# Computational design of the Tiam1 PDZ domain and its ligand binding

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 direct protein-protein interactions and serve as models for protein design. Here, we optimized a protein design energy function for the Tiam1 and Cask PDZ domains that combines a molecular mechanics energy, Generalized Born solvent, and an empirical unfolded state model. Designed sequences were recognized as PDZ domains by the Superfamily fold recognition tool and had similarity scores comparable to natural PDZ sequences. The optimized model was used to redesign two PDZ domains, by gradually varying the chemical potential of hydrophobic amino acids; the tendency of each position to lose or gain a hydrophobic character represents a novel hydrophobicity index. We also redesigned four positions in the Tiam1 PDZ domain involved in peptide binding specificity. The calculated affinity differences between designed variants reproduced experimental data and suggest substitutions with altered specificities.

**Keywords:** protein-protein interactions, molecular modelling, Monte Carlo simulation, Proteus software

# 1 Introduction

PDZ domains ("Postsynaptic density-95/Discs large/Zonula occludens-1") are small, globular protein domains that 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 interaction with target proteins have been extensively studied and computationally engineered. 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].

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

Here, we considered a simple but important class of CPD models. The energy is a physics-based function of the "MMGBSA" type, which combines a molecular mechanics protein energy with a Generalized Born + surface area implicit solvent. 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 $\epsilon_P$, a small set of atomic surface energy coefficients $\sigma_i$, 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, while each reference energy represents the contribution of a single amino acid of type $t$ to the unfolded state energy. The model is implemented in the Proteus software[26–28].

The present physics-based energy function can be compared to more empirical ones,

of which the most successful is the Rosetta energy function[29–31]. The Rosetta function includes a Lennard-Jones repulsion term, a Coulomb term, a hydrogen-bonding term, a Lazaridis-Karplus solvation term[32], and unfolded state reference energies. It has a large number of parameters specifically optimized for CPD, which provide optimal performance, but less transferability and a less transparent physical interpretation. Proteus also provides some specific functionalities, such as Replica Exchange Monte Carlo, various importance sampling methods, and the ability to compute free energies that are formally exact[33–35].

We optimized the reference energies $E_t^r$ for the Tiam1 and Cask PDZ proteins, using a maximum likelihood formalism. We compared two values of the protein dielectric constant, $\epsilon_P = 4$ and 8. These values gave good results in a systematic study that compared dielectric constants in the range 1–32[36]. The performance of the model was tested by generating designed sequences for both proteins and comparing them to natural sequences, as well as sequences generated with the Rosetta energy function and software[37]. The sequence design was performed by running long Monte Carlo simulations where all protein positions except Gly and Pro were allowed to mutate freely, leading to thousands of designed protein variants. The testing included cross-validation, where the reference energies were optimized using one set of PDZ domains, then applied to others. We also performed 100–1000 nanosecond molecular dynamics (MD) simulations for a few of the sequences designed with our optimized CPD model, to help assess their stability. Ten sequences were stable over 100 ns or more and one over 1000 ns of MD simulation.

We then applied the CPD model with optimized parameters to two problems, which are representative of the two main areas we are interested in: exploring the plasticity of sequence space for PDZ domains and designing strong and specific PDZ ligands. Earlier applications in these areas mostly employed empirical, knowledge-based energy functions such as the Rosetta function[9,10,13,14]. First, we performed a series of Monte Carlo simulations of two PDZ domains where the chemical potential of the hydrophobic amino acid types was gradually increased, artificially biasing the protein composition. As the hydrophobic bias was increased, hydrophobic amino acids gradually invaded the protein from the inside out, forming a hydrophobic core that became 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, which provides information

on the designability or plasticity of the protein core. The second application consisted in designing four Tiam1 positions known to be involved in specific target recognition. These four positions were varied through Monte Carlo simulations of either the apo-protein or the protein in complex with two distinct peptide ligands. The simulations were in agreement with experimental sequences and binding affinities, and suggest new variants that could have altered specificities. This application is a step towards the design of strong peptide binders, which could be of use as reagents or inhibitors *in vitro* or *in vivo*.

# 2 The unfolded state model

## 2.1 Maximum likelihood reference energies

The Monte Carlo method employed here generates a Markov chain of states[38,39], such that the states are populated according to a Boltzmann distribution. The energy employed is not the folded protein's energy, but rather its *folding energy*, i.e., the difference between its folded and unfolded state energies[33]. 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. We consider the same mutation in the unfolded state. 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 folding energy change due to a mutation thus 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 $\Delta E^f$ and $\Delta E^u$ 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 $\Delta G_S = G_S^f - E_S^u$ is the folding free energy of $S$, $G_S^f$ 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

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

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 $E_t^r$; we refer to $\mathcal{L}$ as their likelihood[40]. We have

$$kT \ln \mathcal{L} = \sum_S \sum_{t \in \mathrm{aa}} n_S(t) E_t^r - \sum_S G_S^f - NkT \ln Z = \sum_{t \in \mathrm{aa}} N(t) E_t^r - \sum_S G_S^f - NkT \ln Z, \tag{5}$$

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. As usual in statistical mechanics[41], 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

We will use two methods to approach the maximum likelihood $\{E_t^r\}$ values, starting from a current guess $\{E_t^r(n)\}$. With the first method, we step along the gradient of $\ln \mathcal{L}$, using the update rule[40]:

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

Here, $\alpha$ 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 $\delta E$ 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:

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

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 Eq. (13) represent the protein internal energy. They were taken from the Amber ff99SB empirical energy function[42], slightly modified for CPD. The original backbone charges were 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[43]. The last term on the right of Eq. (13), $E_{\text{solv}}$, represents the contribution of solvent. We used a "Generalized Born + Surface Area", or GBSA implicit solvent model[44]:

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

Here, $\epsilon_W$ and $\epsilon_P$ 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 $i$[44,45]. $A_i$ is the exposed solvent accessible surface area of atom $i$; $\sigma_i$ is a parameter that reflects each atom's preference to be exposed or hidden from solvent. The solute atoms were divided into four groups with specific $\sigma_i$ values. The values were -5 (nonpolar), -40 (aromatic), -80 (polar), and -100 (ionic) cal/mol/Å$^2$. Hydrogen atoms were assigned a surface coefficient of 0. Surface areas were computed by the Lee and Richards algorithm[46], implemented in the XPLOR program[47], using a 1.5 Å probe radius. The MC simulations used a protein dielectric of $\epsilon_P = 4$ or 8.

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[44], with model parameters optimized for use with the Amber force field[45]. Since $b_i$ depends on the coordinates of all the solute atoms[44], an additional approximation is needed to make the GB energy term pairwise additive and to define the energy matrix[27,48]. 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[27,48].

The surface energy contribution $E_{\text{surf}}$ 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, we used the method of Street et al.[49]. In this method, 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 the buried surface area, a scaling factor is applied to the contact areas involving buried sidechains. Previous studied showed that a scaling factor of 0.65 works well[45,48].

## 3.2 Reference energies in the unfolded state

In the CPD model, the unfolded state energy depends on the sequence composition through a set of reference energies $E_t^r$ (Eq. 1). Here, the reference energies were 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 were two distinct $E_t^r$ values: a buried and an exposed value. This is so even though the reference energies are used to represent the unfolded, not the folded state. This procedure is supported by three assumptions. 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 contributes indirectly to the reference energies. 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 $\delta E_t^r$ 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. Specifically, we ran MC simulations of an extended peptide (the Syndecan1 peptide; see below) and computed the average energies for each amino acid type at each peptide position (excluding the termini). We took the differences between amino acid types and averaged them over the peptide positions. During likelihood maximization, $E_c^r$ is optimized while $\delta E_t^r$ is held fixed. To optimize the $E_c^r$ values, we apply the linear or logarithmic method while 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 and structural models

We considered the Tiam1 and Cask PDZ domains, whose crystal structures are known (PDB codes 4GVD and 1KWA, respectively). They both belong to the class II binding motif[3], which recognizes the pattern Φ-X-Φ at the C-terminus of its peptide ligand, where Φ is a hydrophobic amino acid. To define the target amino acid frequencies for likelihood maximization, we collected homologous sequences for each PDZ domain. We identified homologous sequences by using the Blast tool to search the Uniprot database, with the sequences taken from the PDB file as the query and the Blosum62 scoring matrix. We retained homologs with a sequence identity, relative to the query, above a 60% threshold and below an 85% threshold. If two homologs had a mutual sequence identity above 95%, one of the two was viewed as redundant and was discarded. This led to 50 Tiam1

and 126 Cask homolog sequences. The two sets of homologs are referred to as $\mathcal{H}_T$ and $\mathcal{H}_C$, respectively. For each of the sets, say H, we average over all homologs and all positions to obtain compute the overall amino acid frequencies. The averaging is done separately for buried and exposed positions. The resulting amino acid frequencies are denoted $\{f_t^b(\mathcal{H}), f_t^e(\mathcal{H})\}$, where the subscript $t$ represents an amino acid type and the superscripts $b$, $e$ refer to buried and exposed positions, respectively. Finally, the sets of mean frequencies derived from $\mathcal{H}_T$ and $\mathcal{H}_C$ were themselves averaged, giving the overall target amino acid frequencies, $f_t^b = \left(f_t^b(\mathcal{H}_T) + f_t^b(\mathcal{H}_C)\right)/2$ for each type $t$, and similarly for the exposed positions. Distinct target frequencies were thus obtained for buried and exposed positions.

Model parameterization and testing were mostly done for the apo state of each protein. However, for Cask, no apo X-ray structure was available at the beginning of this work, so a holo-like structure was used, where the peptide binding site is occupied by the C-terminus of another PDZ domain in the crystal lattice; the apo state was then modelled by removing this peptide. For the PDZ domain of Tiam1, we also used 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; therefore, we expect the CPD model to be transferable between apo/holo Tiam1 states. For additional testing, we also considered two class I PDZ domains, syntenin and DLG2 (second PDZ domain in both cases), which recognize the pattern S/T-X-Φ at the C-terminus of its peptide ligand. Their X-ray structures are 1R6J and 2BYG, respectively. In both these structures, the peptide ligand was not co-crystallized, but the peptide binding site of each PDZ domain was partly occupied by the C-terminus of another protein molecule in the crystal lattice. The structures employed are listed in Table 1.

To carry out the Monte Carlo design calculations, the structures were prepared and energy matrices computed using procedures described previously[15,50]. Two missing segments in the Tiam1 PDZ domain (residues 851-854 and 868-869) were built using the Modeller program[51]. The peptide ligand was removed from the PDB structure for most of the design calculations before computing the energy matrix. For each pair of amino acid side chains, the interaction energy was computed after 15 steps of energy minimization, with the backbone held fixed and only the interactions of the pair with each other and the backbone included[26]. This short minimization alleviates the discrete rotamer

approximation. Side chain rotamers were described by a slightly expanded version of the library of Tuffery et al[52], which has a total of 254 rotamers (summed over all amino acid types). This expanded library includes additional hydrogen orientations for OH and SH groups[48]. This rotamer library was chosen for its simplicity and because it gave very good performance in sidechain placement tests, comparable to the specialized Scwrl4 program (which uses a much larger library)[53,54].

## 3.4 Monte Carlo simulations

Sequence design was performed with Proteus, which runs long Monte Carlo (MC) simulations where selected amino acid positions can mutate freely. The choice of mutating positions is user-defined and depends on the specific design challenge. Four different choices occurred in the present work. First, to optimize the reference energies, we did simulations where about half of the positions could mutate at a time. Second, the optimized models were tested in simulations where all positions except Gly and Pro were free to mutate. Hydrophobic titration of two PDZ domains also employed this choice. Third, to produce designed sequences to test through molecular dynamics, we did MC simulations where Gly, Pro, and 11 positions closely involved in peptide binding were held fixed, while all other positions were allowed to mutate. Fourth, in the second Tiam1 application, only four positions in the protein could mutate. In all these cases (with two exceptions), mutations occurred randomly, subject only to the MMGBSA energy function that drives the simulation. In only two cases, an additional, "experimental" energy term was used to explicitly bias the simulation to stay close to the natural, Pfam sequences.

The Monte Carlo simulations used one- and two-position moves, where either rotamers, amino acid types, or both changed. For two-position moves, the second position was selected among those that had a significant interaction energy with the first (i.e., there was at least one rotamer conformation where their unsigned interaction energy was 10 kcal/mol or more). In addition, sampling was enhanced by Replica Exchange Monte Carlo (REMC), where several MC simulations ("replicas" or "walkers") were run in parallel, at different temperatures. Periodic swaps were attempted between the conformations of two walkers $i$, $j$ (adjacent in temperature). The swap was accepted with the probability

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

where $\beta_i$, $\beta_j$ are the inverse temperatures of the two walkers and $\Delta E_i$, $\Delta E_j$ are the changes in their folding energies due to the conformation change[55,56]. We used eight walkers, with thermal energies $kT_i$ that range from 0.125 to 3 kcal/mol, spaced in a geometric progression: $T_{i+1}/T_i$ = constant[55]. Simulations were done with the proteus program (which is part of the Proteus package)[27]. REMC was implemented with an efficient, shared-memory, OpenMP parallelization[33].

One simulation of Tiam1 and one of Case were done that included an "experimental", biasing energy term, which penalized sequences that had a low similarity to a reference, experimental set. The bias energy had the form

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

where the sum extends 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; $S_i^{\text{rand}}$ is the mean score (versus the same Pfam column) for a random type (where all types are equiprobable), and $c = 0.5$ kcal/mol.

## 3.5 Rosetta sequence generation

Monte Carlo simulations were also performed using the Rosetta program and energy function[37]. 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 ex1 and ex2 options activate an enhanced rotamer search for buried sidechains, the last option (linmem_ig) corresponds to on-the-fly energy calculation, and default parameters were used otherwise. Gly and Pro residues present in the wildtype protein were not allowed to mutate, and positions that do mutate could not change into Gly or Pro (as with the Proteus design simulations). Simulations were run for each PDZ domain until 10,000 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 was comparable to the cost of the Proteus calculations (energy matrix plus Monte Carlo simulations).

### 3.6 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. This matrix is appropriate for comparing rather distant homologs (CPD and Pfam sequences in this case). Each Pfam sequence was also compared to the Pfam alignment, which allowed for comparison between the designed sequences and a typical pair of natural PDZ domains. For these Pfam/Pfam comparisons, if a test PDZ domain 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, consisting of 12,255 sequences. Similarities were computed separately for the 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[57,58], which attempts to classify sequences according to the Structural Classification Of Proteins, or SCOP[59]. 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. The hmmscan program was executed using an E-value threshold of $10^{-10}$ and a total of 15,438 models to represent the SCOP database.

To compare the diversity in the designed sequences with the diversity in natural sequences, we used the standard, position-dependent sequence entropy[60], 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 employed 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[61], and by analyzing residue-residue contact energies in proteins[62]. To obtain a sense for how many amino acid types appeared at a typical position, we report the residue entropy in its exponential form, $\exp(S)$ (which ranges from 1 to 6), averaged over the protein chain.

## 3.7 Protein:peptide binding free energies

For the Tiam1 PDZ domain, we used design calculations in the presence and absence of a bound peptide to obtain estimates of the binding free energy differences between protein variants. If a given sequence $S$ was sampled in both the apo and holo states, we computed the mean energy $\langle E_{holo}(S)\rangle$, $\langle E_{apo}(S)\rangle$ in each of the two states by averaging over the sampled conformations. Then, we took the difference

$$\Delta\Delta E(S, S') = (\langle E_{holo}(S')\rangle - \langle E_{apo}(S')\rangle) - (\langle E_{holo}(S)\rangle - \langle E_{apo}(S)\rangle) \tag{19}$$

as our estimate of the binding free energy difference between the variants $S$ and $S'$. We also computed binding free energy differences between *groups* of homologous sequences, say $\mathcal{S}$ and $\mathcal{S}'$, by pooling the homologous sequences sampled in either the apo or holo state, then averaging over the conformations sampled and taking the energy difference $\Delta\Delta E(\mathcal{S}, \mathcal{S}')$.

## 3.8 Molecular dynamics simulations

Wildtype and a quadruple mutant Tiam1 and ten sequences designed with Proteus were subjected to MD simulations with explicit solvent and no peptide ligand. The starting structures were taken from the MC trajectory or the crystal structure (wildtype protein and quadruple mutant: PDB codes 4GVD and 4NXQ) and slightly minimized with harmonic restraints to maintain the backbone geometry. The protein was immersed in a large box of non-overlapping waters. The solvated system was truncated to the shape of a truncated octahedral box using the Charmm graphical interface or GUI[63]. The minimum distance between protein atoms and the box was 15 Å and the final models included about 11,000 water molecules. A few sodium or chloride ions were included to ensure overall electroneutrality. The 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[64,65]. Long-range electrostatic interactions were treated with a Particle Mesh Ewald approach[66]. The Amber ff99SB force field and the TIP3P model[67] were used for the protein and water, respectively. Simulations were run for 100–1000 nanoseconds, depending on the sequence, using the Charmm and NAMD programs[68,69].

# 4 Results

## 4.1 Experimental structures and sequences

Three dimensional (3D) structures of the four test PDZ domains are shown in Fig. 1A. Fourteen core residues (identified visually) superimposed well between the structures, while loops and chain termini displayed large deviations. The Tiam1 $\alpha_2$ helix is rotated slightly outwards compared to the other three structures[70]. Fig. 1B illustrates the similarity between pairs of PDZ domains, as determined by the rms deviation between structurally-aligned $C_\alpha$ atoms and the pairwise sequence identities. The rms deviations are between 1.0 and 2.1 Å and the sequence identities between 17 and 33%. The Tiam1/Cask sequence identity is 33% and their structural deviation is 1.7 Å based on 42 aligned $C_\alpha$ atoms. The syntenin and DLG2 structures are more similar, with a structural deviation of 1.0 Å based on 60 aligned $C_\alpha$ atoms.

Sequence conservation within the four PDZ domains and a subset of the Pfam seed alignment is shown in Fig. 2. The 14 positions used to define the hydrophobic core are highly, although not totally conserved within the Pfam seed alignment. Arginine, Lys and Gln appear at some of the positions, since in small proteins like PDZ domains, the long hydrophobic portion of these side chains 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.

## 4.2 Optimizing the unfolded state model

We optimized the reference energies $E_t^r$ for Tiam1 and Cask, using their natural homologs to define the target amino acid frequencies. The protein dielectric constant $\epsilon_P$ was either 4 or 8, with $\epsilon_P$=8 giving the best results. The $E_t^r$ optimizations all converged to within 0.05 kcal/mol after about 20 iterations for most amino acid types, and to 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 2 indicates the final reference energies. The $E_t^r$ values were compared to, and agreed qualitatively with the energies computed from an extended peptide structure, which provides a less empirical model of the unfolded state. Table 3

compares the amino acid frequencies from the natural homologs and the simulations using parameters optimized with $\epsilon_P$=8. Results obtained using parameters optimized with $\epsilon_P$=4 are given in Supplementary Material. The theoretical population of the different amino acid classes agreed well with experiment, with rms deviations of about 1%, for both the exposed and buried positions. The agreement for the amino acid types was less good, with rms deviations of 3.9%/2.4% (buried/exposed positions). The intra-class frequency distributions depend explicitly on the energy offsets $\delta E_t^r$ defined within each class, which were computed with molecular mechanics (see Methods, Eq. 15).

## 4.3 Assessing designed sequence quality

**Family recognition tests** Proteus design simulations used Replica Exchange Monte Carlo with eight replicas and 750 million steps per replica, at thermal energies $kT$ that ranged from 0.125 to 3 kcal/mol. All positions (except Gly and Pro) were allowed to mutate freely into all amino acid 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 10,000 sequences with the lowest energies among those sampled by any of the MC replicas were retained for analysis, along with the 10,000 Rosetta sequences. These sequences were analyzed by the Superfamily fold recognition tool[58,71] (Table 4). With a protein dielectric constant of 8, we obtained a high percentage of sequences assigned to the correct family: 91% for Tiam1 and 100% for Cask, with E-values of around $10^{-3}$ for the family assignments. These values are similar to Rosetta (90 and 98% family recognition for Tiam1 and Cask). Changing the protein dielectric constant to $\epsilon_P$=4 gave somewhat poorer results for Tiam1, with 53% of the sequences designed with Proteus correctly recognized by Superfamily.

**Sequences and sequence diversity** Tiam1 and Cask sequences predicted by Proteus and Rosetta as well as natural sequences are shown in Fig. 3 for the fourteen core residues and in Fig. 4 for the sixteen surface residues (Tiam1 only). The sequences are represented as sequence logos. As seen in many previous CPD studies[30,72], agreement with experiment for the core residues is very good, while agreement for the surface residues is much poorer. The behavior of surface positions was also probed by designing each position individually,

with the rest of the protein free to explore rotamers but not mutations ("mono-position" design). The corresponding logo (Fig. 4) shows an excess of Arg and Lys residues, suggesting that the CPD reference energies are not yet fully optimal, despite the extensive empirical $E_t^r$ tuning. Sequence similarity scores are given in the next subsection.

The diversity of the natural and designed sequences was characterized by a mean, exponential sequence entropy (see Methods), which corresponds to a mean number of sampled sequence classes per position. For example, a value of 2 at a particular position indicates that amino acids from two of the six classes are present at that position within the set of analyzed sequences. An overall average value of two indicates that on average, two amino acid classes are present at any position within the analyzed sequences. For reference, the Pfam RP55 set of 12,255 natural sequences has a mean entropy of 3.4. Pooling the designed Tiam1 and Cask sequences gave an entropy of 2.2 with Rosetta and 2.0 with Proteus, indicating that these two backbone geometries cannot accomodate as much diversity as the much larger RP55 set. Taking the 10,000 lowest energy sequences sampled with the room temperature Monte Carlo replica (instead of the 10,000 lowest energies sampled collectively by all replicas at all temperatures) and pooling Tiam1 and Cask as before gave a higher overall entropy of 2.9 with Proteus. With Rosetta, entropy in the core was only slightly below the average over all positions. With Proteus, it was distinctly lower (1.25). For the Pfam-RP55 sequences, it was 1.8.

**Blosum similarity scores** Fig. 5 shows the computed Blosum40 similarity scores between designed and natural sequences. With Proteus, for both Tiam1 and Cask, the overall similarities overlapped with the bottom of the peak of the natural scores, and were comparable to the values for the Rosetta sequences. For the surface residues, shown separately, similarity to the natural sequences was low (scores below zero), both for Proteus and Rosetta. With a protein dielectric constant of 4, Proteus performed about as well as with $\epsilon_P$=8, giving almost the same similarity averaged over all Tiam1 and Cask positions, for example.

While the similarity scores vs. Pfam with Proteus were comparable to Rosetta (Fig. 5), the identity scores vs. the wildtype sequence were significantly higher with Rosetta. Identity scores excluding (respectively, including) Gly and Pro positions (which did not mutate) were 20% (28%) for Proteus vs. 26% (34%) for Rosetta. Evidently, for Tiam1

and Cask, Rosetta performed ≈5 fewer mutations than Proteus.

For certain applications, we may need to specifically explore a sequence space region very similar to Pfam, beyond the similarity provided by an MMGBSA energy. This can be achieved by adding to the energy an "experimental," or bias energy term that explicitly favors high sequence scores. Fig. 5 includes results that use such a biased energy term: by construction, it leads to very high similarity scores. A bias energy term could also be used to limit the total number of mutations.

**Cross-validation tests** As a first cross-validation test, we applied the reference energies optimized using Tiam1 and Cask homologs (with $\epsilon_P$=8) to two other PDZ domains: DLG2 and syntenin. The Superfamily scores were comparable to those obtained for Tiam1 and Cask, with 100% family recognition (Table 4). Sequences designed with Rosetta for DLG2 and syntenin also gave 100% family recognition. For further cross-validation, we optimized reference energies using an alternate set of PDZ domains: DLG2, syntenin, PSD95, GRIP, INAD, and NHERF. Target frequencies were defined by a small set of their natural homologs. We used $\epsilon_P$=8. To distinguish the new and initial model variants, we refer to the new variant as the $n$=6 model (it uses six PDZ domains for parameterization), and the initial model as the "T+C" model (it used Tiam1 and Cask). The new, $n$=6 reference energies were then used to produce designed Tiam1 and Cask sequences, which were subjected to Superfamily tests and similarity calculations. The Superfamily performance for Tiam1 was slightly degraded, compared to the previous, T+C model. The Tiam1 Superfamily score decreased from 90.6% to 76.6% for family recognition. The Cask score was unchanged. Histograms of Blosum similarity scores (Supplementary Material) show that the overall scores for Tiam1 and Cask with $n$=6 were very similar to the T+C model, while the scores for the core positions were actually shifted to higher, not lower values. For DLG2 and syntenin, we also computed similarity scores using both the initial, T+C parameterization and the new, $n$=6 parameterization. The similarity scores with the T+C model were slightly poorer than with the $n$=6 model, as expected. The overall score decreased by about 20 points for syntenin and about 10 points for DLG2 (Supplementary Material). Overall, the cross-validated models gave slightly degraded performance. Thus, for any PDZ domain of interest, it may be preferable to optimize reference energies specifically for that domain, rather than transferring values parameterized

using other PDZ domains.

## 4.4 Stability of designed sequences in molecular dynamics simulations

As another test of the design model, ten Tiam1 sequences designed with Proteus were subjected to molecular dynamics simulations (MD) using an explicit solvent environment. These sequences were obtained using Proteus with either $\epsilon_P$=8 or the less polarizable value $\epsilon_P$=4. Although no peptide ligand was present during the design simulations, 11 positions in the binding pocket 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 2,500 lowest energy designed sequences, we narrowed down the choice of sequences using the following four criteria: (a) sequences should have a nonneutral isoelectric point, (b) they should be assigned to the correct SCOP family by Superfamily with good E-values, (c) they should have good Pfam similarity scores, and (d) they should have at most 15 mutations that drastically change the amino acid type compared to the wildtype protein (such a change is defined by a Blosum62 similarity score between the two amino acid types of -2 or less). Applying these criteria reduced the number of sequences to 66 from the $\epsilon_P$=8 model and 45 from the $\epsilon_P$=4 model. In addition, we eliminated sequences that had two mutations that created a buried cavity and those that had net protein charges of +6 or more (which could lead to protein instability). A total of six sequences were chosen for further analysis. We refer to them as sequences 1–6 or seq-1, ..., seq-6. Sequences 1, 2, 4, and 5 were modified further manually to eliminate charged residues in the exposed loop 852–856 (lysines were changed manually to alanine), giving sequences 1', 2', 4', and 5'. The ten sequences are shown in Fig. 6A. Using these sequences 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 $10^{-10}$).

All ten sequences were subjected to MD simulations with explicit solvent. Initial simulations were run for 100 ns, with all ten sequences exhibiting good stability. Six were extended to lengths of 500 or 1000 ns. The wildtype protein (WT) was also simulated

for 1000 ns. The WT sequence appeared stable over the entire simulation, judging by its rms deviations from the wildtype (WT) X-ray structure and from its own mean MD structure (Fig. 6B). The mean MD structure had a backbone rms deviation of 1.0 Å from the WT X-ray structure (excluding 3–4 residues at each terminus and one very flexible loop, residues 850–857). During the MD trajectory, the rms deviation from the mean MD structure varied in the range 1–1.5 Å, without any visible drift (Fig. 6B). A weakly stable quadruple mutant (QM) with an unfolding free energy of just 1 kcal/mol[70] was also simulated for 1000 ns. The mean MD structure of QM had a backbone rms deviation of 1.6 Å relative to the QM X-ray structure (4NXQ). Note that the X-ray structure includes a peptide ligand, whereas the MD simulation represents the apo state. The average MD structure of QM (Fig. 6C) exhibited some unwinding of the N-terminus of the $\alpha_2$ helix. During the QM simulation, the structure had deviations from its average MD structure in the 0.8–1.2 Å range (Fig. 6B) and appeared stable.

Sequences 1, 3, 4, and 5 were simulated for 500 ns; sequence 3 moved away from the mean MD structure towards the end of the simulation; the other three sequences appeared stable (Fig. 6B). Sequence 2 (or seq-2) appeared stable up to almost 1000 ns (Fig. 6B). The mean seq-2 structure exhibited a shortening of the $\beta$ strands 2 and 3. Note that in the holo state, strand 2 makes direct contacts with the peptide ligand. The rms deviations between the average MD structure of seq-2 and the WT and QM X-ray structures were 1.5 and 1.6 Å, respectively. During the seq-2 MD simulation, the rms deviation of seq-2 from its average MD structure varied in the range 1.3–2 Å up to almost 1000 ns. At this point, just before 1000 ns, seq-2 underwent a large fluctuation. Extending the simulation for another 100 ns, the fluctuation largely regressed. More data are needed to determine if this fluctuation signals instability of this designed sequence.

Sequence 6 appeared stable throughout the microsecond MD simulation (Fig. 6B). The mean MD structure had a backbone rms deviation from the WT X-ray structure of just 1.0 Å, the same deviation as the mean WT MD structure. The mean MD structures of seq-6 and WT are superimposed in Fig. 6C and are very similar to each other, with a 1.2 Å backbone deviation between them. During the seq-6 MD trajectory, the deviations of seq-6 away from its mean MD structure fluctuated between about 1 and 1.5 Å, without any visible drift over the microsecond MD simulation.

### 4.5 Application: growing the PDZ hydrophobic core

As a first application of our optimized models, we examined the designability of the Tiam1 and Cask hydrophobic cores. Each PDZ domain was subjected to Replica Exchange Monte Carlo simulations with a succession of biased energy functions that increasingly favored hydrophobic residues. The first simulation included a bias energy term $\delta = 0.4$ kcal/mol (per position) that *penalized* hydrophobic amino acid types (ILMVAWFY). The final simulation included a bias energy term $\delta$ = -0.4 kcal/mol (per position) that *favored* hydrophobic types. Intermediate bias energy values $\delta$ = 0.2, 0, and -0.2 kcal/mol were also simulated. By gradually decreasing the bias energy value $\delta$, we effectively "titrate in" hydrophobic residues.

The results for Tiam1 are illustrated in Fig. 7. At the largest $\delta$ value, the Tiam1 hydrophobic core was depleted, with 10 amino acid positions (out of 94) changed to polar types. The changed positions mostly lie on the outer edge of the core. At the intermediate $\delta$ values, the hydrophobic core remained native-like. At the most negative $\delta$ value, the hydrophobic core became larger, expanding out towards surface regions, with 14 polar positions changed to hydrophobic types. Thus, the numbers of positions changed were approximately symmetric (around $\pm$12 changes), reflecting the bias. About 2/3 of the changes were in secondary structure elements. Overall, the observed propensities of each position to become polar or hydrophobic in the presence of a large or small penalty bias energy $\delta$ can be thought of as a hydrophobic designability index. Here, 11 of the 14 conserved core positions (all but positions 884, 898 and 903) remained hydrophobic even at the highest level of polar bias, along with 13 other positions, indicating that these positions have the highest hydrophobic propensity. Furthermore, 14 positions changed from polar to hydrophobic at the highest bias level, indicating that these positions also have a certain hydrophobic propensity. Results for Cask were similar, with 11 positions changed to polar at the highest polar bias and 9 changed to hydrophobic at the highest hydrophobic bias.

We also derived a parameter to describe the relative number of amino acid type changes per unit bias energy. This parameter was defined as the number $\delta N$ of residue positions changed from nonpolar to polar, divided by the product of the change $\delta E$ in bias energy and the mean number $N$ of nonpolar positions at zero bias. We call it the

hydrophobic susceptibility, $\chi_h$. For the Tiam1 PDZ domain, this calculation amounts to $\chi_h = \frac{1}{N}\frac{\delta N}{\delta E} = 0.88$ changes (per position) per kcal/mol. For Cask, the susceptibility was $\chi_h = 0.71$ changes per kcal/mol.

### 4.6 Application to Tiam1: designing specificity positions

As a second application, we redesigned four amino acid positions in the Tiam1 PDZ domain known to contribute to peptide binding specificity. Modifying these four positions (quadruple mutant or QM) in the protein altered the binding specificity such that QM preferentially bound a Caspr4 peptide relative to a syndecan-1 (Sdc1) peptide[70,73]. The four mutations in the QM PDZ domain were L911M, K912E, L915F, and L920V. All four positions but Lys912 are part of the conserved hydrophobic core. The four single and two double mutants were also characterized experimentally[70,73]. For simplicity, we denote the native (WT) sequence as LKLL and the quadruple mutant (QM) as MEFV. Other variants are denoted similarly. Replica Exchange MC simulations were conducted on several structural templates, where all four positions could mutate simultaneously, into all amino acid types except Gly and Pro. We used the Proteus model with the lower dielectric constant, $\epsilon_P$=4, which gave similarity scores equivalent to the $\epsilon_P$=8 model but had a reduced tendency to bury polar sidechains, thanks to its lower dielectric constant. In addition, no bias energy term was used, only the MMGBSA energy function. The CPD model used either the wildtype or the quadruple mutant crystal structure as backbone template for the design (PDB codes 4GVD and 4NXQ, respectively), shown in Fig. 8. Although these two structures were determined with the Sdc1 and Caspr4 ligands, they were used here for both holo *and* apo design simulations. The backbone rms deviation between these structures is 0.9 Å, with the main differences in the flexible 850–857 loop near the peptide C-terminus and in helix $\alpha_2$. This helix is pushed slightly outwards in the mutant complex, to accomodate Phe side chains both at protein position 915 and at the peptide C-terminus. One expectation is 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.

We studied six systems: the Tiam1 PDZ domain with either its wildtype or QM backbone X-ray structure, with the syndecan1 or the Caspr4 peptide ligand or no ligand.

Results are shown in Fig. 8. For all six systems, the native or native-like amino acid types were sampled at all four designed positions. For example, using the wildtype backbone structure (4GVD), Leu911 was preserved in the apo simulations and changed to Ile or Val in the holo simulations. Similarly, holo simulations with the mutant backbone structure (4NXQ) sampled Ile, Leu and Met. With the mutant backbone, holo simulations sampled somewhat different types at position 911 (Trp, Arg, Lys), which all appear in low amounts at the corresponding position in the Pfam seed alignment. All the simulations sampled 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 was better with the wildtype X-ray structure, while agreement with the mutant sequence was better with the mutant X-ray structure.

Recovery of the precise native and quadruple mutant sequences in the different states (apo and holo) was qualitative. Thus, using the wildtype backbone structure and in the apo state, the MC simulation recovered the wildtype sequence LKLL just 2 kcal/mol above the lowest energy 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 did any close homologs, probably because the wildtype backbone structure cannot accomodate Phe at position 915. Similar results were obtained with the Sdc1 ligand, with the LKLL, IKLL, VKLL, and MKLL sequences all having energies just 1–2 kcal/mol above the best sequence. The MKLL:Sdc1 affinity is known experimentally, and is within 0.1 kcal/mol of the wildtype value[73]. Experimentally, the wildtype and mutant sequences have the same binding free energy for Caspr4, and stabilities just 2 kcal/mol apart, suggesting that both should be sampled. Instead, neither was sampled. The closest homolog sequence was IEAV (similar to MEFV), at +2 kcal/mol (relative to the best sequence). This was 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 recovered the mutant sequence MEFV at an energy of +5 kcal/mol (relative to the best sequence) and the wildtype sequence LKLL at +7 kcal/mol. Both protein variants are thermodynamically stable; a slightly higher energy to produce LKLL seems reasonable, since the design simulation used the mutant backbone structure, which presumably should favor MEFV. With the Sdc1 ligand, MEFV appeared at

an energy of +6 kcal/mol, relative to the best sequence, which was the close wildtype homolog IKLV. VKVL was just 3 kcal/mol higher. With the Caspr4 ligand, the mutant sequence appeared at an energy of +7 kcal/mol, compared to the best sequence, TKMV. Its homologs MQMV and MEMV appeared at +5 kcal/mol. The wildtype LKLL and its close homologs did not appear (indicating poorer energies), while MAFI was the second best sequence overall.

A more detailed comparison is possible with the binding affinities of the experimentally characterized mutants[73]. 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 Leu915 to Phe affects binding differently depending on the residue type at position 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 gave information on correlations between the four mutating positions. Fig. 8C shows covariance plots between positions 911 and 912 for the apo

and holo simulations. Position 911 was more diverse in the apo than in either holo state (Sdc1 or Caspr4), while 912 was not very sensitive to the peptide. The computed pairwise correlations between all four protein positions were weak, so that the covariance plots mostly exhibit horizontal and vertical lines or bands, without noticeable "diagonal" features. This agrees with the experimental affinities of the single, double, and quadruple mutants, where the affinity changes associated with each point mutation were only weakly coupled to the other positions[73].

# 5 Discussion

## 5.1 Model limitations

We have parameterized a simple CPD model for PDZ design, suitable for high-throughput design applications and 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 $\sigma_i$ and two protein dielectric constants $\epsilon_P$, quantities that characterize the solvent model and are the main empirical parameters in the folded state energy function. We used a specific set of X-ray structures as our test proteins, each with a specific backbone conformation. For the sidechains, we used a simple, discrete rotamer library and a short minimization of each sidechain pair during the energy matrix calculation to alleviate the discrete rotamer approximation. Both the energy function and the rotamer description have been extensively tested and shown to give very good performance for sidechain reconstruction tests[54] (comparable to the popular Scwrl4 program[53]) and good performance for a large set of protein acid/base constants[74] (superior to the Rosetta software[75], despite extensive *ad hoc* parameter tuning).

The unfolded state representation used a simple, implicit model, characterized by a set of empirical, amino acid chemical potentials or reference energies. These energies were 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 used here is more refined than previously[76], since distinct reference energy values were used for amino acid positions that are buried or exposed in the *folded* state.

This method assumes 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. In principle, this doubles the number of adjustable reference energies. However, we reduced this number by introducing amino acid similarity classes, with one adjustable reference energy per class. To optimize the reference energies, we performed 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 that used over a half billion steps per simulation (per replica), and produced thousands of designed sequences in a single simulation. Reference energy values were optimized with two different choices for the protein dielectric constant $\epsilon_P$. The performance levels were similar with both values.

The model has several limitations, most of which are widespread in CPD implementations and applications. The first is the use of protein stability as the sole design criterion, without explicitly accounting for fold specificity[77], protection against aggregation, or functional considerations like ligand binding. We note, however, that the Superfamily tests did not lead to any fold misassignments (sequences that prefer another SCOP fold), so that in practice, fold specificity was achieved. Functional criteria can also be introduced in an *ad hoc* way; for example, the sequences tested by MD were designed with 11 peptide binding residues fixed, to facilitate future experimental studies.

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)[78]. 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[78], 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 backbone conformation of the protein and peptide. Specifically, with the wildtype backbone structure, there was no room to insert a Phe sidechain at position 915, even though Phe915 is present in the experimental quadruple mutant (which has a slightly different backbone structure). Therefore, the choice of the initial X-ray structural model is important, and several strategies are possible. Here, to parameterize the CPD model, we used X-ray structures solved with 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 could have been to parameterize the model using all apo structures, then switch to holo structures for the Tiam1 application.

For whole protein design (such as the hydrophobic titration application), the use of a fixed backbone can be partly counterbalanced by designing two or more PDZ structures. For example, pooling the designed Tiam1 and Cask sequences gave a mean sequence entropy comparable to the experimental Pfam set, and allowed us to recapitulate more sequences than design with just one backbone. In the application to 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 915. Notice that a new method for multibackbone design was recently developed in Proteus, based on a novel, non-heuristic hybrid Monte Carlo method that preserves Boltzmann sampling[35]. 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 choices that are partly arbitrary. These include the choice of a set of protein domains to represent the protein or family of interest. 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 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[26,76]. 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 (so-called "Native Environment approximation" or NEA[74,76]). This leads to an energy function that has the form of a sum over pairs of residues and that can be pre-calculated and stored in an energy matrix, which then serves as a lookup table during the subsequent Monte Carlo simulations. Despite this approximation, the model gave good results for a large acid/base benchmark, a problem that is very sensitive to the electrostatic treatment[74].

Some of these limitations could be removed in future work. In particular, since the 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 due to the recent implementation (manuscript in preparation) of a more exact Generalized Born calculation, whose efficiency is comparable to the pairwise approximation[79]. We have also implemented an improved model for hydrophobic solvation[80], which is faster and more accurate than our current surface area energy term (manuscript in preparation).

## 5.2 Model testing and applications

Designed sequences were extensively compared to natural sequences, through fold recognition tests, similarity calculations, and entropy calculations. In the test simulations, we designed the entire protein sequence, so that all positions (except Gly and Pro) could 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 had a high overall similarity to the natural, Pfam sequences, as measured by the Blosum40 similarity scores. The scores obtained were 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 was very strong for residues in the core of the protein, as observed in previous CPD studies[30,72]. In contrast, for residues at the protein surface, similarity scores were close to zero, the score one would obtain if one picked amino acid types randomly. Notice that many surface residues are involved in functional interactions, such as the eleven peptide-binding residues in PDZ domains. Surface residues are also selected by evolution to avoid aggregation or unwanted adhesion. Most of these functional constraints are not explicitly accounted for in our design protocol (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[15,50]. Evidently, the combined use of an improved protein force field, Generalized Born solvent, and family-specific reference energies leads to designed sequences that are more native-like and presumably better.

The Proteus sequences were also compared to sequences designed with the Rosetta software (using default parameters), which has itself been extensively tested. Based on the Blosum similarity scores (vs. natural sequences in Pfam) 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 6 percent points higher. This means that Proteus mutates about five 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 (see the "biased Proteus" results in the two upper panels of Fig. 5), 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 sim-

ulations. 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. Seven of the sequences remained stable over 200 ns simulation lengths and two were extended up to a microsecond, which represents a very stringent test of the designs. Sequence 6 was stable over the whole microsecond. Its mean deviation from the starting, experimental wildtype structure was 1 Å, which is the same as the mean deviation in the MD simulation of the wildtype sequence itself. The deviation between the mean sequence 6 MD structure and the mean wildtype MD structure was also small, just 1.2 Å. Sequence 2 was also simulated and remained stable until just before the end of the microsecond simulation, at which point it underwent a larger fluctuation. The fluctuation regressed 100 ns later. An even longer simulation would be needed to determine if this fluctuation is harmless or signals the beginning of domain unfolding. Note that in the presence of a peptide ligand, we expect the structural stability of the designed domains would increase further. MD simulations of additional designed sequences would also be of interest. Direct experimental testing of the designed proteins remains to be done.

The CPD model was used for two applications. "Hydrophobic titration" of two PDZ domains illustrated a novel way to characterize protein designability. The cost or availability of hydrophobic sidechain types was controlled by a bias energy term that was gradually varied. One result was that the mean overall hydrophobic "susceptibility", the number of type changes per kcal/mol and per position, differed by a factor of two between Tiam1 and Cask. 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. A comparison to other domain families would be of interest, and is left for future work.

Redesign of four specificity positions in Tiam1 allowed us to test the design model in a different way. It revealed some of the limitations of fixed backbone design, but also gave semi-quantitative agreement with available binding free energies. This agreement predicts new mutations that could be of interest for obtaining 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 the future, 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 allow different backbone structures for the apo and each holo system. This could be done in the future, since recent hybrid Monte Carlo schemes[35,81] 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[34]). 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 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. 77–97.

[29] Kuhlman B, Baker D. Native protein sequences are close to optimal for their structures. Proc Natl Acad Sci USA. 2000;97:10383–10388.

[30] 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.

[31] Rohl CA, Strauss CEM, S MKM, Baker D. Protein structure prediction using Rosetta. Methods Enzym. 2004;383:10383–1038866–93.

[32] Lazaridis T, Karplus M. Effective energy function for proteins in solution. Proteins. 1999;35:133–152.

[33] 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.

[34] 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.

[35] Druart K, Bigot J, Audit E, Simonson T. A hybrid Monte Carlo method for multibackbone protein design. J Chem Theory Comput. 2016;12:6035–6048.

[36] Gaillard T, Simonson T. Full protein sequence redesign with an MMGBSA energy function. J Comput Chem. 2017;submitted:0000.

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

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

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

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

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

[42] 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.

[43] 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.

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

[45] 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.

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

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

[48] Gaillard T, Simonson T. Pairwise decomposition of an MMGBSA energy function for computational protein design. J Comput Chem. 2014;35:1371–1387.

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

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

[51] 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.

[52] 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.

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

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

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

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

[57] 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.

[58] 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.

[59] 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.

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

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

[62] 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.

[63] 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.

[64] Nose S. A unified formulation of the constant temperature molecular dynamics method. J Chem Phys. 1984;81:511–519.

[65] Hoover WG. Canonical dynamics: equilibrium phase-space distributions. Phys Rev A. 1985;31:1695–1697.

[66] 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.

[67] 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.

[68] 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.

[69] 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.

[70] 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.

[71] 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.

[72] 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.

[73] 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.

[74] 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.

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

[76] 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.

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

[78] Simonson T. What is the dielectric constant of a protein when its backbone is fixed? J Chem Theory Comput. 2013;9:4603–4608.

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

[80] 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.

[81] 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.

Table 1: Test proteins

| protein name[a] | PDB code | residue numbers | # active positions[c] |
|---|---|---|---|
| syntenin(2) | 1R6J | 192-273 | 72 |
| DLG2(2) | 2BYG | 186-282 | 82 |
| Cask | 1KWA[b] | 487-568 | 74 |
| Tiam1 | 4GVD[b] | 837-930 | 84 |

[a]In parentheses: number of the PDZ domain within the protein. [b]Holo or holo-like structures. [c]The number of non-Gly, non-Pro positions, which can mutate during the design simulations.

Table 2: Unfolded state reference energies $E_t^r$ (kcal/mol)

| | | Design, $\epsilon_P$=8 | | Design, $\epsilon_P$=4 | |
|---|---|---|---|---|---|
| Residues | Peptide[a] | Buried | Exposed | Buried | Exposed |
| ALA | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| CYS | -0.85 | -0.85 | -0.85 | -0.60 | -0.60 |
| THR | -5.44 | -5.44 | -5.44 | -8.22 | -8.22 |
| SER | -6.43 | -3.71 | -4.74 | -5.68 | -6.44 |
| ASP | -17.28 | -11.90 | -15.88 | -20.31 | -25.21 |
| GLU | -17.35 | -11.97 | -15.95 | -17.67 | -22.57 |
| ASN | -12.25 | -7.82 | -10.22 | -17.70 | -20.67 |
| GLN | -11.50 | -7.07 | -9.47 | -14.35 | -17.32 |
| [b]$HIS^+$ | 9.02 | 12.53 | 9.73 | 13.52 | 9.80 |
| [b]$HIS_\epsilon$ | 6.98 | 10.49 | 7.69 | 9.90 | 6.18 |
| [b]$HIS_\delta$ | 7.35 | 10.86 | 8.06 | 10.62 | 6.89 |
| ARG | -36.90 | -32.00 | -35.18 | -54.08 | -57.90 |
| LYS | -11.71 | -6.76 | -10.17 | -8.41 | -12.35 |
| ILE | 4.22 | 4.63 | 3.63 | 5.30 | 4.00 |
| VAL | -0.15 | 0.26 | -0.74 | -0.89 | -2.19 |
| LEU | -0.53 | -0.12 | -1.12 | -0.97 | -2.27 |
| MET | -1.78 | -2.05 | -2.40 | -1.11 | -2.17 |
| PHE | -3.98 | -0.23 | -4.17 | 0.66 | -4.01 |
| TRP | -5.96 | -2.21 | -6.15 | 0.17 | -4.50 |
| TYR | -10.09 | -5.80 | -9.82 | -7.87 | -12.52 |

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

Table 3: Amino acid composition (%) of natural and designed PDZ proteins

| type | Natural sequences | | | | Designed sequences | | | |
|---|---|---|---|---|---|---|---|---|
| | Buried | | Exposed | | Buried | | Exposed | |
| | type | class | type | class | type | class | type | class |
| A | 5.9 | | 4.6 | | 4.1 | | 7.2 | |
| C | 1.5 | 11.2 | 1.2 | 13.4 | 8.6 | 12.7 | 5.8 | 13.6 |
| T | 3.8 | | 7.6 | | 0.0 | [1.5] | 0.6 | [0.2] |
| S | 4.7 | 4.7 | 10.2 | 10.2 | 4.9 | 4.9 [0.2] | 10.7 | 10.7 [0.5] |
| D | 3.5 | 9.6 | 6.2 | 16.7 | 7.4 | 9.4 | 8.0 | 16.1 |
| E | 6.1 | | 10.5 | | 2.0 | [-0.2] | 8.1 | [-0.6] |
| N | 1.9 | 2.7 | 7.4 | 16.1 | 1.8 | 2.8 | 8.6 | 17.1 |
| Q | 0.8 | | 8.7 | | 1.0 | [0.1] | 8.5 | [1.0] |
| $H^+$ | 0.7 | | 4.7 | | 0.1 | | 1.8 | |
| $H_\epsilon$ | 0.0 | 0.7 | 0.0 | 4.7 | 0.6 | 0.9 [0.2] | 2.2 | 4.5 [-0.2] |
| $H_\delta$ | 0.0 | | 0.0 | | 0.2 | | 0.5 | |
| I | 15.7 | | 4.1 | | 25.1 | | 8.4 | |
| V | 13.5 | 49.6 | 5.5 | 14.4 | 12.8 | 46.7 [-2.9] | 3.3 | 15.3 [0.9] |
| L | 20.4 | | 4.8 | | 8.8 | | 3.6 | |
| M | 5.0 | 5.0 | 1.4 | 1.4 | 5.9 | 5.9 [0.9] | 1.4 | 1.4 [0.0] |
| K | 6.5 | 6.5 | 10.1 | 10.1 | 5.5 | 5.5 [-1.0] | 10.8 | 10.8 [0.7] |
| R | 1.8 | 1.8 | 9.5 | 9.5 | 2.2 | 2.2 [0.4] | 9.1 | 9.1 [-0.4] |
| F | 5.0 | 5.0 | 0.4 | 0.4 | 3.2 | 5.5 | 0.3 | 0.5 |
| W | 0.0 | | 0.0 | | 2.3 | [0.5] | 0.2 | [0.1] |
| Y | 2.9 | 2.9 | 0.9 | 0.9 | 3.4 | 3.4 [0.5] | 0.9 | 0.9 [0.0] |
| G | 0.0 | 0.3 | 1.7 | 2.1 | 0.0 | 0.0 | 0.0 | 0.0 |
| P | 0.3 | | 0.4 | | 0.0 | [-0.3] | 0.0 | [-2.1] |
| | 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 experimental target set included the Tiam1 and Cask homologs.

Table 4: 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 # |
|---|---|---|---|---|---|---|
| Tiam1 | Proteus, $\epsilon_P$=4 | 53/94 | 1.0e-4 | 10000 | 7.0e-2 | 5259 |
| Cask | Proteus, $\epsilon_P$=4 | 76/83 | 5.1e-7 | 10000 | 1.6e-2 | 10000 |
| syntenin | Proteus, $\epsilon_P$=8 | 69/91 | 1.3e-2 | 9999 | 4.0e-3 | 9999 |
| DLG2 | Proteus, $\epsilon_P$=8 | 85/97 | 8.0e-9 | 10000 | 5.0e-3 | 10000 |
| Tiam1 | Proteus, $\epsilon_P$=8 | 64/94 | 1.2e-4 | 9920 | 5.2e-2 | 9058 |
| Cask | Proteus, $\epsilon_P$=8 | 71/83 | 3.2e-7 | 10000 | 8.2e-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 |
| syntenin | Rosetta | 76/82 | 7.3e-13 | 10000 | 1.8e-3 | 10000 |
| DLG2 | Rosetta | 86/97 | 1.3e-9 | 10000 | 9.6e-4 | 10000 |

[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.

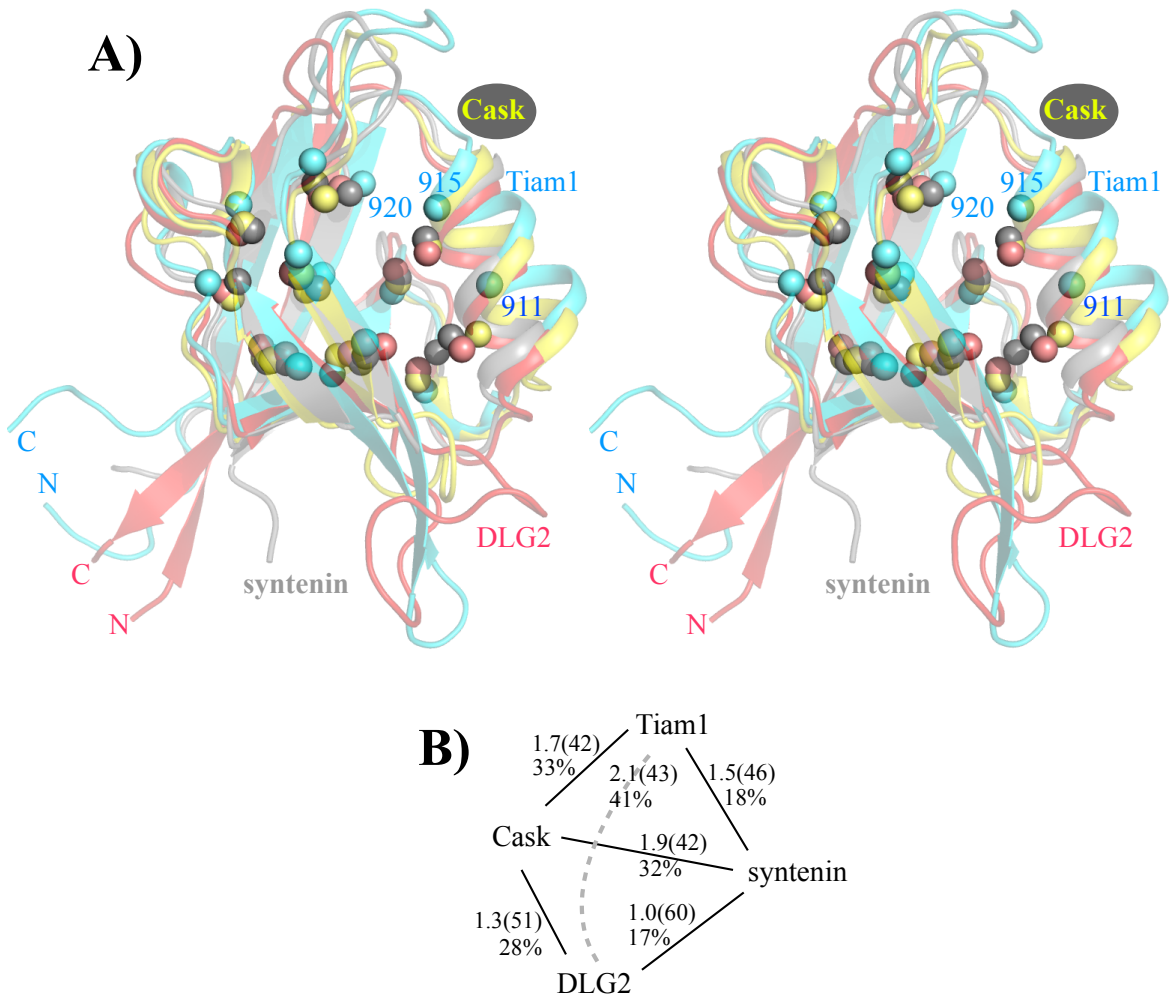


Figure 1: **A)** Three dimensional view of four PDZ domains. The $C_\beta$ atoms of fourteen hydrophobic core residues are shown as spheres. Three core positions designed in this work are labelled (Tiam1 numbers). **B)** Cluster representation of the PDZ domains studied. The links between domains are labeled with the percent identity scores and backbone rms deviations (Å); the number in parentheses is the number of aligned $C_\alpha$ atoms used to compute the rms deviation.

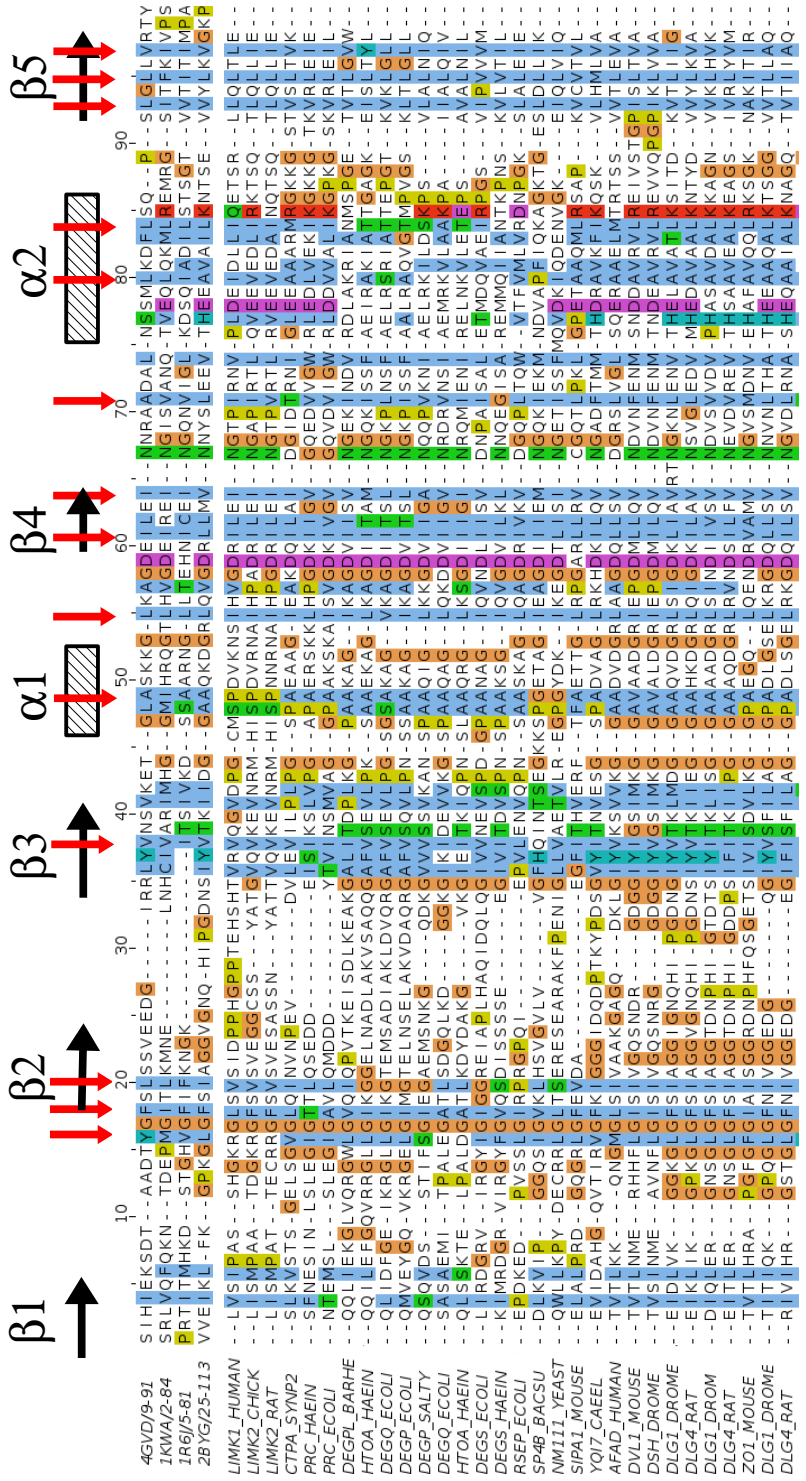


Figure 2: Alignment of natural PDZ sequences. The top four sequences were tested in this work. The others are the first 30 sequences from the Pfam seed alignment. Fourteen hydrophobic core positions are indicated by red arrows and the secondary structure elements are shown for reference. The Clustal color scheme is used, where conserved amino acids are colored according to their physical chemical properties.

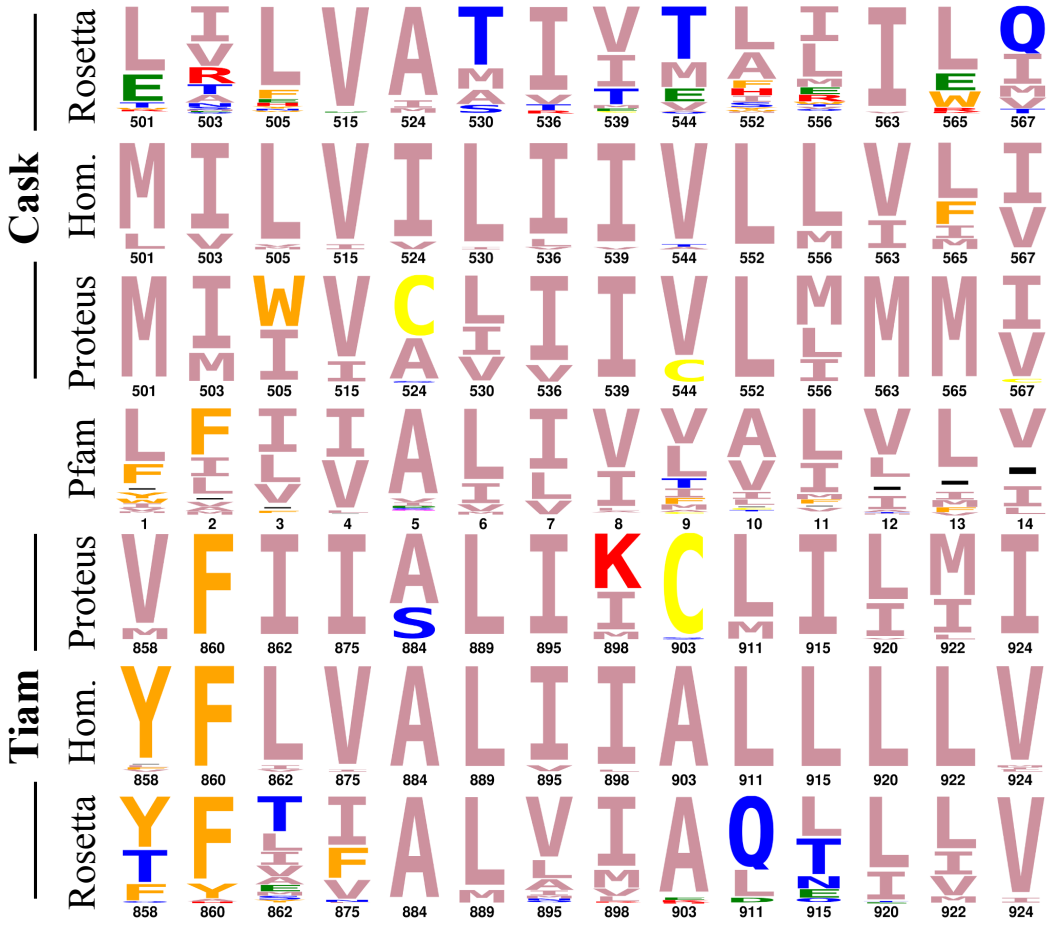


Figure 3: Sequence logos for the conserved hydrophobic core of designed and natural Tiam1 and Cask sequences. "Hom." corresponds to the homologs that make up our target set of sequences (used for $E_t^r$ optimization). "Pfam" corresponds to the Pfam seed alignment. Proteus sequences were generated with model $\epsilon_P$=8. 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. The color of each letter is determined by the physical chemical properties of each amino acid type.


Figure 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 could mutate, the rest of the protein having its native sequence (see text). Amino acid colors as in Fig. 3.

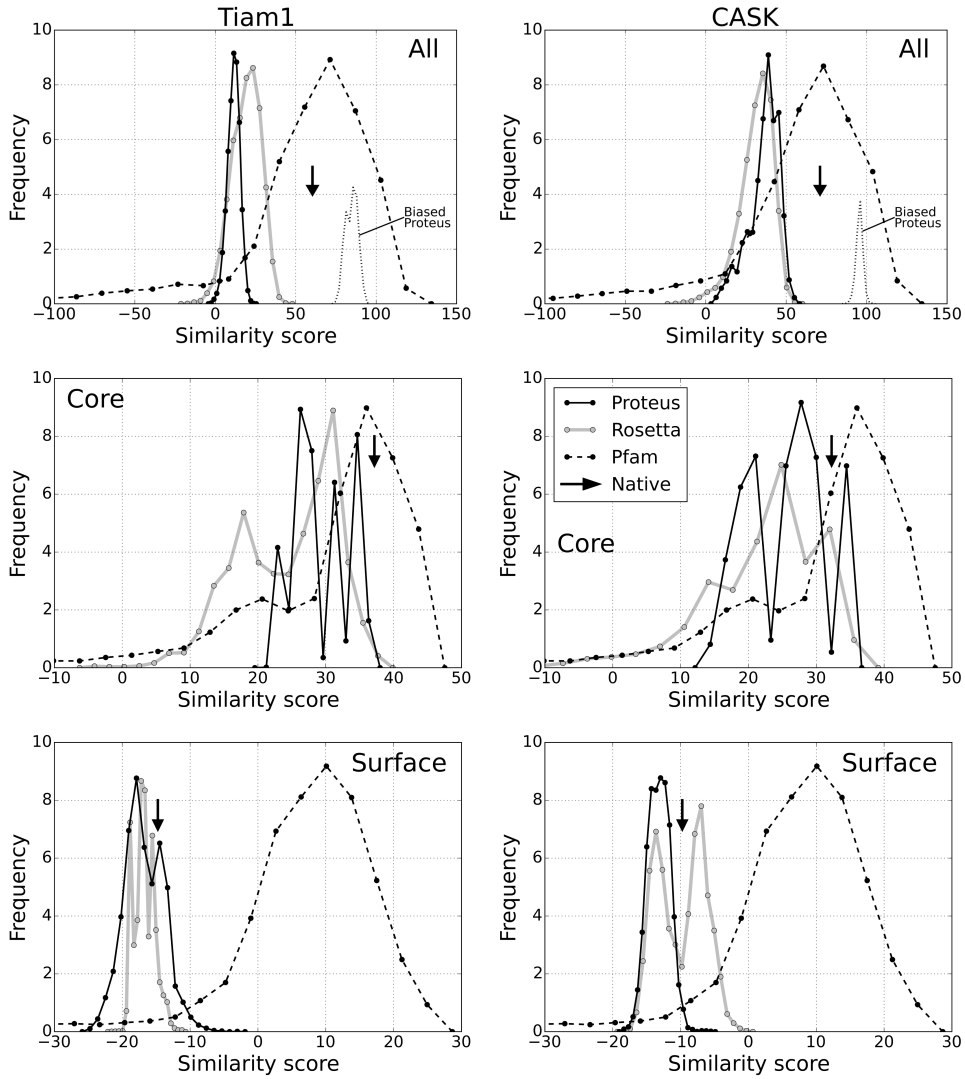


Figure 5: Histogram plots showing similarity scores for designed PDZ sequences. Similarity scores for Tiam1 (left) and Cask (right), relative to the Pfam-RP55 alignment. The scores were computed for all positions (top), 14 core positions (middle), and 16 surface positions (bottom). Values are shown for 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 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 were the same; only the positions included in the similarity score calculation differ between panels.

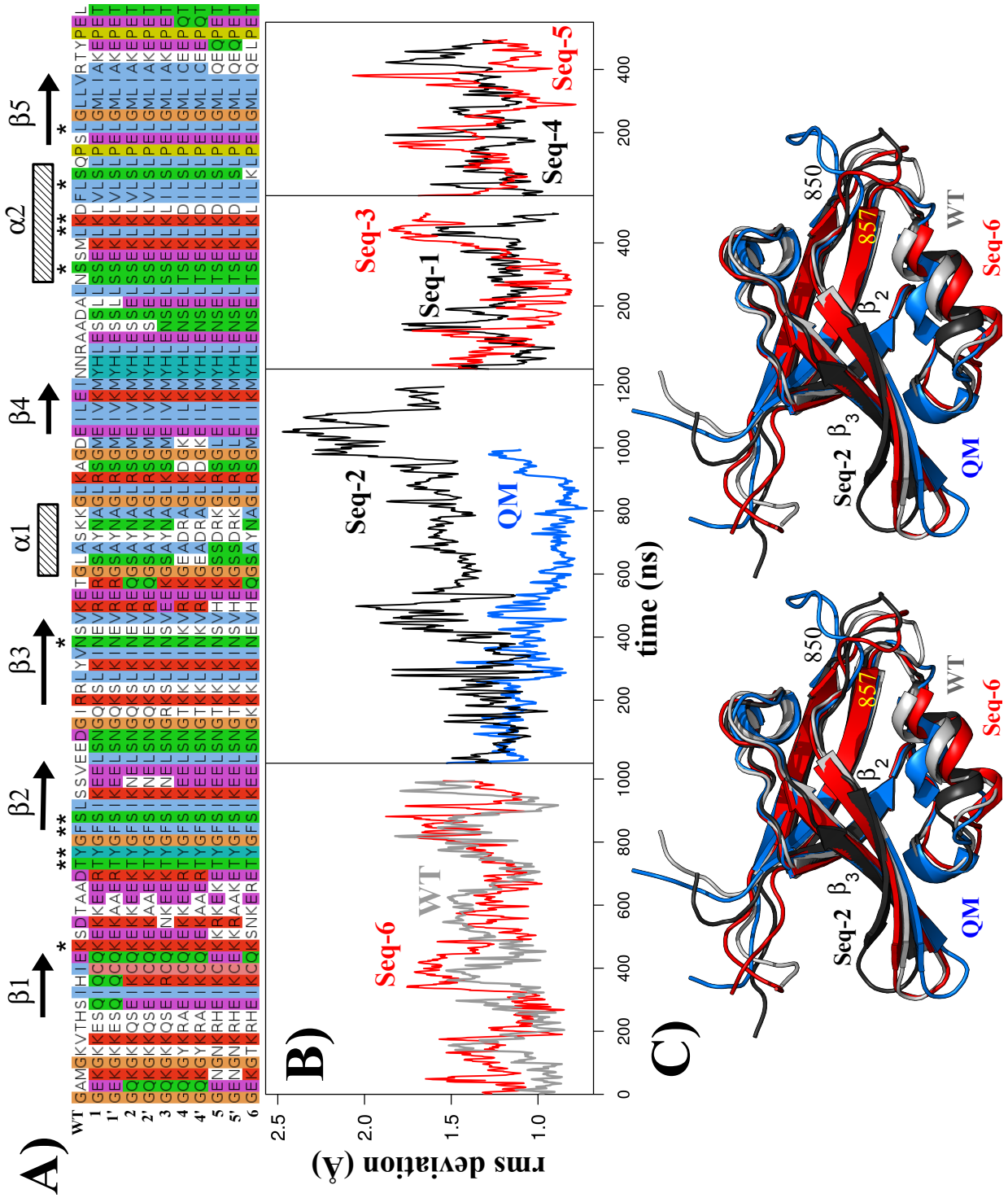


Figure 6: MD simulations of Tiam1 variants designed with Proteus. **A)** Sequences of the wildtype (WT) PDZ domain and the ten designed variants simulated by MD. Asterisks indicate peptide binding residues held fixed during the design simulations. **B)** Backbone rms deviations over the course of an MD simulation for WT, QM, and 6 designed variants relative to the corresponding mean MD structure. **C)** Mean MD structures of WT, QM, seq-6, seq-2.

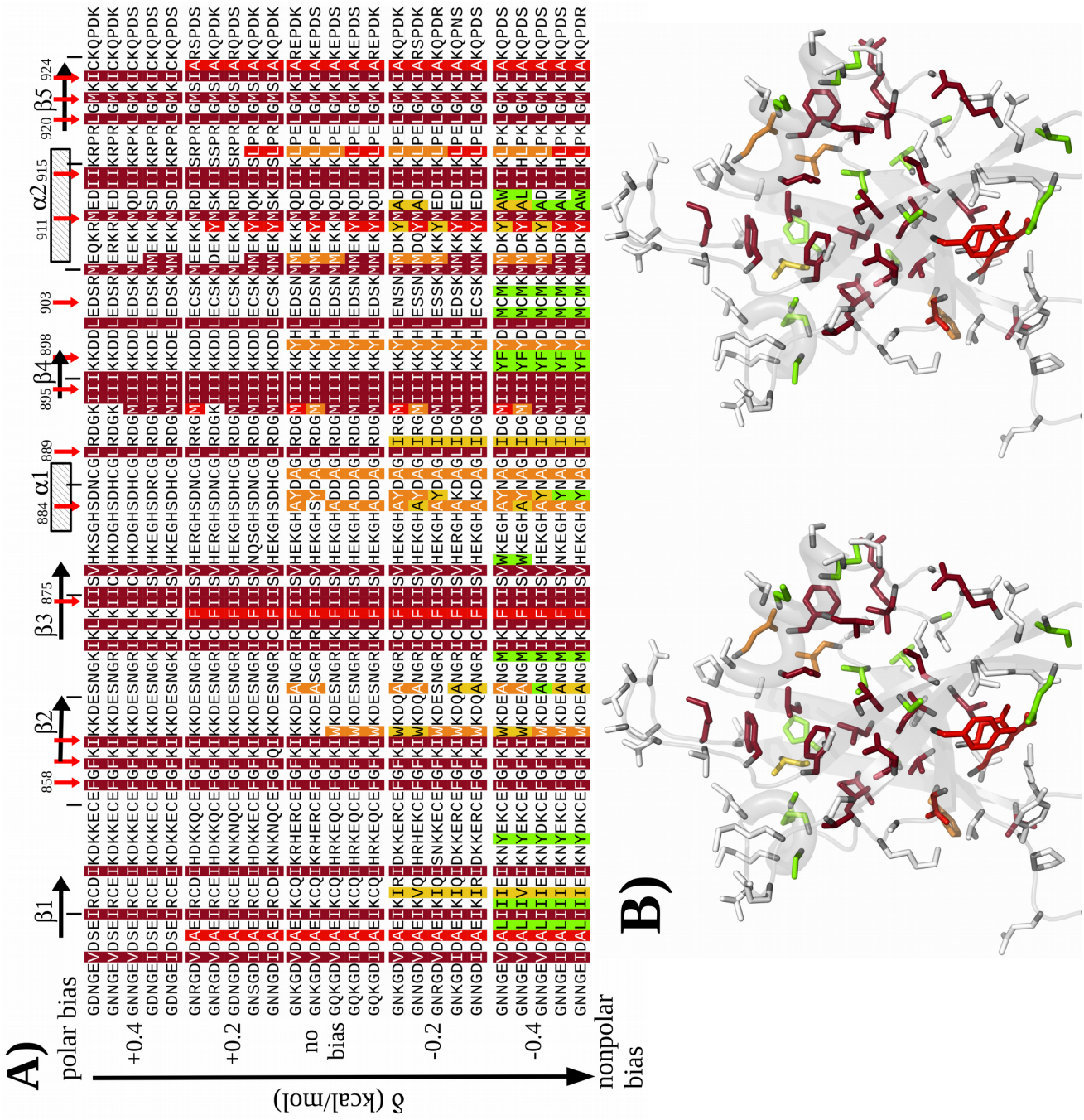


Figure 7: **A)** Tiam1 sequences designed with different levels of hydrophobic bias $\delta$, increasing from top to bottom. The middle sequences were obtained from Proteus simulations without any bias ($\delta = 0$). For each bias level, five low energy sequences are shown. Hydrophobic positions are colored according to the simulation where they appear first: from brick red (top) to light green (bottom). The 14 hydrophobic core positions are indicated by red arrows. **B)** 3D Tiam1 structure (stereo) with colors as in **A)**.

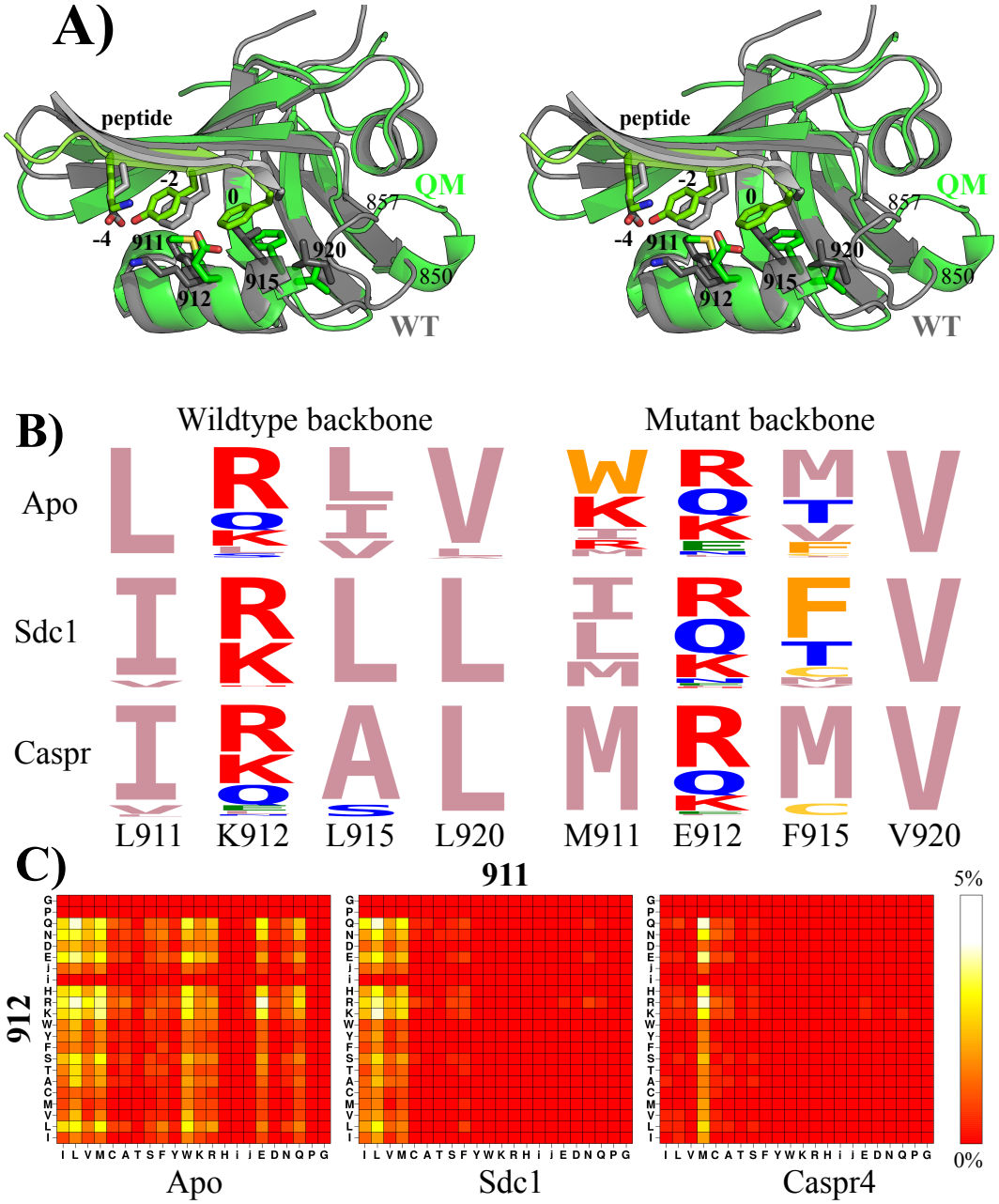


Figure 8: Design of four Tiam1 specificity positions. **A)** X-ray structures (stereo) with the wildtype sequence (LKLL; labelled WT; PDB code 4GVD) or the quadruple mutant sequence (MEFV; labelled QM; PDB code 4NXQ), with bound Sdc1 and Caspr4, respectively. The four designed sidechains are shown and labelled (both WT and QM 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 with the wildtype backbone.

# Computational design of the Tiam1 PDZ domain and its ligand binding

David Mignon, Nicolas Panel, Xingyu Chen, Ernesto J. Fuentes and Thomas Simonson

## Table of contents graphic