6	6.2	16.7	7.4	9.4	8.0	16.1
E	2.2	[-0.1]	11.1	[-1.2]	2.5		11.9		6.1		10.5		2.0	[-0.2]	8.1	[-0.6]
N	2.5	4.7	7.5	14.0	2.6	4.7	6.7	12.2	1.9	2.7	7.4	16.1	1.8	2.8	8.6	17.1
Q	2.2	[0.0]	6.5	[1.8]	2.1		5.5		0.8		8.7		1.0	[0.1]	8.5	[1.0]
H^+	1.0	1.1	5.2	5.6	1.2	1.2	5.0	5.0	0.7	0.7	4.7	4.7	0.1	0.9	1.8	4.5
H_ϵ	0.1	[-0.1]	0.4	[0.6]	0.0		0.0		0.0		0.0		0.6	[0.2]	2.2	[-0.2]
H_δ	0.0		0.0		0.0		0.0		0.0		0.0		0.2		0.5	
I	16.9	52.1	4.1	14.0	16.0	50.7	4.2	14.0	15.7	49.6	4.1	14.4	25.1	46.7	8.4	15.3
V	16.7	[1.4]	5.6	[0.0]	16.5		5.4		13.5		5.5		12.8	[-2.9]	3.3	[0.9]
L	18.5		4.3		18.2		4.4		20.4		4.8		8.8		3.6	
M	1.6	1.6 [0.7]	1.4	1.4 [-0.1]	0.9	0.9	1.5	1.5	5.0	5.0	1.4	1.4	5.9	5.9 [0.9]	1.4	1.4 [0.0]
K	1.5	1.5 [-1.0]	13.0	13.0 [2.1]	2.5	2.5	10.9	10.9	6.5	6.5	10.1	10.1	5.5	5.5 [-1.0]	10.8	10.8 [0.7]
R	2.5	2.5 [-0.3]	6.1	6.1 [-2.6]	2.8	2.8	8.7	8.7	1.8	1.8	9.5	9.5	2.2	2.2 [0.4]	9.1	9.1 [-0.4]
F	4.5	4.6	2.1	2.1	4.1	4.1	2.4	2.4	5.0	5.0	0.4	0.4	3.2	5.5	0.3	0.5
W	0.1	[0.5]	0.0	[-0.3]	0.0		0.0		0.0		0.0		2.3	[0.5]	0.2	[0.1]
Y	2.2	2.2 [-0.4]	0.4	0.4 [-0.8]	2.6	2.6	1.2	1.2	2.9	2.9	0.9	0.9	3.4	3.4 [0.5]	0.9	0.9 [0.0]
G	0.0	0.0	0.0	0.0	0.8	0.9	3.1	4.9	0.0	0.3	1.7	2.1	0.0	0.0	0.0	0.0
P	0.0	[-0.9]	0.0	[-4.9]	0.1		1.8		0.3		0.4		0.0	[-0.3]	0.0	[-2.1]
	type	class	type	class	type	class	type	class	type	class	type	class	type	class	type	class

Compositions are given for buried/exposed positions, for individual amino acid types (left) and for classes (right); values in brackets (right) are the deviations between design and experiment per class. The n=2 experimental target set includes the Tiam1 and Cask homologs; the n=6 set includes the homologs of the other 6 test proteins. Design results are shown for models A, B.

4.3 Assessing designed sequence quality

Family recognition tests Proteus simulations used Replica Exchange Monte Carlo with eight replicas at temperatures between 0.236 and 3 kcal/mol; 750 million steps (per replica) were run. All positions (except Gly and Pro) were allowed to mutate freely into all types (except Gly and Pro). The simulations were done with the MMGBSA energy function, without any bias towards natural sequences or any limit on the number of mutations. The 10000 lowest energies among those sampled by any of the the MC replicas were retained for analysis, along with the 10000 Rosetta sequences. These sequences were submitted to the Superfamily fold recognition tool [53, 63]. Results are given in Table 5. For six of eight proteins, all 10000 Rosetta sequences are assigned by Superfamily to the correct SCOP superfamily and family, with E-values between 0.5 10^{-3} and 4 10^{-3} for the family assignments. For Cask and Tiam1, the family success rates are 90/98%, with slightly higher E-values. With model A = (ϵ=8,S1,n=6), the Proteus results are also good, with 6 out of 8 proteins giving 100% of correct family assignments, with E-values between 2 10^{-3} and 15 10^{-3} for the family assignments. For the syntenin (1R6J) and Tiam1 PDZ domains, only a few of the top sequences were assigned to the correct family (13% and 4 %, respectively), and the Superfamily tool only recognized about half of the sequence length. In contrast, for the other six proteins, the Superfamily match lengths were similar to those seen with the Rosetta sequences. Notice that neither Tiam1 nor Cask were part of the reference energy optimization set. Specifically optimizing the reference energies for Tiam1 and Cask (model A') did not improve the Superfamily performance, with just 0.1% of correct family assignments for Tiam1 and 63% for Cask (vs. 4% and 100% with model A). Apparently, target frequencies averaged over Tiam1 and Cask (and their close homologs) alone do not improve the model, compared to more generic PDZ target frequencies. This may be due to the rather low similarity between Tiam1 and Cask, which means that their mean sequence is not very close to either one.

Changing surface coefficients (set 2, model B) gave significantly improved Proteus results for Tiam1 and Cask, even though these values were optimized earlier using a *different* solvent model (Coulombic instead of Generalized Born electrostatics) [64]. Using these coefficients and E_t^r values optimized for Tiam1 and Cask (model B or ϵ=8,S2,n=2), we obtained a high percentage of sequences assigned to the correct family: 91% for Tiam1 and 100% for Cask, slightly better than Rosetta. Changing the protein dielectric constant to ϵ_P=4 (model B') gave results for Tiam1 that were poorer than model B but still much better than model A.

Table 5: Fold recognition of designed sequences by Superfamily

Protein	Design model	[a]Match/seq length	[b]Superfamily E-value	[c]Superfamily success #	[b]Family E-value	[c]Family success #
1G9O	A	78/91	2.5e-3	10000	3.0e-3	10000
1R6J	A	41/82	1.5	1350	2.6e-2	1350
2BYG	A	77/97	1.0e-2	10000	2.3e-3	10000
3K82	A	79/97	5.8e-10	10000	3.6e-3	10000
1IHJ	A	86/94	5.6e-7	10000	2.3e-3	10000
1N7E	A	81/95	1.1e-6	10000	2.4e-3	10000
Tiam1	A	43/94	1.3	442	4.0e-2	374
Cask	A	72/83	2.3e-4	10000	1.5e-2	10000
Tiam1	B'	53/94	1.0e-4	10000	7.0e-2	5259
Cask	B'	76/83	5.1e-7	10000	1.6e-2	10000
Tiam1	B	64/94	1.2e-4	9920	5.2e-2	9058
Cask	B	71/83	3.2e-7	10000	8.2e-3	10000
1G9O	Rosetta	79/91	1.3e-13	10000	2.2e-3	10000
1R6J	Rosetta	76/82	7.3e-13	10000	1.8e-3	10000
2BYG	Rosetta	86/97	1.3e-9	10000	9.6e-4	10000
3K82	Rosetta	90/97	3.7e-23	10000	5.2e-4	10000
1IHJ	Rosetta	85/94	7.4e-14	10000	3.7e-3	10000
1N7E	Rosetta	84/95	2.2e-10	10000	1.2e-3	10000
Tiam1	Rosetta	65/94	4.4e-4	9035	2.8e-2	9030
Cask	Rosetta	68/83	2.8e-5	9832	7.5e-3	9832

[a]The average match length for sequences recognized by Superfamily and the total sequence length. [b]Average E-values for Superfamily assignments to the correct SCOP superfamily/family. [c]The number of designed sequences (out of 10000 tested) assigned to the correct SCOP superfamily/family.

Sequences and sequence diversity Tiam1 and Cask sequences predicted by Proteus, using model B, and by Rosetta are shown as sequence logos, both for the 14 core residues (Fig. 3) and for 16 surface residues (Tiam1 only; Fig. 4), and compared to natural sequences. Agreement with experiment for the core residues is very good, while agreement for the surface residues is much poorer, as seen in previous CPD studies [65, 66]. The behavior of the surface positions is also illustrated by designing each position individually, with the rest of the protein free to explore rotamers but not mutations ("mono-position" design). The corresponding logo shows an excess of Arg and Lys residues, suggesting that the model B reference energies are not yet optimal, despite the extensive empirical E_t^r tuning. Sequence similarity scores are given in the next subsection.

$\star\star\star$ insert Figure 3 near here $\star\star\star$

$\star\star\star$ insert Figure 4 near here $\star\star\star$

The diversity of the natural and designed sequences can be characterized by a mean, exponentiated sequence entropy (see Methods), which corresponds to a mean number of sampled sequence classes per position. The Pfam RP55 set of 12,255 natural sequences has a mean entropy of 3.4. Pooling the designed Tiam1 and Cask sequences gives an entropy of 2.2 with Rosetta and 2.2 with Proteus and model A, or 2.0 with model B, indicating that these two backbone geometries cannot accomodate as much diversity as the much larger RP55 set. Taking the 10000 lowest energy sequences sampled with the room temperature Monte Carlo replica (instead of the 10000 lowest energies sampled by all replicas) and pooling Tiam1 and Cask as before gives a higher overall entropy of 3.0/2.9 with Proteus models A/B. Entropy in the core is only slightly below average with Rosetta and model A (2.1 and 2.0), but is lower (1.25) with Proteus model B, compared to 1.8 for Pfam-RP55.

Blosum similarity scores We also computed Blosum40 similarity scores between designed and natural sequences, shown in Figs. 5 and 6. With model A = (ϵ=8,S1,n=6), for the 14 core residues, the scores for all but one protein overlap with the scores seen within the RP55 set of natural sequences, with values between 20 and 40 (Fig. 5). The syntenin protein, which gave low Superfamily scores, does well in terms of Blosum40 scores. Only for PSD95 are the Proteus scores in the very low range of the experimental scores. Rosetta does somewhat better for this protein. For the seven others, results for the Rosetta sequences are similar on average to Proteus, with some cases a bit better and

others a bit worse. The Proteus sequences for Tiam1 and Cask score mostly above the Rosetta sequences, even though these proteins were not part of model A's E_r^t optimization set.

⋆⋆⋆ insert Figure 5 near here ⋆⋆⋆

⋆⋆⋆ insert Figure 6 near here ⋆⋆⋆

With model B (ϵ=8,S2,n=2), the Cask results are similar to model A and the Tiam1 results slightly improved (whereas the Superfamily results were greatly improved). With model B, we also computed surface and overall similarities (Fig. 6). The overall similarities overlap with the bottom of the peak of experimental scores, and are comparable to the values for the Rosetta sequences. For the surface residues, similarity to the natural sequences is low (scores below zero), both for Proteus and Rosetta. Model B' (not shown) performs about as well as model B, giving the same similarity averaged over all Tiam1 and Cask positions, for example.

While the similarity scores vs. Pfam with our models are comparable to Rosetta, the identity scores vs. the wildtype sequence are significantly higher with Rosetta. Identity scores excluding (respectively, including) Gly and Pro positions (which do not mutate) are 27±6% (37%) for model A vs. 41±9% (49%) for Rosetta. Evidently, for a given level of similarity to Pfam, Rosetta performs ≈10 fewer mutations than Proteus. Sequence identities for Tiam1 and Cask are distinctly lower with all models: 26% and 28% with models A and B, 34% with Rosetta (including Gly, Pro positions).

For certain applications, we may need to specifically explore a sequence space region that is very similar to Pfam, beyond the similarity that is provided by an approximate energy function such as the MMGBSA one. This can be achieved by adding to the energy function an "experimental" bias energy term that explicitly favors high sequence scores. Fig. 6 includes results (dotted line, labelled) that use such a bias energy term (see Methods): by construction, it leads to very high similarity scores, above the mean similarity between Pfam sequences in this particular case. A bias energy term could also be used to limit the total number of mutations with respect to the wildtype sequence.

We also analyzed the similarity between designed sequences generated with the different Proteus models. With model B, overall similarity scores between the Proteus sequences (model B vs. itself) are 453±22 for Tiam1 and 491±20 for Cask. Changing the dielectric constant to 4 (model B') gives sequences less similar to model B, with B-B' scores of 389±18 and 412±30 for Tiam1 and Cask, respectively. Changing the surface

coefficients to set 1 (model A) changes the sequences more substantially, with A-B scores of 173±11 and 150±14 for Tiam1 and Cask, respectively.

4.4 Stability of designed sequences in molecular dynamics simulations

Ten Tiam1 sequences designed with Proteus were chosen for testing in molecular dynamics simulations with an explicit solvent environment. These sequences were obtained from the best design model, model B, and the less polarizable model B'. Although no peptide ligand was present during the design simulations, 11 positions that make close contact with the peptide when it is present were not allowed to mutate. This was done to allow future experimental testing of designed sequences by a peptide binding assay. Among the 2500 lowest energy designed sequences, we narrowed down our choice using the following four criteria: we chose sequences (a) that had a nonneutral isoelectric point, (b) that were assigned to the correct SCOP family by Superfamily with good E-values, (c) that had good Pfam similarity scores, and (d) that did not have too many mutations that drastically change the amino acid type compared to the wildtype protein. This left us with 66 sequences from model B and 45 from model B'. In addition, we eliminated sequences that had two mutations that created a buried cavity and several that had net protein charges of +6 or more (which could lead to instability). Six sequences were chosen; four were modified further by hand to eliminate charged residues in the exposed loop 852–856 (lysines changed manually to alanine), for a final set of ten sequences (Fig. 7A). When these sequences were used as queries to search Uniprot with Blast, the top hits were either Tiam1 mammalian orthologs (including human Tiam1) or uncharacterized proteins, with identity scores between 35 and 40% and Blast E-values of around 10^{-8}–10^{-7} (except for one sequence which gave hits with lower E-values of around 10^{-10}).

MD simulations were run for 100 ns, then extended to 200 ns for the cases that did not exhibit instability within 100 ns (five sequences; Fig. 7). All five appeared stable after 200 ns; one was extended to 600 ns and remained stable; we call it "sequence 6". Another one, sequence 2, was extended to 400 ns and also remained stable. The native Tiam1 protein was simulated for 600 ns and remained stable. The experimental Tiam1 stability is weak, with an unfolding free energy of about 3 kcal/mol [67]. While 200-600 ns simulation times are still short compared to experimental unfolding times (microseconds), they are long enough to challenge the structural models, in the context of an explicit

solvent environment and a high quality protein force field. Indeed, a very weakly stable quadruple mutant (QM) with an unfolding free energy of just 1 kcal/mol [67] was also simulated for 400 ns; it underwent larger deviations than all five designed proteins, with significant loss of structure in the α_2 helix.

$\star\star\star$ insert Figure 7 near here $\star\star\star$

The backbone rms deviations of each sequence compared to the starting, experimental backbone structure (4GVD, or 4NXQ for the quadruple mutant) are shown as a function of time in Fig. 7B, excluding the chain termini (3–4 residues at each end) and one very flexible loop (positions 850–857). The quadruple mutant undergoes the largest deviation from its starting, X-ray structure (4NXQ) [67]. All five designed sequences and the wildtype protein exhibit modest deviations of 1–2 Å away from the starting, X-ray backbone structure for 200 ns or more, suggesting that the five designed sequences are reasonably stable. Sequence 6 stays within 2 Å of the wildtype structure over 600 ns, similar to the wildtype simulation. The mean sequence-6 MD structure (averaged over the first 400 ns) is actually closer to the wildtype experimental structure than the mean wildtype MD structure (rms deviations of 1.2 and 1.5 Å, respectively; Fig. 7C). In the wildtype simulation, performed in the absence of a peptide ligand, helix α_2 shifts into the ligand binding pocket part of the time. A peak in the rms deviation near the end of the sequence-6 simulation corresponds to a transient bending of helix α_2 near its N-terminus, due to interactions that Asn907 and Ser908 make with Glu864 above the helix. This feature is consistent with nuclear magnetic resonance experiments [68], which show that Asn907 and Ser908 have the lowest order parameters (S^2) of the entire protein (outside of the chain termini), indicating significant local flexibility.

4.5 Application: growing the PDZ hydrophobic core

As an application of our optimized models, we examined the designability of the Tiam1, DLG2, Cask, and PSD95 hydrophobic cores. We submitted each protein to Replica Exchange Monte Carlo simulations with a succession of slightly different energy functions that increasingly favor hydrophobic residues. The first simulation included a bias energy term δ = 0.4 kcal/mol (per position) that *penalized* hydrophobic amino acid types (ILMVAWFY). The last simulation included a bias energy term δ = -0.4 kcal/mol (per position) that *favored* hydrophobic types. Intermediate bias energy values δ = 0.2, 0, and -0.2 kcal/mol were also simulated. Results are illustrated in Fig. 8 (Tiam1 only). With

the largest δ value, the Tiam1 hydrophobic core is depleted, with 17 amino acid positions changed to polar types (out of 94). The changed positions mostly lie on the outer edge of the core. With the intermediate δ values, the hydrophobic core is native-like. With the most negative δ value, the hydrophobic core expanded out towards surface regions, with 17 initially polar positions changed to hydrophobic types. Thus, the numbers of positions changed were symmetric (± 17 changes), reflecting the bias. The changes were divided evenly between loop regions and secondary structure elements. The observed propensities of each position to become polar or hydrophobic in the presence of a large or small penalty δ can be thought of as a hydrophobic designability index. Thus, 12 of the 14 conserved core positions (all but 884 and 903) remain hydrophobic even at the highest level of polar bias, along with 22 other positions, indicating that these positions have the highest hydrophobic propensity. We also derive a parameter that describes the relative number of type changes per unit bias energy, as follows. We take the number $\delta N = 34$ of positions that have changed from nonpolar to polar changes when we go from the lowest to the highest bias. We divide this by the bias energy change $\delta E = 0.8$ kcal/mol and by the mean number $N = 52$ of nonpolar positions at zero bias. The result is $\frac{1}{N}\frac{\delta N}{\delta E} = 0.84$ changes (per position) per kcal/mol. This quantity can be thought of as an overall hydrophobic designability or susceptibility. For DLG2, we obtain 0.36 changes per kcal/mol, roughly half the Tiam1 value. While 15 positions change to polar under the strongest polar bias, only 3 change to nonpolar under the strongest nonpolar bias. Thus, DLG2's response was asymetric with respect to the bias sign, with a stronger response (or susceptibility) in the polar direction and a much weaker one in the nonpolar direction. It is as if the protein is saturated with nonpolar sidechains and has difficulty accepting more. For Cask and PSD95, response to the bias is symmetric, as for Tiam1, but smaller, with susceptibilities of 0.47 changes per kcal/mol (Cask) and 0.58 changes per kcal/mol (PSD95), respectively.

$\star\star\star$ insert Figure 8 near here $\star\star\star$

4.6 Application to Tiam1: designing specificity positions

As a second application, we redesigned four amino acid positions in Tiam1 known to contribute to its specific peptide binding. Experimentally, modifying these four positions (quadruple mutant or QM) altered the binding specificity so that it preferentially binds the Caspr4 peptide instead of the syndecan-1 (Sdc1) peptide [30, 67]. The mutations were

L911M, K912E, L915F, L920V. All single and two double mutants were also characterized experimentally. We denote the native sequence LKVV and the quadruple mutant MEFV, and similarly for other variants. All four positions but Lys912 are part of the conserved hydrophobic core. We did Replica Exchange MC simulations where all four positions could mutate simultaneously, in the absence of a peptide ligand, and in the presence of either the Sdc1 or the Caspr4 peptide. Each position could mutate into all types except Gly, Pro. We used Proteus model B' = (ϵ_P=4, S2, n=2), which gave similarity scores (above) equivalent to model B but has a reduced tendency to bury polar sidechains, thanks to its lower dielectric constant. There is no bias energy term involved, only the MMGBSA energy function. The CPD model was constructed either using either the wildtype or the quadruple mutant crystal structure for the backbone (PDB codes 4GVD and 4NXQ, respectively), shown in Fig. 9. The two structures were determined with the Sdc1 and Caspr4 ligands, respectively; nevertheless, they were used here for both holo *and* apo design simulations. The backbone rms deviation between them is 0.9 Å; the main differences are in the flexible 850–857 loop near the peptide C-terminus and in helix α_2, pushed slightly outwards in the mutant complex, to accomodate Phe at protein position 915 and at the peptide C-terminus (position 0: peptide positions are numbered backwards from its C-terminus, as is common for PDZ ligands). We expect that the mutant backbone model (4NXQ) will better describe variants with Phe at position 915 and the wildtype backbone model (4GVD) will better describe variants with a smaller 915 sidechain.

⋆⋆⋆ insert Figure 9 near here ⋆⋆⋆

Results are shown in Fig. 9 as sequence logos. For all systems, the native or native-like amino acid types were sampled at all four positions. For example, using the wild-type (holo) backbone structure (4GVD), Leu911 is preserved in the apo simulations and changed to Ile or Val in the holo simulations. Using the mutant backbone structure (4NXQ), the holo simulations sample Ile, Leu and Met. With the mutant backbone, the holo simulations sample somewhat different types at position 911 (Trp, Arg, Lys); all these types appear in low amounts at the corresponding position in the Pfam seed alignment. All the simulations sample mostly Arg, Lys, Gln and occasionally Glu at position 912, and mostly Leu and Val at position 920, similar to the wildtype sequence. Not surprisingly, agreement with the wildtype sequence is better when we use the wildtype Xray structure; agreement with the mutant sequence is better when we use the mutant Xray structure.

Recovery of the precise native and quadruple mutant sequences in the different states (apo and holo) is qualitative, not quantitative. Thus, using the wildtype backbone structure and in the apo state, the room temperature Monte Carlo replica recovered the wildtype sequence LKLL just 2 kcal/mol above the best sequence (KKLV). The LKML homolog was the second best sequence overall, and the homologs IKLL and LKLV were just 1–2 kcal/mol higher in energy. The mutant sequence MEFV did not appear, nor do any close homologs, probably because there is not space for a Phe sidechain at position 915 with this backbone structure. With the Sdc1 ligand, results are very similar, with the LKLL, IKLL, VKLL, and MKLL sequences all having energies just 1–2 kcal/mol above the best sequence. Notice that the MKLL:Sdc1 affinity is known experimentally, and is within 0.1 kcal/mol of the wildtype value [30]. Experimentally, the wildtype and mutant sequences have the same binding affinity for Caspr4, and stabilities just 2 kcal/mol apart, suggesting that both should be sampled. Instead, neither one is sampled; the closest homolog is IEAV (similar to MEFV), at +2 kcal/mol (relative to the best sequence). This is probably due to steric conflict between position 915 (L or F) and the Caspr4 Phe0 in this backbone geometry.

Using the mutant backbone structure (4NXQ) and in the apo state, the room temperature Monte Carlo replica recovers the mutant MEFV at an energy of +5 kcal/mol (relative to the best sequence) and the wildtype LKLL at +7 kcal/mol. Both protein variants are stable experimentally; a slightly higher energy to produce LKLL seems reasonable, since the design simulation uses the mutant backbone structure, which presumably should favor MEFV. With the Sdc1 ligand, MEFV appears at an energy of +6 kcal/mol, relative to the best sequence, which is the close wildtype homolog IKLV. VKVL is just 3 kcal/mol higher. With the Caspr4 ligand, the mutant sequence appears at an energy of +7 kcal/mol, compared to the best sequence, TKMV; its homologs MQMV and MEMV appear at +5 kcal/mol. The wildtype LKLL and its close homologs do not appear (indicating poorer energies), while MAFI is the second best sequence overall.

A more detailed comparison is possible with the binding affinities of the experimentally characterized mutants [30]. The experiments show that (1) affinity changes associated with each position are roughly independent of the other positions (coupling free energies of 0.4 kcal/mol or less between positions); (2) homologous changes to Leu911, Leu915, and Val920 have a very small effect on the affinity; (3) changing Lys912 to Glu reduces binding by about 0.5–1 kcal/mol (for both peptides, possibly due to lost interactions with the Lys methylenes); (4) changing Leu920 to Phe affects binding differently

depending on the 912 type and the peptide. These properties are mostly reproduced by our simulations. With the wildtype backbone model, considering sequences of the form NKNN (where N $\in$ {I,L,V,M}), the mean apo and Sdc1-bound energies are 0.9$\pm$0.6 and 1.1$\pm$0.5 kcal/mol, respectively, which leads to a mean affinity of 0.2$\pm$0.8 kcal/mol (relative to IKLL, taken as a reference): mutations between the amino acid types I, L, V, and M (N to N' mutations) change the Sdc1 affinity very little, consistent with experiment. Comparing the apo and holo energies sampled in our design simulations, we predict that NKNN $\rightarrow$ NENN mutations lead to affinity changes of +0.75 kcal/mol for both peptides, compared to 0.94 kcal/mol (Sdc1) and 0.55 kcal/mol (Caspr4) experimentally. Similarly, we predict that NKNN $\rightarrow$ NKFN mutations reduce the affinity by 0.5 kcal/mol for both peptides, compared to 1.2 and 0.8 kcal/mol experimentally. Only for NENN $\rightarrow$ NEFN mutations do we see larger errors: we predict a 0.5 kcal/mol affinity loss for Sdc1 (vs. no loss experimentally) and a 0.9 kcal/mol loss for Caspr4 (vs. a 0.5 kcal/mol *gain* experimentally). Specificity changes are predicted to be small, in qualitative agreement with experiment. For example the MKFV $\rightarrow$ MEFV mutation favors Caspr4, relative to Sdc1, by 0.2 kcal/mol, compared to 0.5 kcal/mol experimentally for the homologous LKLL $\rightarrow$ LELL mutation.

The simulations also give information on the correlations between the four mutating positions, illustrated by covariance plots between positions 911 and 912 in Fig. 9. We see that position 911 is more diverse in the apo than the holo state, while 912 is not very sensitive to the peptide. The predicted pairwise correlations between all four protein positions are weak, so that the covariance plots mostly exhibit horizontal and vertical lines or bands, without noticeable "diagonal" structure. This agrees with the experimental affinities of the single, double, and quadruple mutants, where the affinity changes associated with each point mutation are roughly independent of the other positions [30].

5 Discussion

We have examined several parameterizations of a simple CPD model for PDZ design, suitable for high throughput design applications, implemented in the Proteus software. For the folded state representation, we use a high quality protein force field and Generalized Born solvent model. We tested two sets of atomic surface energy parameters σ_i and two protein dielectric constants ϵ_P; there quantities characterize the solvent model and are the main empirical parameters in the folded state energy function. We use a specific set of

X-ray structures for our test proteins, with a specific conformation for each protein's backbone. For the sidechains, we use a simple, discrete rotamer library; the discrete rotamer approximation is alleviated through a short minimization of each sidechain pair during the energy matrix calculation. Both the energy function and the rotamer description have been extensively tested, giving very good performance for sidechain reconstruction tests [48] (comparable to the popular Scwrl4 program [47]) and good performance for a large set of protein acid/base constants [69] (superior to the Rosetta software [70], despite extensive *ad hoc* parameter tuning).

For the unfolded state representation, we use a simple, implicit model, characterized by a set of empirical, amino acid chemical potentials or reference energies. These energies are chosen by a likelihood maximization procedure, formulated here, in order to reproduce the amino acid composition of carefully selected natural homologs. The unfolded state description is more refined than previously [71], since distinct reference energy values are used for amino acid positions that are buried or exposed in the *folded* state. This method hypothesizes that there is residual structure in the unfolded state, with some positions more buried than others. Furthermore, it should make the parameterization more robust and less sensitive to the size and structure of the natural homologs used to define the target amino acid compositions, because the amino acid frequencies of exposed and buried regions are averaged separately. Although this doubles, in principle, the number of adjustable reference energies, we reduced this number by introducing amino acid similarity classes, with one adjustable reference energy per class. To optimize the reference energies, we did design calculations for each test protein (apo state) where half of the amino positions could mutate at a time (excluding Gly and Pro), with distinct simulations for each half. This way, during parameter optimization, a mutating position was always surrounded by an environment at least 50% identical to the wildtype sequence. The design calculations relied on a powerful and efficient Replica Exchange Monte Carlo exploration method, using over a half billion steps per simulation (per replica), and producing thousands of designed sequences in a single simulation. Overall, reference energy values were optimized with two different sets of natural homologs as targets (n=2 and n=6 sets) and four different parameterizations of the implicit solvent model used for the folded state (different choices of the σ_i and ϵ_P). Encouragingly, performance levels in the subsequent tests were fairly robust across model parameterizations.

The model has several limitations, most of which are widespread in CPD implementations and applications. One is the use of protein stability as the sole design criterion,

without explicitly accounting for fold specificity [72], protection against aggregation, or functional considerations like ligand binding or binding between the PDZ domain and other domains or proteins. We note, however, that the Superfamily tests did not lead to any fold misassignments (sequences that prefer another SCOP fold), so that fold specificity was achieved in practice. Functional criteria can also be introduced in an *ad hoc way*; for example, the sequences tested by MD were originally designed with their 11 peptide binding residues not allowed to mutate.

A second model limitation is the use of a fixed protein backbone. In fact, the backbone is not really fixed: rather, certain motions are allowed but modeled *implicitly*, through the use of a protein dielectric constant greater than one ($\epsilon_P = 4$ or 8). This dielectric value means that the protein structure (including its backbone) is allowed to relax or reorganize in response to charge redistribution associated with mutations or sidechain rotamer changes; however the reorganization is modeled implicitly, not explicitly [73], and it does not involve motion of the atomic centers or their associated van der Waals spheres. Thus, the backbone cannot reorganize in response to steric repulsion produced by mutations or rotamer changes, nor can it shift to fill space left empty by a mutation. The effect of this approximation was apparent in the design of the four Tiam1 specificity positions, where the designed sequences were sensitive to the particular conformation of the protein and peptide backbones. Specifically, with the wildtype backbone structure, there was not room to insert a Phe sidechain at position 920, even though Phe920 is present in the experimental quadruple mutant (which has a slightly different backbone structure). Therefore, the choice of the initial X-ray model is important, and several strategies are possible. Here, to parameterize the CPD model, we used a mixture of X-ray structures solved with and without a peptide ligand, even though the parameterization simulations and most of the testing were done for the apo proteins. This choice was made partly because the apo/holo PDZ structures are quite similar and partly to make the model more transferable and facilitate applications to peptide binding. Another strategy would have been to parameterize the model using all apo structures, then switch to holo structures for the Tiam1 application.

For whole protein design, the use of a fixed backbone is partly counterbalanced by designing two or more PDZ structures. For example, pooling the designed Tiam1 and Cask sequences gives a mean sequence entropy comparable to the experimental Pfam set, and allows us to recapitulate more sequences than design with just one backbone. In the Tiam1 4-position design, the fixed backbone was also counterbalanced by doing

calculations separately with two different backbone structures, a holo wildtype and a holo mutant structure. Simulations with the mutant backbone allowed us to obtain mutants having Phe at position 920. Notice that a new method for multibackbone design was recently developed in Proteus, using a novel, non-heuristic hybrid Monte Carlo method that preserves Boltzmann sampling [74]; this method could be applied in the future.

A third limitation of our model is the need, for optimal results, to parameterize the reference energies specifically for a given set of proteins. This step is well-automated and highly parallel. However, it involves several chioces that are partly arbitrary. These include the choice of a set of protein domains to represent the protein or family of interest. Here we obtained better results with a model parameterized using just two PDZ domains (n=2 models), compared to a model parameterized using six PDZ domains. We also need to choose a similarity threshold to define the target homologs from which we compute the experimental amino acid compositions. Here, we chose to use close homologs of each family member, compute their compositions, then average over the two or six family representatives. This method worked well but other choices are possible, and more work is needed to draw definitive conclusions. Also, the mono-position design of Tiam1 showed evidence of some systematic error (Fig. 4), with a large fraction of Arg, Lys, and Gln residues types on the protein surface, despite the optimized reference energies. In the future, it may be necessary to relax the intra-group constraints towards the end of the reference energy optimization and/or target smaller numbers of mutating positions, instead of one half of the protein at a time.

A fourth limitation of our model is the discrete rotamer approximation, which requires some adaptation of the energy function to avoid exaggerated steric clashes; the method used here is the residue-pair minimization method described earlier [26, 71]. A fifth limitation is the use of a pairwise additive solvation model (as in most CPD models). Specifically, the dielectric environment of each residue pair is assumed here to be native-like (NEA approximation). This leads to an energy function that has the form of a sum over pairs of residues and can be pre-calculated and stored in an energy function, which then serves as a lookup table during the subsequent Monte Carlo simulations. Despite this approximation, our model gave good results for a large acid/base benchmark, a problem that is very sensitive to the electrostatic treatment [69].

Some of these limitations could be removed in future work. In particular, since our energy function is mostly physics-based, it can benefit rapidly from ongoing improvements in protein force fields and solvation models. Thus, the NEA approximation could be

removed in the future thanks to the recent implementation (manuscript in preparation) of a more exact Generalized Born calculation, whose efficiency is not much reduced compared to the pairwise approximation [75]. We have also implemented an improved model for hydrophobic solvation [76], which is faster and more accurate than our current surface area energy term (manuscript in preparation).

Designed sequences were extensively compared to experiment, through fold recognition tests, similarity calculations, and entropy calculations. In the test simulations, we design the entire protein sequence, so that all positions (except Gly and Pro) can mutate freely, subject only to an overall bias towards the mean, experimental amino acid composition (through the reference energies). Despite the lack of experimental bias or constraints, the resulting sequences have a high overall similarity to the natural, Pfam sequences, as measured by the Blosum40 similarity scores. The scores obtained are mostly comparable to the similarity scores between pairs of Pfam sequences themselves. Thus, the sequences designed with Proteus resemble moderately-distant natural homologs. The similarity is very strong for residues in the core of the protein, as observed in previous CPD studies [65, 66]. For the protein surface, similarity scores are close to zero, which is the score we would obtain if we pick amino acid types randomly. Notice that many surface residues are involved in functional interactions, such as the eleven peptide-binding residues. Surface residues are also selected to avoid aggregation or unwanted adhesion. These functional constraints are not accounted for in the design (although the energy function indirectly favors protein solubility). Despite the limited similarity scores for surface regions, fold recognition with the Superfamily tool and the best design models was almost perfect. Earlier fold recognition tests that used a simpler energy function gave a lower fold recognition rate of about 85% (for a larger and more diverse test set) and lower similarities [15, 44]. Evidently, the use of an improved protein force field, Generalized Born solvent, and family-specific reference energies leads to better designed sequences.

The Proteus sequences were also compared to ones designed with the Rosetta software (using default parameters), which has itself been extensively tested. Judging by the Blosum similarity scores (vs. natural sequences) and the fold recognition tests, the Proteus and Rosetta sequences appear to be of about the same quality. However, Rosetta makes fewer mutations than Proteus, so that the identity scores, compared to the corresponding wildtype protein, are about 12 percent points higher. This means that Proteus mutates about ten more positions, on average, per PDZ domain. This number could easily be reduced, by adding to the Proteus energy function an explicit bias energy term that

increases with the number of mutations (away from the wildtype sequence). An equivalent bias energy was used above for just two simulations, of Tiam1 and Cask, respectively (Fig. 6), to illustrate the possibility of using experimentally-restrained sampling. It remains to be seen whether a restraint based on the identity score would lead to more stable and realistic designed sequences.

Another attractive route for testing designed sequences is through high level MD simulations. Here, ten designed Tiam1 sequences were tested in rather long MD simulations, in the apo form, using the same protein force field as in the CPD model (Amber force field) and an established explicit water model. These sequences were designed using Proteus, with Gly, Pro, and eleven peptide-binding positions held fixed but all others free to mutate. Five of the sequences moved away from the starting structure within 100 ns (2–3 Å backbone deviation), suggesting that they are weakly stable. The other five remained stable over much longer simulation lengths (200, 200, 200, 400, and 600 ns, respectively). Their deviations from the starting, experimental structures (1–2 Å) are similar to the deviations observed in the wildtype Tiam1 simulation, and lower than the deviations observed in the simulation of the (weakly stable) quadruple mutant Tiam1 (experimental unfolding free energy of 1 kcal/mol [67]). Direct experimental testing of the designed sequences remains to be done.

The model was used for two applications. Hydrophobic "titration" of four PDZ domains illustrates a novel way to characterize protein designability. The cost or availability of hydrophobic sidechain types is modelled through a bias energy term that is gradually varied. One result is the mean overall hydrophobic "susceptibility", the number of type changes per kcal/mol and per position, which differs by a factor of two between Tiam1 and DLG2 (2BYG), with intermediate values for Cask and PSD95 (3K82). In Tiam1, 12 of the core positions remained hydrophobic even with the largest bias value favoring polar types, while 17 other positions changed to polar (respectively, nonpolar) types at the largest polar (nonpolar) bias energy values. In DLG2, unlike Tiam1, when the nonpolar bias increased, very few positions changed: the wildtype protein is "hydrophobically saturated", and does not tolerate additional nonpolar residues. A position-by-position analysis would be of interest, as well as a comparison to other domain families; these aspects are left for future work.

Redesign of four specificity positions in Tiam1 allowed us to test the design model in a different way, revealing some of the limitations of fixed backbone design, but also giving semi-quantitative agreement with available binding free energies. This agreement suggests

that some of the new, predicted mutations could be of interest and provide new specificity properties. They remain to be studied further and tested experimentally. Here, the apo and two holo states were studied, and designed separately. Information about binding affinities and binding specificity were then obtained by comparing the energies sampled in the different simulations. In principle, we would like to include binding affinity and/or specificity directly in the design calculations, as a property to be designed for or against within a single simulation. In addition, we should in principle allow different backbone structures for the apo and each holo system. This could be done in the future, since recent hybrid Monte Carlo schemes [74, 77] can be used for multi-backbone design, and can be extended to the problem of designing ligand binding specificity. We also note that since our energy function is physics-based, it has transferability to a range of molecule types, such as nonnatural amino acids (considered in an earlier protein-ligand design study [78]). Such amino acids could be of interest for designing PDZ peptide ligands, to provide additional diversity and perhaps enhanced resistance to proteolysis.

Acknowledgements

We are grateful for helpful discussions with Michael Schnieders and Young Joo Sun (University of Iowa). Some of the calculations were done at the CINES supercomputer center of the French Ministry of Research. NAMD was developed by the Theoretical and Computational Biophysics Group in the Beckman Institute at the University of Illinois at Urbana.

References

[1] Harris BZ, Lim WA. Mechanism and role of PDZ domains in signaling complex assembly. J Cell Sci. 2001;114:3219–3231.

[2] Hung AY, Sheng M. PDZ Domains: Structural Modules for Protein Complex Assembly. J Biol Chem. 2002;277:5699–5702.

[3] Tonikian R, Zhang YN, Sazinsky SL, Currell B, Yeh JH, Reva B, et al. A Specificity Map for the PDZ Domain Family. PLoS Biology. 2008;6:2043–2059.

[4] Gfeller D, Butty F, Wierzbicka M, Verschueren E, Vanhee P, Huang H, et al. The multiple-specificity landscape of modular peptide recognition domains. Molec Syst Biol. 2011;7:484.

[5] Subbaiah VK, Kranjec C, Thomas M, Ban L. Structural and thermodynamic analysis of PDZ-ligand interactions. Biochem J. 2011;439:195–205.

[6] Shepherd TR, Fuentes EJ. Structural and thermodynamic analysis of PDZ-ligand interactions. Methods in Enzymology. 2011;488:81–100.

[7] Bacha A, Clausen BH, Moller M, Vestergaard B, Chic CN, Round A, et al. A high-affinity, dimeric inhibitor of PSD-95 bivalently interacts with PDZ1-2 and protects against ischemic brain damage. Proc Natl Acad Sci USA. 2012;109:3317–3322.

[8] Roberts KE, Cushing PR, Boisguerin P, Madden DR, Donald BR. Computational Design of a PDZ Domain Peptide Inhibitor that Rescues CFTR Activity. PLoS Comp Bio. 2012;8:e1002477.

[9] Zheng F, Jewell H, Fitzpatrick J, Zhang J, Mierke DF, Grigoryan G. Computational Design of Selective Peptides to Discriminate between Similar PDZ Domains in an Oncogenic Pathway. J Mol Biol. 2015;427:491–510.

[10] Lockless W, Ranganathan R. Evolutionary Conserved Pathways of Energetic Connectivity in Protein Families. Science. 1999;295:295–299.

[11] Kong Y, Karplus M. Signaling pathways of PDZ2 domain: A molecular dynamics interaction correlation analysis. Proteins. 2009;74:145–154.

[12] McLaughlin Jr RN, Poelwijk FJ, Raman A, Gosal WS, Ranganathan R. The spatial architecture of protein function and adaptation. Nature. 2012;458:859–864.

[13] Melero C, Ollikainen N, Harwood I, Karpiak J, Kortemme T. Quantification of the transferability of a designed protein specificity switch reveals extensive epistasis in molecular recognition. Proc Natl Acad Sci USA. 2014;111:15426–15431.

[14] Reina J, Lacroix E, Hobson SD, Fernandez-Ballester G, Rybin V, Schwab MS, et al. Computer-aided design of a PDZ domain to recognize new target sequences. Nat Struct Mol Biol. 2002;9:621–627.

[15] Schmidt am Busch M, Sedano A, Simonson T. Computational protein design: validation and possible relevance as a tool for homology searching and fold recognition. PLoS One. 2010;5(5):e10410.

[16] Smith CA, Kortemme T. Structure-Based Prediction of the Peptide Sequence Space Recognized by Natural and Synthetic PDZ Domains. J Mol Biol. 2010;402:460–474.

[17] Butterfoss GL, Kuhlman B. Computer-based design of novel protein structures. Ann Rev Biophys Biomolec Struct. 2006;35:49–65.

[18] Lippow SM, Tidor B. Progress in computational Protein Design. Curr Opin Biotech. 2007;18:305–311.

[19] Dai L, Yang Y, Kim HR, Zhou Y. Improving computational protein design by using structure-derived sequence profile. Proteins. 2010;78:2338–2348.

[20] Feldmeier K, Hoecker B. Computational protein design of ligand binding and catalysis. Curr Opin Chem Biol. 2013;17:929–933.

[21] Tinberg CE, Khare SD, Dou J, Doyle L, Nelson JW, Schena A, et al. Computational design of ligand-binding proteins with high affinity and selectivity. Nature. 2013;501:212–218.

[22] Au L, Green DF. Direct Calculation of Protein Fitness Landscapes through Computational Protein Design. Biophys J. 2016;110:75–84.

[23] Pokala N, Handel TM. Energy functions for protein design I: Efficient and accurate continuum electrostatics and solvation. Prot Sci. 2004;13:925–936.

[24] Samish I, MacDermaid CM, Perez-Aguilar JM, Saven JG. Theoretical and computational protein design. Ann Rev Phys Chem. 2011;62:129—149.

[25] Li Z, Yang Y, Zhan J, Dai L, Zhou Y. Energy Functions in De Novo Protein Design: Current Challenges and Future Prospects. Ann Rev Biochem. 2013;42:315–335.

[26] Schmidt am Busch M, Lopes A, Mignon D, Simonson T. Computational protein design: software implementation, parameter optimization, and performance of a simple model. J Comput Chem. 2008;29:1092–1102.

[27] Simonson T. Protein:ligand recognition: simple models for electrostatic effects. Curr Pharma Design. 2013;19:4241–4256.

[28] Polydorides S, Michael E, Mignon D, Druart K, Archontis G, Simonson T. Proteus and the design of ligand binding sites. In: Stoddard B, editor. Methods in Molecular Biology: Design and Creation of Protein Ligand Binding Proteins. vol. 1414. Springer Verlag, New York; 2016. p. 0000.

[29] Baker D. Prediction and design of macromolecular structures and interactions. Phil Trans R Soc Lond. 2006;361:459–463.

[30] Shepherd TR, Hard RL, Murray AM, Pei D, Fuentes EJ. Distinct Ligand Specificity of the Tiam1 and Tiam2 PDZ Domains. Biochemistry. 2011;50:1296–1308.

[31] Frenkel D, Smit B. Understanding molecular simulation, Chapter 3. Academic Press, New York; 1996.

[32] Grimmett GR, Stirzaker DR. Probability and random processes. Oxford University Press; 2001.

[33] Kleinman CL, Rodrigue N, Bonnard C, Philippe H, Lartillot N. A maximum likelihood framework for protein design. BMC Bioinf. 2006;7:Art. 326.

[34] Fowler RH, Guggenheim EA. Statistical Thermodynamics. Cambridge University Press; 1939.

[35] Cornell W, Cieplak P, Bayly C, Gould I, Merz K, Ferguson D, et al. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. J Am Chem Soc. 1995;117:5179–5197.

[36] Hawkins GD, Cramer C, Truhlar D. Pairwise descreening of solute charges from a dielectric medium. Chem Phys Lett. 1995;246:122–129.

[37] Lopes A, Aleksandrov A, Bathelt C, Archontis G, Simonson T. Computational sidechain placement and protein mutagenesis with implicit solvent models. Proteins. 2007;67:853–867.

[38] Lee B, Richards F. The interpretation of protein structures: estimation of static accessibility. J Mol Biol. 1971;55:379–400.

[39] Brünger AT. X-PLOR version 3.1, A System for X-ray crystallography and NMR. Yale University Press, New Haven; 1992.

[40] Gaillard T, Simonson T. Pairwise Decomposition of an MMGBSA Energy Function for Computational Protein Design. J Comput Chem. 2014;35:1371–1387.

[41] Street AG, Mayo SL. Pairwise calculation of protein solvent-accessible surface areas. Folding and Design. 1998;3:253–258.

[42] Aleksandrov A, Polydorides S, Archontis G, Simonson T. Predicting the Acid/Base Behavior of Proteins: A Constant-pH Monte Carlo Approach with Generalized Born Solvent. J Phys Chem B. 2010;114:10634–10648.

[43] Mishra P, Socolich M, Wall MA, Graves J, Wang Z, Ranganathan R. Dynamic Scaffolding in a G Protein-Coupled Signaling System. Cell. 2007;131:80–92.

[44] Schmidt am Busch M, Mignon D, Simonson T. Computational protein design as a tool for fold recognition. Proteins. 2009;77:139–158.

[45] Eswar N, Marti-Renom MA, Webb B, Madhusudhan MS, Eramian D, Shen M, et al. Comparative Protein Structure Modeling With MODELLER. Curr Prot Bioinf. 2006;Suppl. 15:5.6.1–5.6.30.

[46] Tuffery P, Etchebest C, Hazout S, Lavery R. A New Approach to the Rapid Determination of Protein Side Chain Conformations. J Biomol Struct Dyn. 1991;8:1267.

[47] Krivov GG, Shapalov MV, Dunbrack RL. Improved prediction of protein side-chain conformations with SCWRL4. Proteins. 2009;77:778–795.

[48] Gaillard T, Panel N, Simonson T. Protein sidechain conformation predictions with an MMGBSA energy function. Proteins. 2016;84:803–819.

[49] Kofke DA. On the acceptance probability of replica-exchange Monte Carlo trials. J Chem Phys. 2002;117:6911–6914.

[50] Earl DJ, Deem MW. Parallel tempering: theory, applications, and new perspectives. Phys Chem Chem Phys. 2005;7:3910–3916.

[51] Mignon D, Simonson T. Comparing three stochastic search algorithms for computational protein design: Monte Carlo, Replica Exchange Monte Carlo, and a multistart, steepest-descent heuristic. J Comput Chem. 2016;37:1781–1793.

[52] Gough J, Karplus K, Hughey R, Chothia C. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. J Mol Biol. 2001;313:903–919.

[53] Wilson D, Madera M, Vogel C, Chothia C, Gough J. The SUPERFAMILY database in 2007: families and functions. Nucl Acids Res. 2007;35:D308—D313.

[54] Andreeva A, Howorth D, Brenner SE, Hubbard JJ, Chothia C, Murzin AG. SCOP database in 2004: refinements integrate structure and sequence family data. Nucl Acids Res. 2004;32:D226–229.

[55] Durbin R, Eddy SR, Krogh A, Mitchison G. Biological sequence analysis. Cambridge University Press; 2002.

[56] Murphy LR, Wallqvist A, Levy RM. Simplified amino acid alphabets for protein fold recognition and implications for folding. Prot Eng. 2000;13:149–152.

[57] Launay G, Mendez R, Wodak SJ, Simonson T. Recognizing protein-protein interfaces with empirical potentials and reduced amino acid alphabets. BMC Bioinf. 2007;8:270–291.

[58] Jo S, Kim T, Iyer VG, Im W. CHARMM-GUI: a web-based graphical user interface for CHARMM. J Comput Chem. 2008;29:1859–1865.

[59] Darden T. Treatment of long-range forces and potential. In: Becker O, Mackerell Jr AD, Roux B, Watanabe M, editors. Computational Biochemistry & Biophysics. Marcel Dekker, N.Y.; 2001.

[60] Jorgensen W, Chandrasekar J, Madura J, Impey R, Klein M. Comparison of simple potential functions for simulating liquid water. J Chem Phys. 1983;79:926–935.

[61] Brooks B, Brooks III CL, Mackerell Jr AD, Nilsson L, Petrella RJ, Roux B, et al. CHARMM: The biomolecular simulation program. J Comp Chem. 2009;30:1545–1614.

[62] Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, et al. Scalable molecular dynamics with NAMD. J Comput Chem. 2005;26:1781–1802.

[63] Madera M, Vogel C, Kummerfeld SK, Chothia C, Gough J. The SUPERFAMILY database in 2004: additions and improvements. Nucl Acids Res. 2004;32:D235–D239.

[64] Schmidt am Busch M, Lopes A, Amara N, Bathelt C, Simonson T. Testing the Coulomb/Accessible Surface Area solvent model for protein stability, ligand binding, and protein design. BMC Bioinformatics. 2008;9:148–163.

[65] Jaramillo A, Wernisch L, Héry S, Wodak S. Folding free energy function selects native-like protein sequences in the core but not on the surface. Proc Natl Acad Sci USA. 2002;99:13554–13559.

[66] Dantas G, Kuhlman B, Callender D, Wong M, Baker D. A Large Test of Computational Protein Design: Folding and Stability of Nine Completely Redesigned Globular Proteins. J Mol Biol. 2003;332:449–460.

[67] Liu X, Speckhard DC, Shepherd TR, Sun YJ, Hengel SR, Yu L, et al. Distinct roles for conformational dynamics in protein-ligand interactions. Structure. 2016;24:0000.

[68] Liu X, Shepherd TR, Murray AM, Xu Z, Fuentes EJ. The Structure of the Tiam1 PDZ Domain/Phospho-Syndecan1 Complex Reveals a Ligand Conformation that Modulates Protein Dynamics. Structure. 2013;21:342–354.

[69] Polydorides S, Simonson T. Monte Carlo simulations of proteins at constant pH with generalized Born solvent, flexible sidechains, and an effective dielectric boundary. J Comput Chem. 2013;34:2742–2756.

[70] Kilambi K, Gray JJ. Rapid calculation of protein pK_a values using Rosetta. Biophys J. 2012;103:587–595.

[71] Simonson T, Gaillard T, Mignon D, Schmidt am Busch M, Lopes A, Amara N, et al. Computational protein design: the Proteus software and selected applications. J Comput Chem. 2013;34:2472–2484.

[72] Mach P, Koehl P. Capturing protein sequence–structure specificity using computational sequence design. Proteins. 2013;81:1556–1570.

[73] Simonson T. What Is the Dielectric Constant of a Protein When Its Backbone Is Fixed? J Chem Theory Comput. 2013;9:4603–4608.

[74] Druart K, Bigot J, Audit E, Simonson T. A hybrid Monte Carlo method for multibackbone protein design. J Chem Theory Comput. 2017;in press:0000.

[75] Archontis G, Simonson T. A residue-pairwise Generalized Born scheme suitable for protein design calculations. J Phys Chem B. 2005;109:22667–22673.

[76] Aguilar B, Shadrach R, Onufriev AV. Reducing the secondary structure bias in the generalized Born model via R6 effective radii. J Chem Theory Comput. 2011;6:3613–3630.

[77] Ollikainen N, de Jong RM, Kortemme T. Coupling Protein Side-Chain and Backbone Flexibility Improves the Re-design of Protein-Ligand Specificity. PLoS Comp Bio. 2015;1:e1004335.

[78] Druart K, Palmai Z, Omarjee E, Simonson T. Protein:ligand binding free energies: a stringent test for computational protein design. J Comput Chem. 2016;37:404–415.

Figure captions

1. **A)** Three dimensional view of four PDZ domains. The C_β atoms of fourteen hydrophobic core residues are shown as spheres. Three core positions designed further on are labelled (Tiam1 numbers). **B)** Cluster representation: links connect pairs that have backbone rms deviations over at least 60 aligned residues of 1 Å or less, except for Cask (1.3 Å) and Tiam1. The links are labeled with the backbone rms deviations (Å) and percent identity scores. Tiam1 is connected to its closest neighbors by dashed gray lines, because its rms deviations are above the 1 Å threshold and computed over fewer (40–45) residues.

2. Alignment of natural PDZ sequences. The top eight are our test proteins; the others are the first 22 sequences from the Pfam seed alignment. Fourteen hydrophobic core positions are indicated by red arrows and the secondary structure elements are shown for reference.

3. Sequence logos for the conserved hydrophobic core of designed and natural Tiam1 and Cask sequences. "Hom." corresponds to the close homologs that make up our target set of sequences (used for E_t^r optimization); "Pfam" corresponds to the Pfam seed alignment. Proteus sequences were generated with model B. The height of each letter is proportional to the abundance of each type at the corresponding position in the Proteus/Rosetta simulations or the natural sequences.

4. Sequence logos for sixteen surface positions in Tiam1. Same representation as Fig. 3. The "mono-position" results are from a set of simulations where only one amino acid at a time can mutate, the rest of the protein having its native sequence (see text).

5. Histogram plots showing similarity scores for designed PDZ sequences. Similarity scores for the eight tested PDZ domains and fourteen conserved core positions were determined using the Blosum40 matrix, relative to the Pfam-RP55 reference alignment. See text for details. Similarity scores are shown for Proteus (model A), Rosetta, and Pfam sequences. The similarity score of the wildtype sequence is indicated in each panel by a vertical arrow.

6. Histogram plots showing similarity scores for designed PDZ sequences. Similarity scores for Tiam1 and Cask, relative to the Pfam-RP55 reference alignment. The similarity scores are computed either for all positions (top), for fourteen core positions (middle), and for sixteen surface positions (bottom). Values are shown for Proteus, Rosetta, and Pfam sequences (all compared to RP55). The similarity score of the wildtype sequence is indicated in each panel by a vertical arrow. The top panels include results for Proteus simulations where a bias energy term was included, which explicitly favors sequences that are similar to Pfam (dotted lines, labeled "Biased Proteus"). Notice that the designed

sequences represented in each top, middle, or bottom panel are the same; only the positions included in the score calculation differ between panels.

7. Molecular dynamics simulations of the wildtype Tiam1 PDZ domain, a quadruple mutant (QM) and five designed sequences. **A)** The wildtype sequence and five sequences designed with Proteus are shown. Eleven peptide binding residues that were held fixed during the design simulations are labeled with stars. **B)** Backbone rms deviations (Å) are shown for each protein over the course of an MD simulation. "WT" corresponds to the wildtype simulation; "QM" corresponds to a quadruple mutant (see text). Deviations are computed relative to the starting structure, excluding seven residues at the chain termini and the flexible 850–857 loop. **C)** The mean structures from the wildtype MD simulation (gray) and the sequence-6 MD simulation (red), compared to the starting, X-ray structure (green). Twenty instantaneous snapshots taken every 20 ns from the sequence-6 MD simulation are also shown (light, semi-transparent pink).

8. **A)** Tiam1 sequences designed with different levels of hydrophobic bias. The top (respectively, bottom) sequences were obtained in the presence of a bias δ opposing (favoring) hydrophobic types. The middle sequences were obtained from Proteus simulations without any bias ($\delta = 0$). For each hydrophobic bias level, five low energy sequences from the corresponding simulation are shown. Hydrophobic positions are colored according to the simulation where they appear first: from brick red (top) to light green (bottom). The fourteen hydrophobic core positions are indicated by red arrows. **B)** Three dimensional Tiam1 structure (stereo) with residue colors as in **A)**. The backbone is shown in light, semi-transparent gray.

9. Design of four Tiam1 specificity positions. **A)** X-ray structures (stereo) with the wildtype sequence (LKLL; labelled WT; PDB code 4GVD) or the quadruple mutant sequence (MEFL; labelled QM; PDB code 4NXQ), with bound Sdc1 and Caspr4, respectively. The four designed sidechains are shown and labelled (both wildtype and quadruple mutant types). **B)** Logo representation of designed sequences with no ligand (apo) or Sdc1 or Caspr4, using the wildtype (left) or quadruple mutant (right) X-ray structure. **C)** Covariance plots for the 911-912 pair: populations of each pair of types are shown as levels of color, with yellow the most highly-populated (5%) and red the lowest (0%). Results correspond to design simulations that used the wildtype backbone.

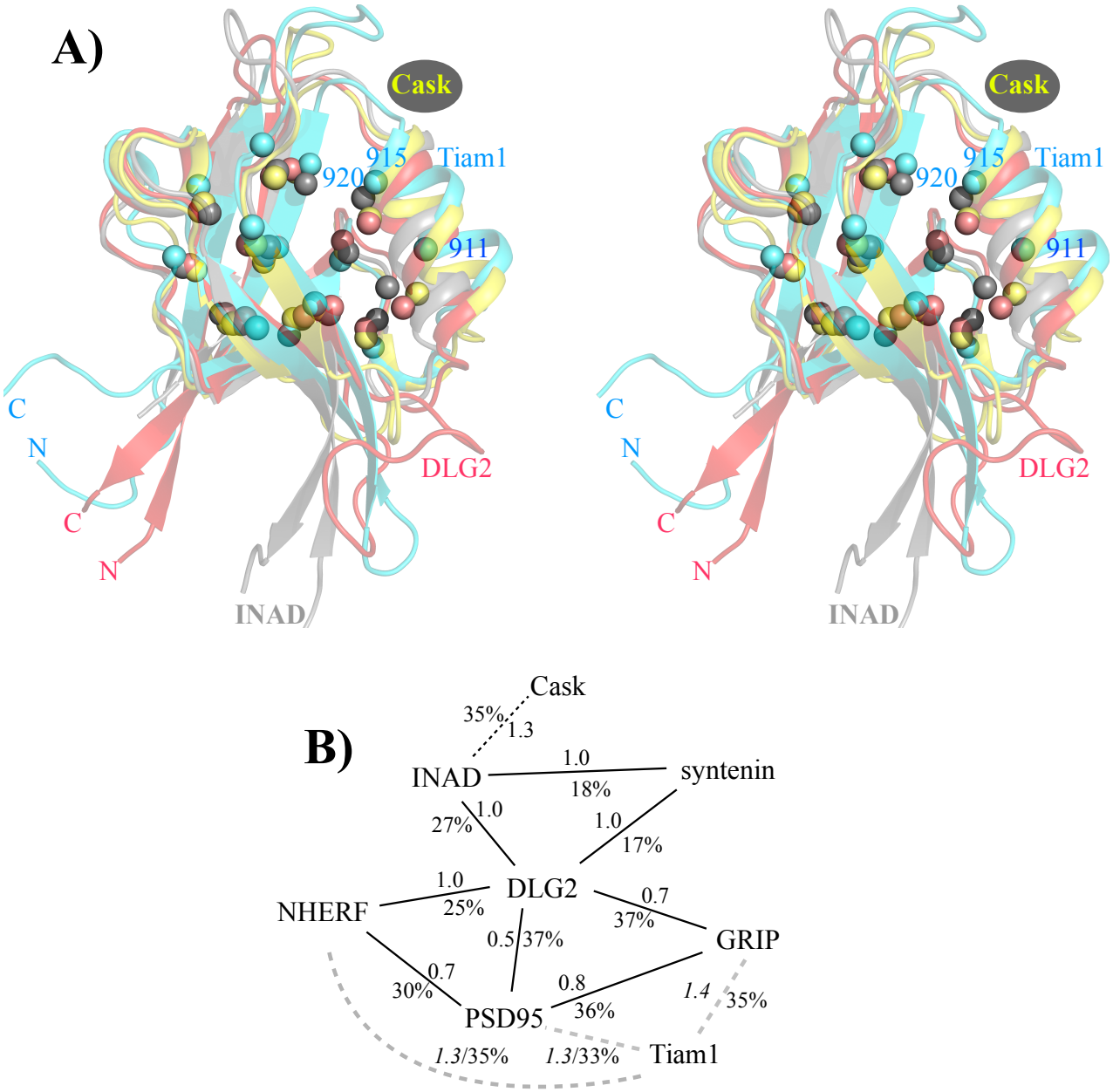

Figure 1: **A)** 3D view of four PDZ domains. **B)** Cluster representation.

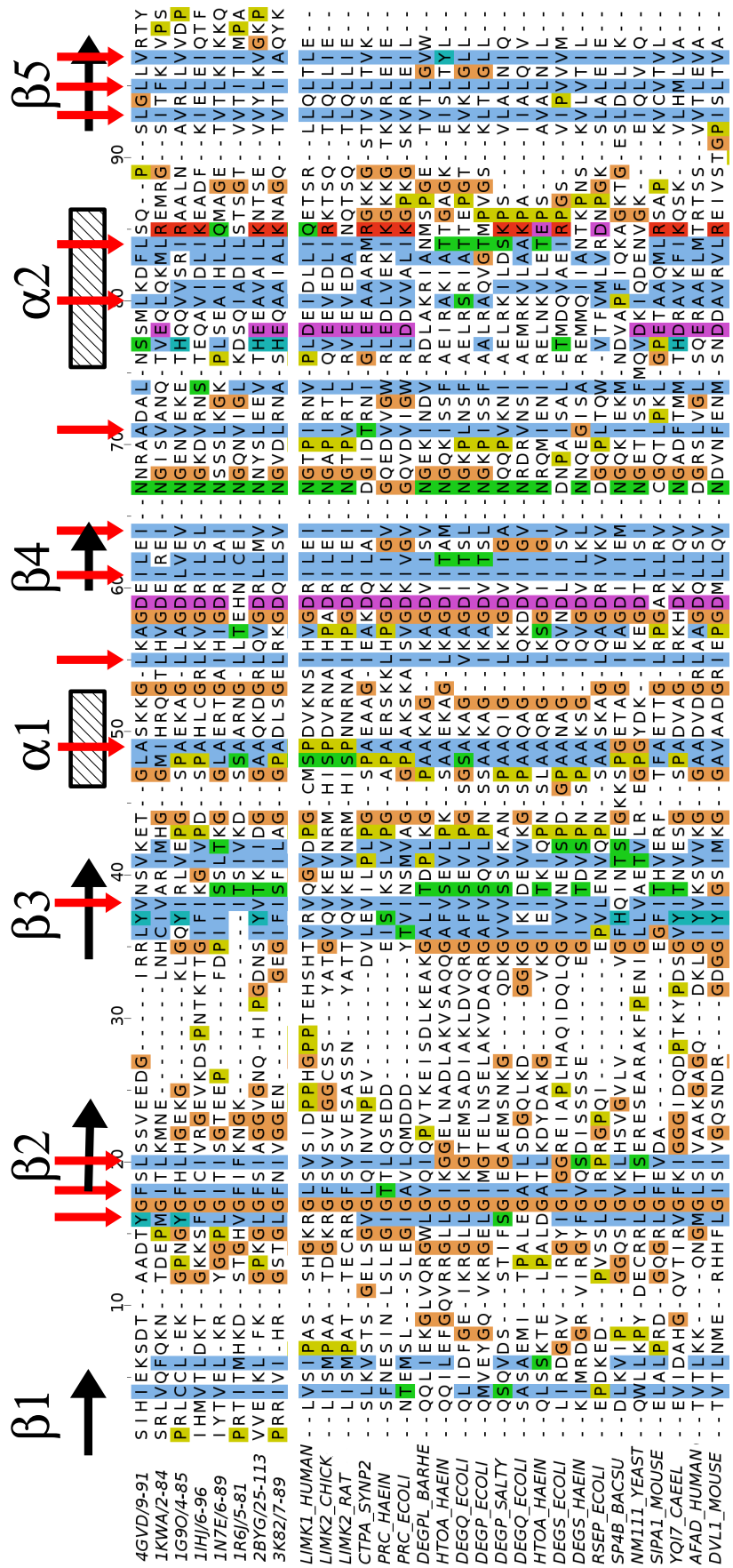

Figure 2: Alignment of experimental PDZ sequences.

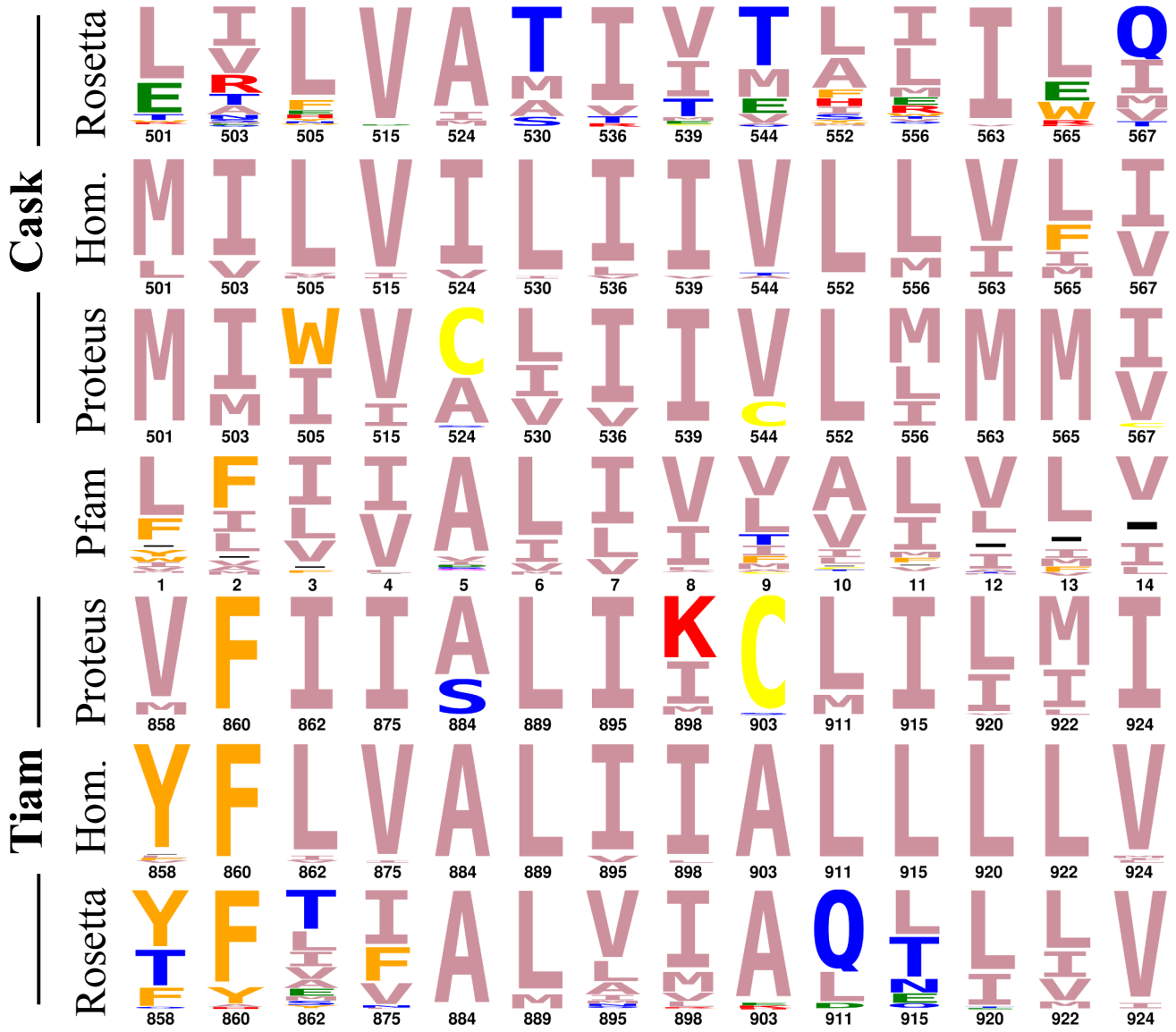

Figure 3: Core logos for Tiam1 and Cask.

Figure 4: Surface logos for Tiam1.

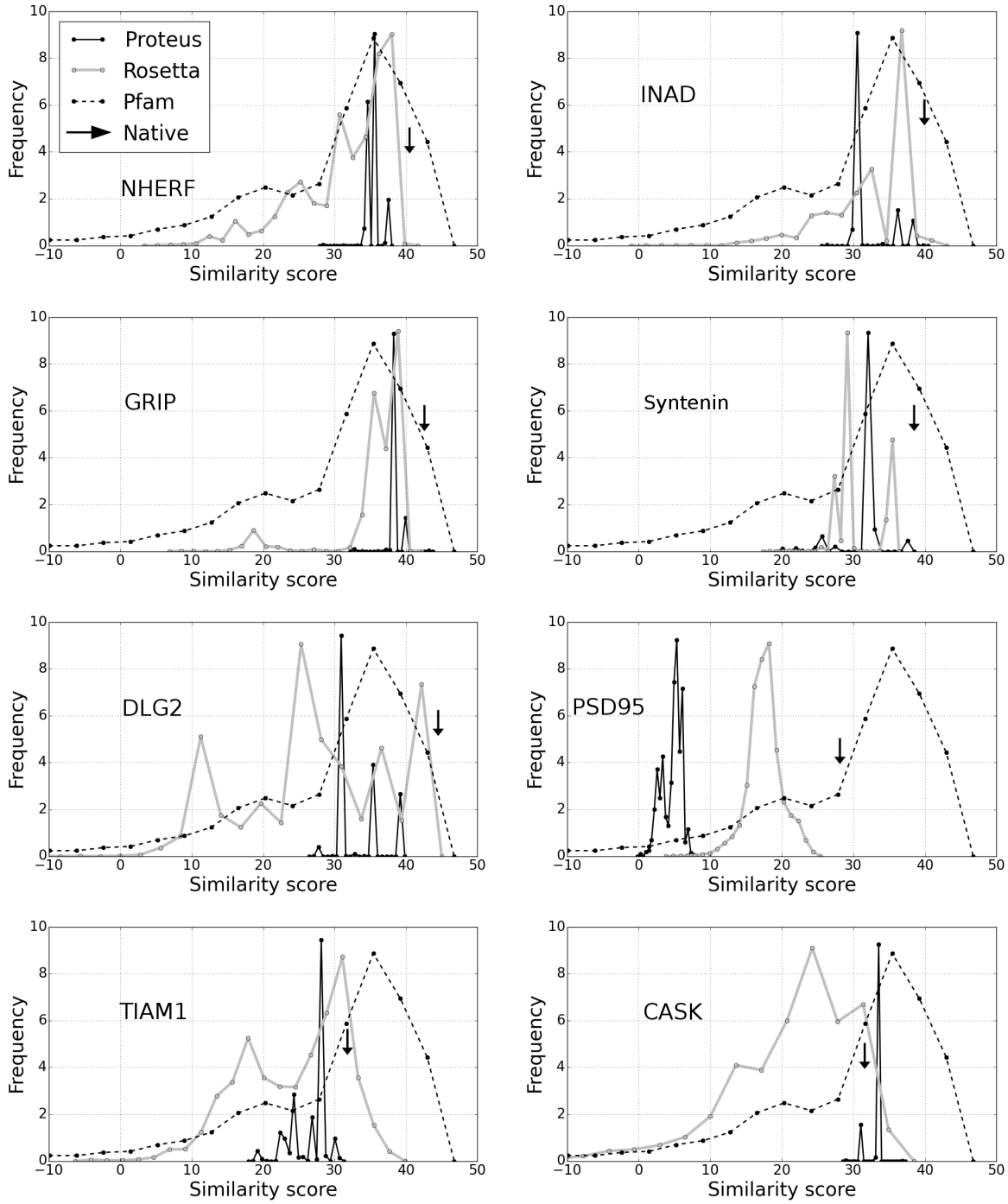

Figure 5: Similarity scores vs. Pfam for hydrophobic core.

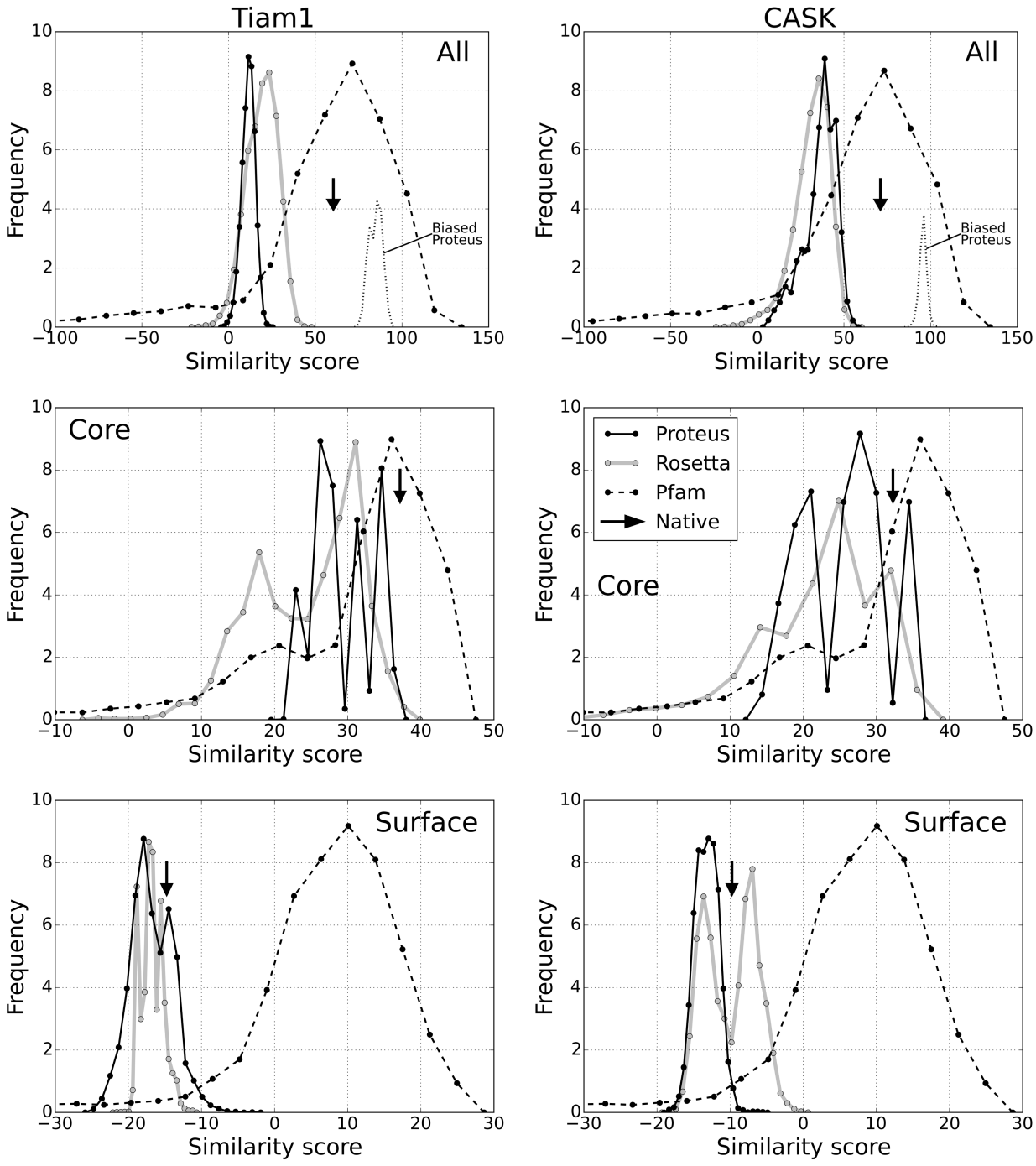

Figure 6: Similarity scores vs. Pfam for Tiam1 and Cask.

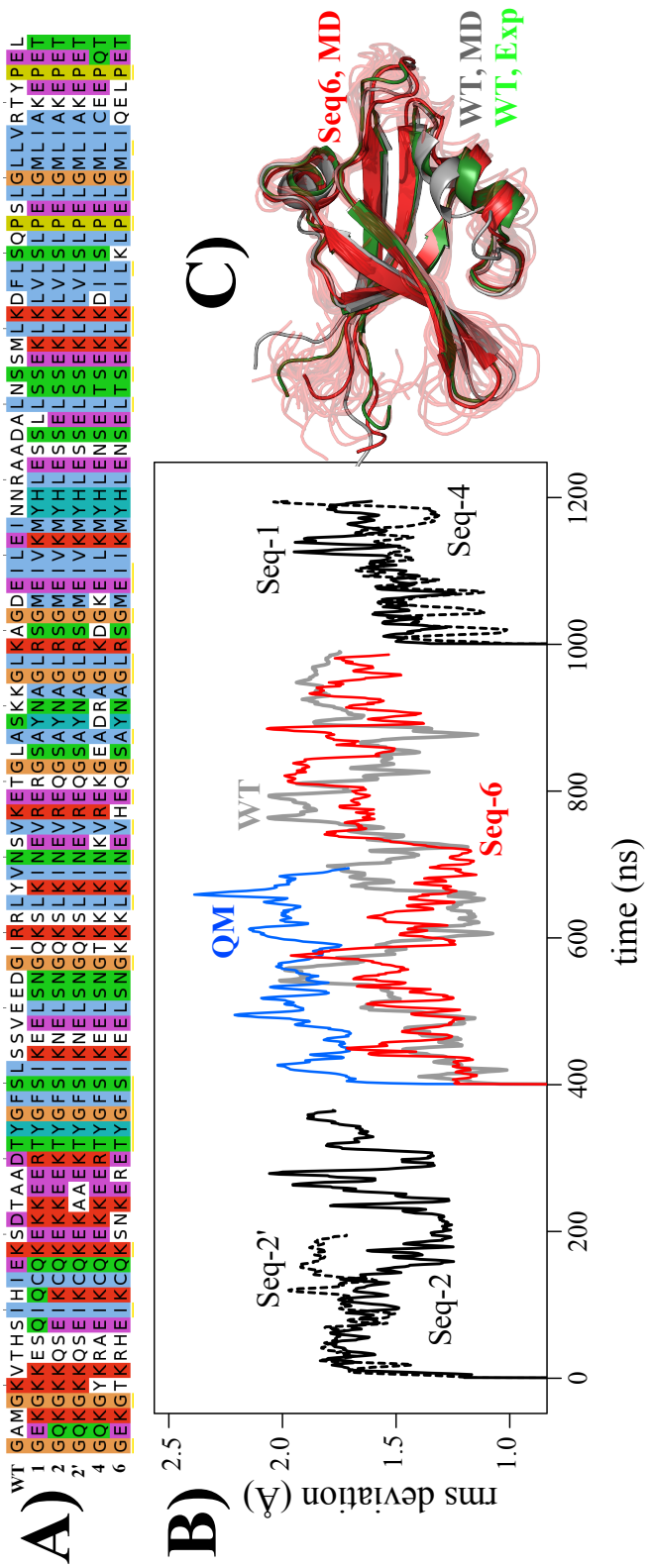

Figure 7: MD simulations of selected Tiam1 designs.

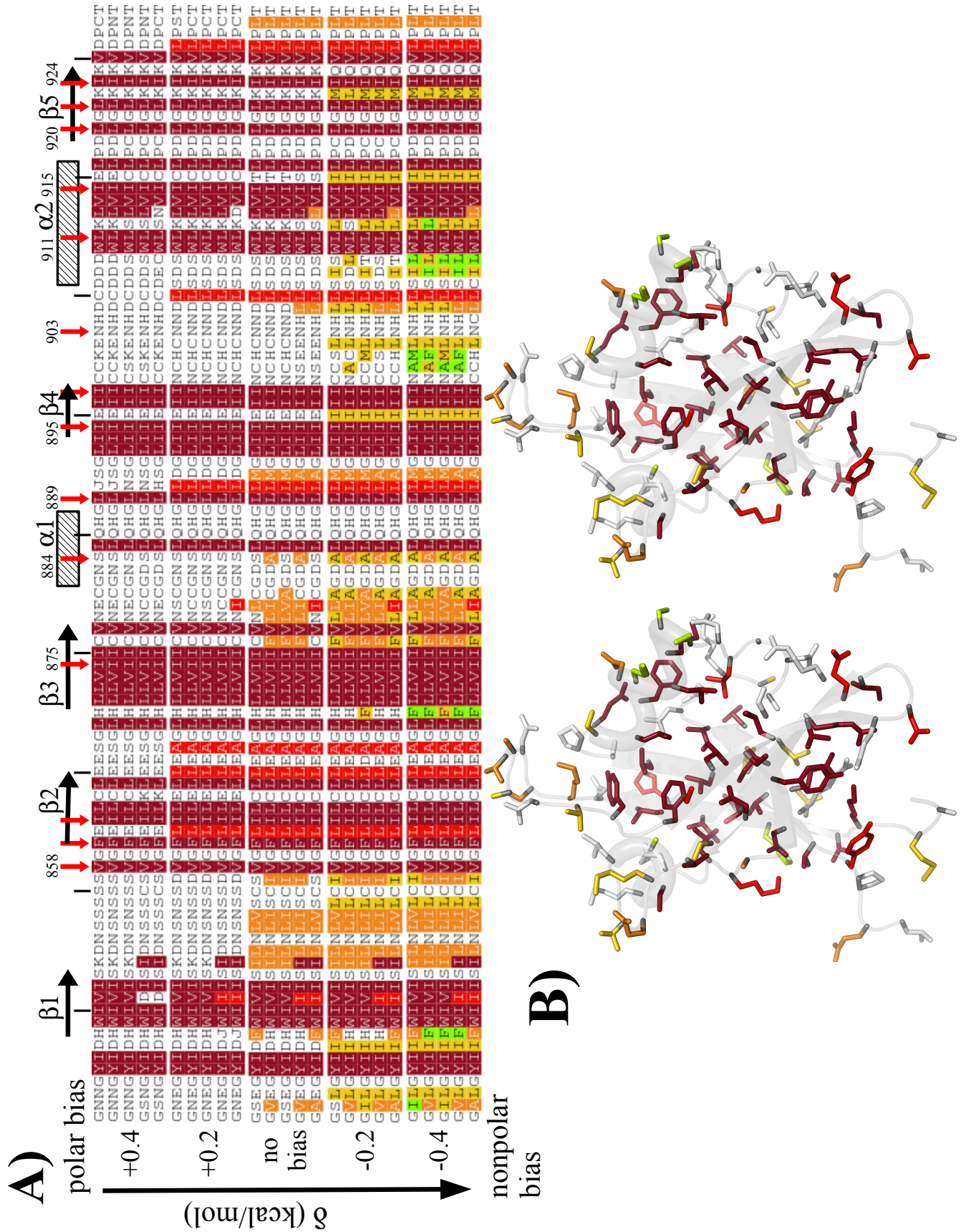

Figure 8: Hydrophobic titration of Tiam1.

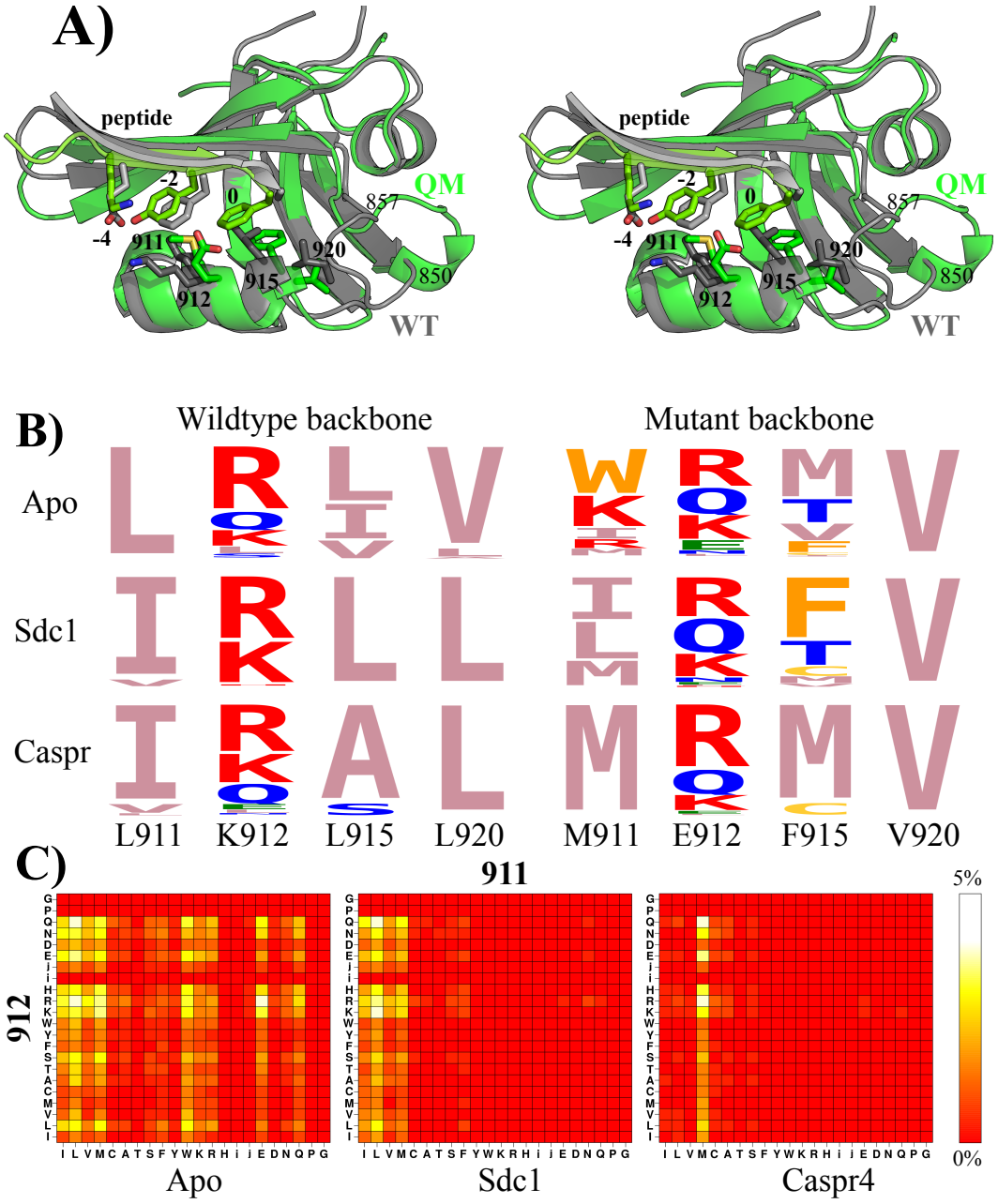

Figure 9: Designing four Tiam1 specificity positions.