

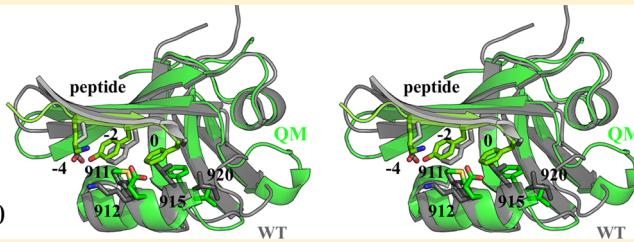
Computational Design of the Tiam1 PDZ Domain and Its Ligand Binding

David Mignon,[†] Nicolas Panel,[†] Xingyu Chen,[†] Ernesto J. Fuentes,[‡] and Thomas Simonson*,[†]

[†]Laboratoire de Biochimie (CNRS UMR7654), Ecole Polytechnique, Palaiseau, France

[‡]Department of Biochemistry, Roy J. & Lucille A. Carver College of Medicine and Holden Comprehensive Cancer Center, University of Iowa, Iowa City, Iowa 52242-1109, United States

Supporting Information



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 the 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.

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. Because of 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 side chain 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.⁴¹

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 side chain rotamer library. The unfolded state energy depends only on sequence composition, not an explicit structural model. The main adjustable model parameters are the protein dielectric constant ϵ_p , a small set of atomic surface energy coefficients σ_p , 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

Received: December 28, 2016

Published: April 10, 2017

63 unfolded state reference energies. It has a large number of
 64 parameters specifically optimized for CPD, which provide
 65 optimal performance, but less transferability and a less
 66 transparent physical interpretation. Proteus also provides
 67 some specific functionalities, such as Replica Exchange Monte
 68 Carlo, various importance sampling methods, and the ability to
 69 compute free energies that are formally exact.^{33–35}

70 We optimized the reference energies E_t^r for the Tiam1 and
 71 Cask PDZ proteins, using a maximum likelihood formalism. We
 72 compared two values of the protein dielectric constant, $\epsilon_p = 4$
 73 and 8. These values gave good results in a systematic study that
 74 compared dielectric constants in the range 1–32.³⁶ The
 75 performance of the model was tested by generating designed
 76 sequences for both proteins and comparing them to natural
 77 sequences, as well as sequences generated with the Rosetta
 78 energy function and software.³⁷ The sequence design was
 79 performed by running long Monte Carlo simulations in which
 80 all protein positions except Gly and Pro were allowed to mutate
 81 freely, leading to thousands of designed protein variants. The
 82 testing included cross-validation, for which the reference
 83 energies were optimized using one set of PDZ domains, then
 84 applied to others. We also performed 100–1000 ns molecular
 85 dynamics (MD) simulations for a few of the sequences
 86 designed with our optimized CPD model, to help assess their
 87 stability. Ten sequences were stable over 100 ns or more and
 88 two over 1000 ns of MD simulation.

89 We then applied the CPD model with optimized parameters
 90 to two problems, which are representative of the two main
 91 areas we are interested in exploring: the plasticity of sequence
 92 space for PDZ domains and designing strong and specific PDZ
 93 ligands. Earlier applications in these areas mostly employed
 94 empirical, knowledge-based energy functions such as the
 95 Rosetta function.^{9,10,13,14} First, we performed a series of
 96 Monte Carlo simulations of two PDZ domains where the
 97 chemical potential of the hydrophobic amino acid types was
 98 gradually increased, artificially biasing the protein composition.
 99 As the hydrophobic bias was increased, hydrophobic amino
 100 acids gradually invaded the protein from the inside out, forming
 101 a hydrophobic core that became larger than the natural one.
 102 The propensity of each core position to become hydrophobic at
 103 a high or low level of bias can be seen as a structure-dependent
 104 hydrophobicity index, which provides information on the
 105 designability or plasticity of the protein core. The second
 106 application consisted in designing four Tiam1 positions known
 107 to be involved in specific target recognition. These four
 108 positions were varied through Monte Carlo simulations of
 109 either the apoprotein or the protein in complex with two
 110 distinct peptide ligands. The simulations were in agreement
 111 with experimental sequences and binding affinities, and suggest
 112 new variants that could have altered specificities. This
 113 application is a step toward the design of strong peptide
 114 binders, which could be of use as reagents or inhibitors *in vitro*
 115 or *in vivo*.

2. THE UNFOLDED STATE MODEL

116 **2.1. Maximum Likelihood Reference Energies.** The
 117 Monte Carlo method employed here generates a Markov chain
 118 of states,^{38,39} such that the states are populated according to a
 119 Boltzmann distribution. The energy employed is not the folded
 120 protein's energy, but rather its *folding* energy, that is, the
 121 difference between its folded and unfolded state energies.³³
 122 One possible elementary move is a “mutation”, we modify the
 123 side chain type $t \rightarrow t'$ at a chosen position i in the folded

protein, assigning a particular rotamer r' to the new side chain.¹²⁴
 We consider the same mutation in the unfolded state. For a
 125 particular sequence S , the unfolded state energy has the form:¹²⁶

$$E^u = \sum_{i \in S} E^r(t_i) \quad (1) \quad 127$$

The sum is over all amino acids; t_i represents the side chain
 128 type at position i . The type-dependent quantities $E^r(t) \equiv E_t^r$ are
 129 referred to as “reference energies”; they can be thought of as
 130 effective chemical potentials of each amino acid type. The
 131 energy change due to a mutation has the form:¹³²

$$\begin{aligned} \Delta E &= \Delta E^f - \Delta E^u \\ &= (E^f(\dots t'_i, r'_i \dots) - E^f(\dots t_i, r_i \dots)) - (E^r(t'_i) - E^r(t_i)) \end{aligned} \quad (2) \quad 133$$

where ΔE^f and ΔE^u are the energy changes in the folded and
 134 unfolded state, respectively. The reference energies are essential
 135 parameters in the simulation model. Our goal here is to choose
 136 them empirically so that the simulation produces amino acid
 137 frequencies that match a set of target values, for example
 138 experimental values in the Pfam database. Specifically, we will
 139 choose them so as to maximize the probability, or likelihood of
 140 the target sequences.¹⁴¹

Let S be a particular sequence. Its Boltzmann probability is¹⁴²

$$p(S) = \frac{1}{Z} \exp(-\beta \Delta G_S) \quad (3) \quad 143$$

where $\Delta G_S = G_S^f - E_S^u$ is the folding free energy of S , G_S^f is the
 144 free energy of the folded form, $\beta = 1/kT$ is the inverse
 145 temperature, and Z is a normalizing constant (the partition
 146 function). We then have¹⁴⁷

$$\begin{aligned} kT \ln p(S) &= \sum_{i \in S} E^r(t_i) - G_S^f - kT \ln Z \\ &= \sum_{t \in aa} n_S(t) E_t^r - G_S^f - kT \ln Z \end{aligned} \quad (4) \quad 148$$

where the sum on the right is over the amino acid types and¹⁴⁹
 $n_S(t)$ is the number of amino acids of type t within the sequence¹⁵⁰
 S .¹⁵¹

We now consider a set \mathcal{S} of N target sequences S ; we denote¹⁵²
 \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 have¹⁵³

$$\begin{aligned} kT \ln \mathcal{L} &= \sum_S \sum_{t \in aa} n_S(t) E_t^r - \sum_S G_S^f - NkT \ln Z \\ &= \sum_{t \in aa} N(t) E_t^r - \sum_S G_S^f - NkT \ln Z \end{aligned} \quad (5) \quad 154$$

where $N(t)$ is the number of amino acids of type t in the whole
 155 data set \mathcal{S} . The normalization factor or partition function Z is a
 156 sum over all possible sequences R :¹⁵⁷

$$\begin{aligned} Z &= \sum_R \exp(-\beta \Delta G_R) \\ &= \sum_R \exp(-\beta \Delta G_R^f) \prod_{t \in aa} \exp(\beta n_R(t) E_t^r) \end{aligned} \quad (6) \quad 158$$

In view of maximizing \mathcal{L} , we consider the derivative of Z with
 159 respect to one of the E_t^r :¹⁶⁰

$$\frac{\partial Z}{\partial E_t^r} = \sum_R \beta n_R(t) \exp(-\beta \Delta G_R^f) \prod_{s \in aa} \exp(\beta n_R(s) E_s^r) \quad (7) \quad 161$$

162 We then have

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

164 The quantity on the right is the Boltzmann average of the
165 number $n(t)$ of amino acids t over all possible sequences. In
166 practice, this is the average population of t we would obtain in a
167 long MC simulation. As usual in statistical mechanics,⁴¹ the
168 derivative of $\ln Z$ with respect to one quantity (E_t^r) is equal to
169 the ensemble average of the conjugate quantity ($\beta n_S(t)$).
170 A necessary condition to maximize $\ln \mathcal{L}$ is that its derivatives
171 with respect to the E_t^r should all be zero. We see that

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

172

173 so that

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

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

178 **2.2. Searching for the Maximum Likelihood.** We will
179 use two methods to approach the maximum likelihood $\{E_t^r\}$
180 values, starting from a current guess $\{E_t^r(n)\}$. With the first
181 method, we step along the gradient of $\ln \mathcal{L}$, using the update
182 rule:⁴⁰

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

184 Here, α is a constant; $n_t^{\text{exp}} = N(t)/N$ is the mean population of
185 amino acid type t in the target database; $\langle \cdot \rangle_n$ indicates an average
186 over a simulation done using the current reference energies
187 $\{E_t^r(n)\}$, and δE is an empirical constant with the dimension of
188 an energy, referred to as the update amplitude. This update
189 procedure is repeated until convergence. We refer to this
190 method as the linear update method.

191 The second method, used previously,^{26,27} employs a
192 logarithmic update rule:

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

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

198 In the later iterations, some E_t^r values tended to converge
199 slowly, with an oscillatory behavior. Therefore, we sometimes
200 used a modified update rule, where the $E_t^r(n+1) - E_t^r(n)$
201 value computed with the linear or logarithmic method for
202 iteration n was mixed with the value computed at the previous
203 iteration, with the $(n-1)$ value having a weight of $1/3$ and the
204 current value a weight of $2/3$. At each iteration, we typically ran
205 500 million steps (per replica) of Replica Exchange Monte
206 Carlo.

3. COMPUTATIONAL METHODS

3.1. Effective Energy Function for the Folded State.²⁰⁷

The energy matrix was computed with the following effective
208 energy function for the folded state:
209

$$\begin{aligned} E &= E_{\text{bonds}} + E_{\text{angles}} + E_{\text{dihes}} + E_{\text{impr}} + E_{\text{vdw}} + E_{\text{Coul}} \\ &\quad + E_{\text{solv}} \end{aligned} \quad (13)$$

The first six terms in eq 13 represent the protein internal
211 energy. They were taken from the Amber ff99SB empirical
212 energy function,⁴² slightly modified for CPD. The original
213 backbone charges were replaced by a unified set, obtained by
214 averaging over all amino acid types and adjusting slightly to
215 make the backbone portion of each amino acid neutral.⁴³ The
216 last term on the right of eq 13, E_{solv} , represents the contribution
217 of solvent. We used a “Generalized Born + Surface Area”, or
218 GBSA implicit solvent model.⁴⁴

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

Here, ϵ_W and ϵ_P are the solvent and protein dielectric
221 constants; r_{ij} is the distance between atoms i,j and b_i is the
222 “solvation radius” of atom i .^{44,45} A_i is the exposed solvent
223 accessible surface area of atom i ; σ_i is a parameter that reflects
224 each atom’s preference to be exposed or hidden from solvent.
225 The solute atoms were divided into four groups with specific σ_i
226 values. The values were -5 (nonpolar), -40 (aromatic), -80
227 (polar), and -100 (ionic) cal/mol/Å². Hydrogen atoms were
228 assigned a surface coefficient of 0. Surface areas were computed
229 by the Lee and Richards algorithm,⁴⁶ implemented in the
230 XPLOR program,⁴⁷ using a 1.5 Å probe radius. Most of the MC
231 simulations used a protein dielectric of $\epsilon_P = 4$ or 8 (see
232 Results).

In the GB energy term, the atomic solvation radius b_i
234 approximates the distance from i to the protein surface and is
235 a function of the coordinates of all the protein atoms. The
236 particular b_i form corresponds to a GB variant we call GB/
237 HCT, after its original authors,⁴⁴ with model parameters
238 optimized for use with the Amber force field.⁴⁵ Since b_i
239 depends on the coordinates of all the solute atoms,⁴⁴ an
240 additional approximation is needed to make the GB energy
241 term pairwise additive and to define the energy matrix.^{27,48} We
242 use a “Native Environment Approximation”, or NEA, in which
243 the solvation radius b_i of each particular group (backbone, side
244 chain or ligand) is computed ahead of time, with the rest of the
245 system having its native sequence and conformation.^{27,48}

The surface energy contribution E_{surf} is not pairwise additive
247 either, because in a protein structure, surface area buried by one
248 side chain may also be buried by another. To make this energy
249 pairwise, we used the method of Street et al.⁴⁹ In this method,
250 the buried surface of a side chain is computed by summing over
251 the neighboring side chain and backbone groups. For each
252 neighboring group, the contact area with the side chain of
253 interest is computed, independently of other surrounding
254 groups. The contact areas are then summed. To avoid
255 overcounting the buried surface area, a scaling factor is applied
256 to the contact areas involving buried side chains. Previous
257 studies 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 homologues (see below). As a result, the model should be more transferable within a protein family.

Distinguishing buried/exposed positions double 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

$$E_t^r = E_c^r + \delta E_t^r \quad (15)$$

Here, E_c^r is an adjustable parameter while δE_t^r is a constant, computed as the molecular mechanics energy difference between amino acid types within the class c , assuming an unfolded conformation where each amino acid interacts only with itself and with solvent. 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 was optimized while δE_t^r is held fixed. To optimize the E_c^r values, we applied the linear or logarithmic method while the target frequencies corresponded 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 $\Phi\text{-X}\text{-}\Phi$ 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 homologues with a sequence identity, relative to the query, above a 60% threshold and below an 85% threshold. If two homologues 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 homologue sequences. The two sets of homologues are referred to as \mathcal{H}_T and \mathcal{H}_C , respectively. For each of the sets, say \mathcal{H} , we average over all homologues and all positions to obtain a computation of 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, and 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, $f_t^b = (f_t^b(\mathcal{H}_T) + f_t^b(\mathcal{H}_C))/2$ for each type t , and similarly for the exposed positions. Distinct target frequencies were thus obtained for buried and exposed positions.

Model parametrization 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 modeled by removing this peptide. For the PDZ domain of Tiam1, we also used a holo structure then modeled 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 cocrystallized, 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.

Table 1. Test Proteins

protein name ^a	PDB code	residue numbers	no. 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

^aIn parentheses: number of the PDZ domain within the protein.

^bHolo or holo-like structures. ^cThe number of non-Gly, non-Pro positions, which can mutate during the design simulations.

To carry out the Monte Carlo design calculations, the structures were prepared and energy matrices were 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 Modeler 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 side chain

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

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

where β_i, β_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: $T_{i+1}/T_i = \text{constant}$.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

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

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

3.5. 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-instruct 10000-ex1-ex2-linmem_ig 10 where the ex1 and ex2 options activate an enhanced rotamer search for buried side chains, 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 min per sequence on a single core of a recent Intel processor, for a total of 10 h (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 homologues (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:

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

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

488 the binding free energy differences between protein variants. If
 489 a given sequence S was sampled in both the apo and holo
 490 states, we computed the mean energy $\langle E_{\text{holo}}(S) \rangle$, $\langle E_{\text{apo}}(S) \rangle$ in
 491 each of the two states by averaging over the sampled
 492 conformations. Then, we took the difference

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

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

500 **3.8. Molecular Dynamics Simulations.** Wildtype and a
 501 quadruple mutant Tiam1 and 10 sequences designed with
 502 Proteus were subjected to MD simulations with explicit solvent
 503 and no peptide ligand. The starting structures were taken from
 504 the MC trajectory or the crystal structure (wildtype protein and
 505 quadruple mutant: PDB codes 4GVD and 4NXQ) and slightly
 506 minimized with harmonic restraints to maintain the backbone
 507 geometry. The protein was immersed in a large box of
 508 nonoverlapping waters. The solvated system was truncated to
 509 the shape of a truncated octahedral box using the Charmm
 510 graphical interface or GUI.⁶³ The minimum distance between
 511 protein atoms and the box was 15 Å and the final models
 512 included about 11 000 water molecules. A few sodium or
 513 chloride ions were included to ensure overall electroneutrality.
 514 The protonation states of histidines were assigned to be neutral,
 515 based on visual inspection. MD was done at room temperature
 516 and pressure, using a Nose-Hoover thermostat and baro-
 517 stat.^{64,65} Long-range electrostatic interactions were treated with
 518 a Particle Mesh Ewald approach.⁶⁶ The Amber ff99SB force
 519 field and the TIP3P model⁶⁷ were used for the protein and
 520 water, respectively. Simulations were run for 100–1000 ns,
 521 depending on the sequence, using the Charmm and NAMD
 522 programs.^{68,69}

4. RESULTS

523 **4.1. Experimental structures and sequences.** Three
 524 dimensional (3D) structures of the four test PDZ domains are
 525 shown in Figure 1A. Fourteen core residues (identified visually)
 526 superimposed well between the structures, while loops and
 527 chain termini displayed large deviations. The Tiam1 α_2 helix is
 528 rotated slightly outward compared to the other three
 529 structures.⁷⁰ Figure 1B illustrates the similarity between pairs
 530 of PDZ domains, as determined by the rms deviation between
 531 structurally aligned C_α atoms and the pairwise sequence
 532 identities. The rms deviations are between 1.0 and 2.1 Å and
 533 the sequence identities between 17 and 33%. The Tiam1/Cask
 534 sequence identity is 33% and their structural deviation is 1.7 Å
 535 based on 42 aligned C_α atoms. The syntenin and DLG2
 536 structures are more similar, with a structural deviation of 1.0 Å
 537 based on 60 aligned C_α atoms.

538 Sequence conservation within the four PDZ domains and a
 539 subset of the Pfam seed alignment is shown in Figure 2. The 14
 540 positions used to define the hydrophobic core are highly,
 541 although not totally conserved within the Pfam seed alignment.
 542 Arginine, Lys, and Gln appear at some of the positions, since in
 543 small proteins such as PDZ domains, the long hydrophobic

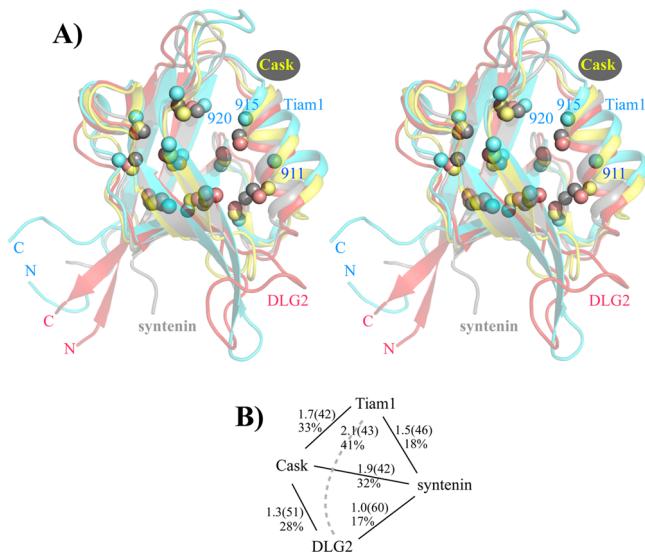


Figure 1. (A) Three dimensional view of four PDZ domains. The C_β atoms of 14 hydrophobic core residues are shown as spheres. Three core positions designed in this work are labeled (Tiam1 numbers). (B) Cluster representation of the PDZ domains studied. The links between domains are labeled with the percent identity scores and backbone rms deviations (Å); the number in parentheses is the number of aligned C_α atoms used to compute the rms deviation.

portion of these side chains can be buried in the core while still allowing the polar tip of the side chain 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 side chain superposition.

4.2. Optimizing the Unfolded State Model. We optimized the reference energies E_t^r for Tiam1 and Cask, using their natural homologues to define the target amino acid frequencies. The protein dielectric constant ϵ_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 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 homologues and the simulations using parameters optimized with $\epsilon_p = 8$. Results obtained using parameters optimized with $\epsilon_p = 4$ are given in Supporting Information. 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 intraclass frequency distributions depend explicitly on the energy offsets δE_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,

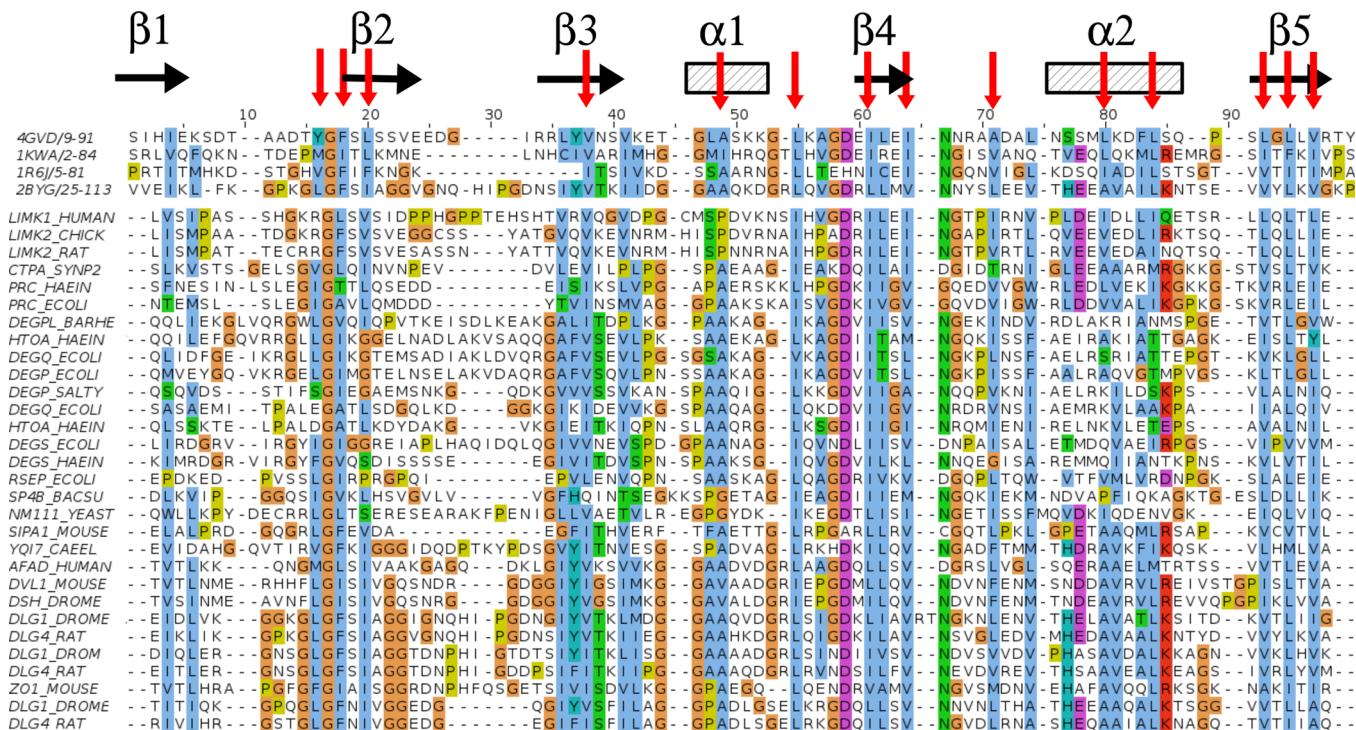


Figure 2. Alignment of natural PDZ sequences. The top four sequences were tested in this work. The others are the first 30 sequences from the Pfam seed alignment. Fourteen hydrophobic core positions are indicated by red arrows and the secondary structure elements are shown for reference. The Clustal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

Table 2. Unfolded State Reference Energies E_t^r (kcal/mol)

residues	peptide ^a	design	
		buried	exposed
ALA	0.00	0.00	0.00
CYS	-1.04	-0.85	-0.85
THR	-3.82	-5.44	-5.44
SER	-2.85	-3.71	-4.74
ASP	-9.17	-11.90	-15.88
GLU	-7.88	-11.97	-15.95
ASN	-5.64	-7.82	-10.22
GLN	-4.42	-7.07	-9.47
HIS ^b	15.72	12.53	9.73
HIS _ε ^b	12.62	10.49	7.69
HIS _δ ^b	13.16	10.86	8.06
ARG	-25.30	-32.00	-35.18
LYS	-4.21	-6.76	-10.17
ILE	1.63	4.63	3.63
VAL	-2.25	0.26	-0.74
LEU	-1.92	-0.12	-1.12
MET	-2.44	-2.05	-2.40
PHE	-1.42	-0.23	-4.17
TRP	-2.66	-2.21	-6.15
TYR	-4.56	-5.80	-9.82

^aEnergies within an extended peptide structure (averaged over positions). ^bHis protonation states.

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 Figure 3 for the 14 core residues and in Figure 4 for the 16 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 (Figure 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 RPSS set of 12 255 natural sequences has a mean entropy of 3.4. Pooling the designed Tiam1 and Cask sequences gave an

without any bias toward 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

Table 3. Amino Acid Composition (%) of Natural and Designed PDZ Proteins^a

type	natural sequences				designed sequences			
	buried		exposed		buried		exposed	
	type	class	type	class	type	class	type	class
A	5.9	11.2	4.6	13.4	4.1	12.7 [1.5]	7.2	13.6 [0.2]
C	1.5		1.2		8.6		5.8	
T	3.8		7.6		0.0		0.6	
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 [-0.2]	8.0	16.1 [-0.6]
E	6.1		10.5		2.0		8.1	
N	1.9	2.7	7.4	16.1	1.8	2.8 [0.1]	8.6	17.1 [1.0]
Q	0.8		8.7		1.0		8.5	
H ⁺	0.7	0.7	4.7	4.7	0.1	0.9 [0.2]	1.8	4.5 [-0.2]
H _e	0.0		0.0		0.6		2.2	
H _d	0.0		0.0		0.2		0.5	
I	15.7	49.6	4.1	14.4	25.1	46.7 [-2.9]	8.4	15.3 [0.9]
V	13.5		5.5		12.8		3.3	
L	20.4		4.8		8.8		3.6	
M	5.0	5.0	1.4	1.4	5.9 [0.9]	5.9	1.4 [0.0]	1.4
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 [0.5]	0.3	0.5 [0.1]
W	0.0		0.0		2.3		0.2	
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 [-0.3]	0.0	0.0 [-2.1]
P	0.3		0.4		0.0		0.0	

^aCompositions 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 homologues.

Table 4. Fold Recognition of Designed Sequences by Superfamily

protein	design model	match/seq length ^a	superfamily		family	
			E-value ^b	success no. ^c	E-value ^b	success no. ^c
Tiam1	Proteus, $\epsilon_p = 4$	53/94	1.0×10^{-4}	10000	7.0×10^{-2}	5259
Cask	Proteus, $\epsilon_p = 4$	76/83	5.1×10^{-7}	10000	1.6×10^{-2}	10000
syntenin	Proteus, $\epsilon_p = 8$	69/91	$1.3e \times 10^{-2}$	9999	$4. \times 10^{-3}$	9999
DLG2	Proteus, $\epsilon_p = 8$	85/97	8.0×10^{-9}	10000	5.0×10^{-3}	10000
Tiam1	Proteus, $\epsilon_p = 8$	64/94	1.2×10^{-4}	9920	5.2×10^{-2}	9058
Cask	Proteus, $\epsilon_p = 8$	71/83	3.2×10^{-7}	10000	8.2×10^{-3}	10000
Tiam1	Rosetta	65/94	4.4×10^{-4}	9035	$2.8e \times 10^{-2}$	9030
Cask	Rosetta	68/83	2.8×10^{-5}	9832	7.5×10^{-3}	9832
syntenin	Rosetta	76/82	7.3×10^{-13}	10000	1.8×10^{-3}	10000
DLG2	Rosetta	86/97	1.3×10^{-9}	10000	9.6×10^{-4}	10000

^aThe average match length for sequences recognized by superfamily and the total sequence length. ^bAverage E-values for superfamily assignments to the correct SCOP superfamily/family. ^cThe number of designed sequences (out of 10000 tested) assigned to the correct SCOP superfamily/family.

entropy of 2.2 with Rosetta and 2.0 with Proteus, indicating that these two backbone geometries cannot accommodate 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. Figure 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 (Figure 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

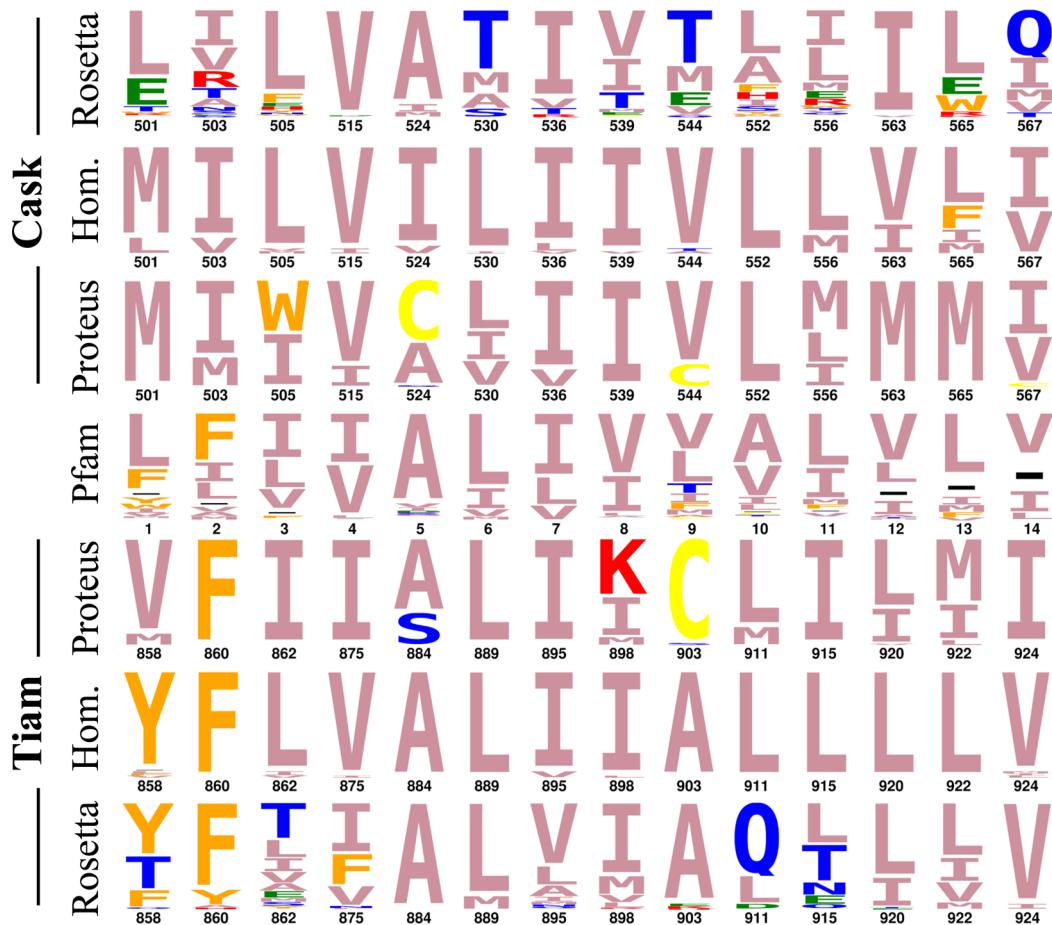


Figure 3. Sequence logos for the conserved hydrophobic core of designed and natural Tiam1 and Cask sequences. “Hom.” corresponds to the homologues that make up our target set of sequences (used for E_t optimization). “Pfam” corresponds to the Pfam seed alignment. Proteus sequences were generated with 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.

652 achieved by adding to the energy an “experimental,” or bias
653 energy term that explicitly favors high sequence scores. Figure 5
654 includes results that used such a biased energy term: by
655 construction, it led to very high similarity scores. A bias energy
656 term could also be used to limit the total number of mutations.
657

4.4. Cross-Validation Tests. As a first cross-validation test,
658 we applied the reference energies optimized using Tiam1 and
659 Cask homologues (with $\epsilon_p = 8$) to two other PDZ domains:
660 DLG2 and syntenin. The superfamily scores were comparable
661 to those obtained for Tiam1 and Cask, with 100% family
662 recognition (Table 4). Sequences designed with Rosetta for
663 DLG2 and syntenin also gave 100% family recognition. For
664 further cross-validation, we optimized reference energies using
665 an alternate set of PDZ domains: DLG2, syntenin, PSD95,
666 GRIP, INAD, and NHERF. Target frequencies were defined by
667 a small set of their natural homologues. We used $\epsilon_p = 8$. To
668 distinguish the new and initial model variants, we refer to the
669 new variant as the $n = 6$ model (it uses six PDZ domains for
670 parametrization), and the initial model as the “T+C” model (it
671 used Tiam1 and Cask). The new, $n = 6$ reference energies were
672 then used to produce designed Tiam1 and Cask sequences,
673 which were subjected to superfamily tests and similarity
674 calculations. The superfamily performance for Tiam1 was
675 slightly degraded, compared to the previous, T + C model. The
676 Tiam1 superfamily score decreased from 90.6% to 76.6% for

677 family recognition. The Cask score was unchanged. Histograms 678 of Blosum similarity scores (Supporting Information) show that 679 the overall scores for Tiam1 and Cask with $n = 6$ were very 680 similar to the T + C model, while the scores for the core 681 positions were actually shifted to higher, not lower values. For 682 DLG2 and syntenin, we also computed similarity scores using 683 both the initial, T+C parametrization and the new, $n = 6$ 684 parametrization. The similarity scores with the T+C model 685 were slightly poorer than with the $n = 6$ model, as expected. 686 The overall score decreased by about 20 points for syntenin 686 and about 10 points for DLG2 (Supporting Information). 687 Overall, the cross-validated models degraded performance 688 slightly. Thus, for any PDZ domain of interest, it may be 689 preferable to optimize reference energies specifically for that 690 doomain, rather than transferring values parametrized using 691 other PDZ domains. 692

4.5. Stability of Designed Sequences in Molecular 693 Dynamics Simulations. As another test of the design model, 694 10 Tiam1 sequences designed with Proteus were subjected to 695 molecular dynamics simulations (MD) using an explicit solvent 696 environment. These sequences were obtained using Proteus 697 with either $\epsilon_p = 8$ or the less polarizable value $\epsilon_p = 4$. Although 698 no peptide ligand was present during the design simulations, 11 699 positions in the binding pocket that make close contact with 700 the peptide when it is present were not allowed to mutate. This 701

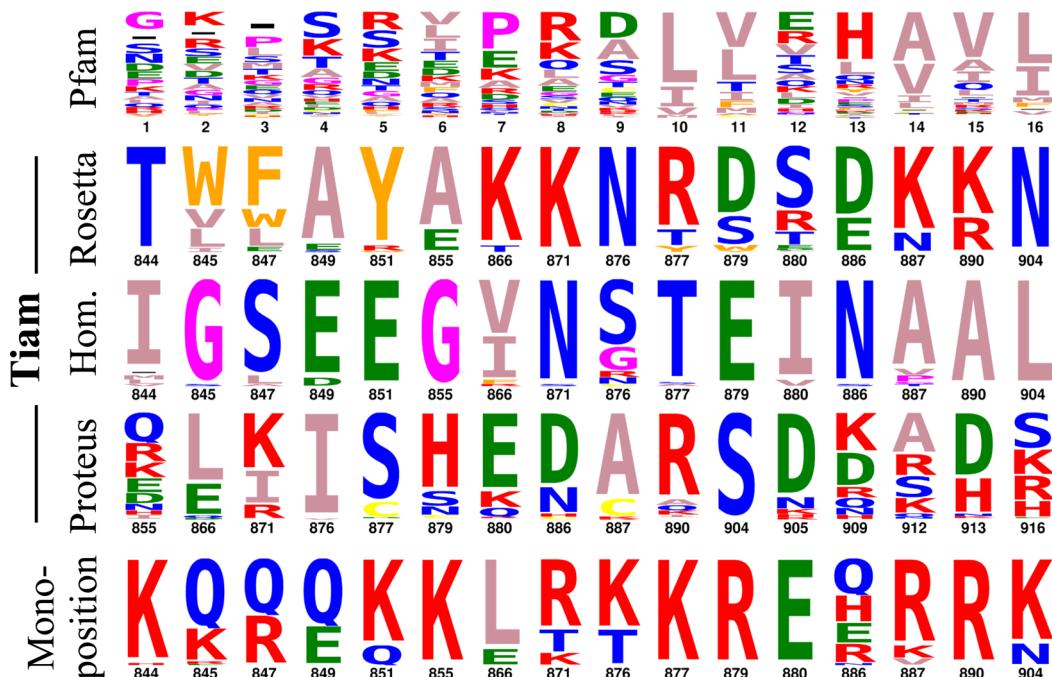


Figure 4. Sequence logos for 16 surface positions in Tiam1. Same representation as Figure 3. The “mono-position” results are from a set of simulations where only one amino acid at a time could mutate, the rest of the protein having its native sequence (see text). Amino acid colors as in Figure 3.

702 was done to allow future experimental testing of designed
703 sequences by a peptide binding assay. Among the 2500 lowest
704 energy designed sequences, we narrowed down the choice of
705 sequences using the following four criteria: (a) sequences
706 should have a nonneutral isoelectric point, (b) they should be
707 assigned to the correct SCOP family by superfamily with good
708 *E*-values, (c) they should have good Pfam similarity scores, and
709 (d) they should have at most 15 mutations that drastically
710 change the amino acid type compared to the wildtype protein
711 (such a change is defined by a Blosum62 similarity score
712 between the two amino acid types of -2 or less). Applying
713 these criteria reduced the number of sequences to 66 from the
714 $\epsilon_p = 8$ model and 45 from the $\epsilon_p = 4$ model. In addition, we
715 eliminated sequences that had two mutations that created a
716 buried cavity and those that had net protein charges of +6 or
717 more (which could lead to protein instability). A total of six
718 sequences were chosen for further analysis. We refer to them as
719 sequences 1–6 or seq-1, ..., seq-6. Sequences 1, 2, 4, and 5 were
720 modified further manually to eliminate charged residues in the
721 exposed loop 852–856 (lysines were changed manually to
722 alanine), giving sequences 1', 2', 4', and 5'. The 10 sequences
723 are shown in Figure 6A. Using these sequences as queries to
724 search Uniprot with Blast, the top hits were either Tiam1
725 mammalian orthologs (including human Tiam1) or uncharac-
726 terized proteins, with identity scores between 35 and 40% and
727 Blast *E*-values of around 10^{-8} – 10^{-7} (except for one sequence
728 which gave hits with lower *E*-values of around 10^{-10}).

729 All 10 sequences were subjected to MD simulations with
730 explicit solvent. Initial simulations were run for 100 ns, with all
731 10 sequences exhibiting good stability. Six were extended to
732 lengths of 500 or 1000 ns. The wildtype protein (WT) was also
733 simulated for 1000 ns. The WT sequence appeared stable over
734 the entire simulation, judging by its rms deviations from the
735 wildtype (WT) X-ray structure and from its own mean MD
736 structure (Figure 6B). The mean MD structure had a backbone

737 rms deviation of 1.0 Å from the WT X-ray structure (excluding
738 3–4 residues at each terminus and one very flexible loop,
739 residues 850–857). During the MD trajectory, the rms
740 deviation from the mean MD structure varied in the range
741 1–1.5 Å, without any visible drift (Figure 6B). A weakly stable
742 quadruple mutant (QM) with an unfolding free energy of just 1
743 kcal/mol⁷⁰ was also simulated for 1000 ns. The mean MD
744 structure of QM had a backbone rms deviation of 1.6 Å relative
745 to the QM X-ray structure (4NXQ). Note that the X-ray
746 structure includes a peptide ligand, whereas the MD simulation
747 represents the apo state. The average MD structure of QM
748 (Figure 6C) exhibited some unwinding of the N-terminus of
749 the α_2 helix. During the QM simulation, the structure had
750 deviations from its average MD structure in the 0.8–1.2 Å
751 range (Figure 6B) and appeared stable.

752 Sequences 1, 3, 4, and 5 were simulated for 500 ns; sequence
753 3 moved away from the mean MD structure toward the end of
754 the simulation; the other three sequences appeared stable
755 (Figure 6B). Sequence 2 (or seq-2) appeared stable up to
756 almost 1000 ns (Figure 6B). The mean seq-2 structure
757 exhibited a shortening of the β strands 2 and 3. Note that in
758 the holo state, strand 2 makes direct contact with the peptide
759 ligand. The rms deviations between the average MD structure
760 of seq-2 and the WT and QM X-ray structures were 1.5 and 1.6
761 Å, respectively. During the seq-2 MD simulation, the rms
762 deviation of seq-2 from its average MD structure varied in the
763 range 1.3–2 Å up to almost 1000 ns. At this point, just before
764 1000 ns, seq-2 underwent a large fluctuation. When the
765 simulation was extended for another 100 ns, the fluctuation
766 largely regressed. More data are needed to determine if this
767 fluctuation signals instability of this designed sequence.

768 Sequence 6 appeared stable throughout microsecond MD
769 simulations (Figure 6B). Its mean MD structure had a
770 backbone rms deviation from the WT X-ray structure of just
771 1.0 Å, the same deviation as the mean WT MD structure. The

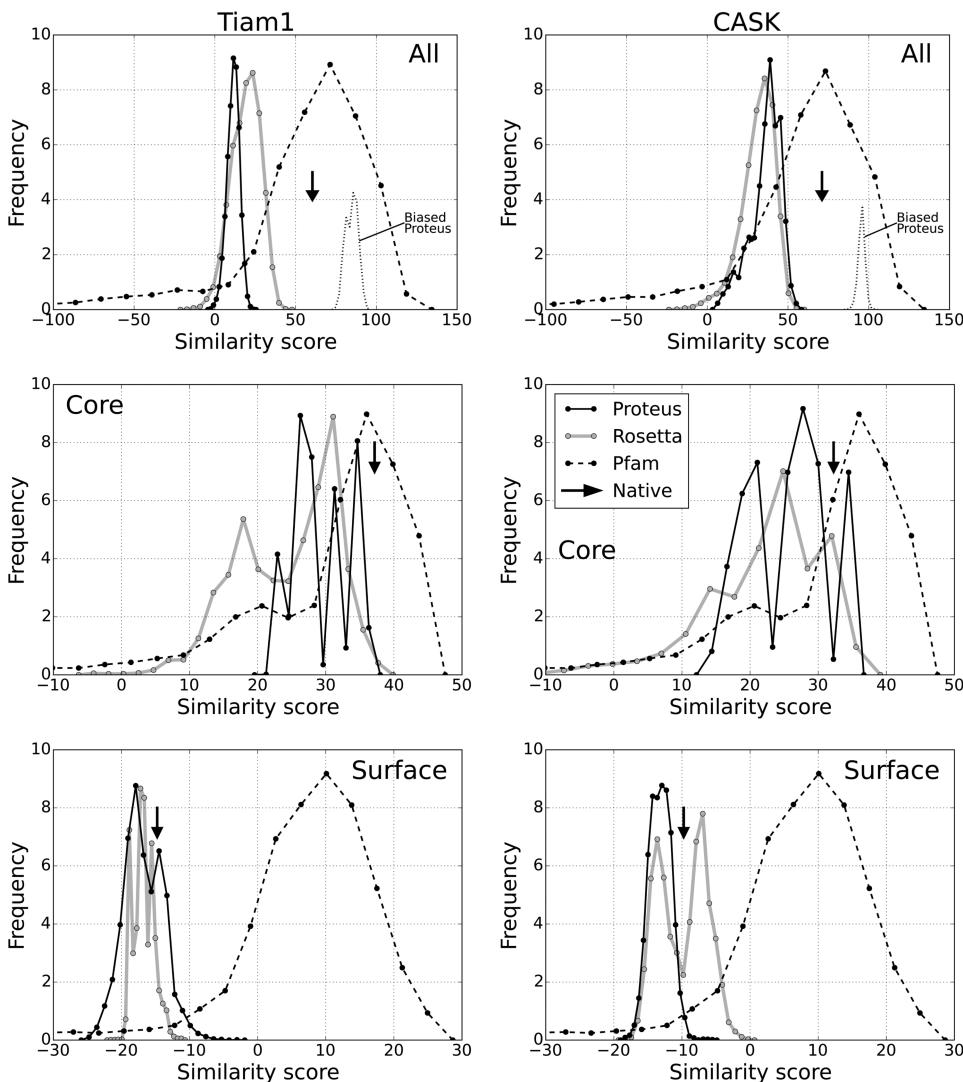


Figure 5. Histogram plots showing similarity scores for designed PDZ sequences. Similarity scores for Tiam1 (left) and Cask (right), relative to the Pfam-RP55 alignment. The scores were computed for all positions (top), 14 core positions (middle), and 16 surface positions (bottom). Values are shown for 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.

mean MD structures of seq-6 and WT are superimposed in Figure 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.

4.6. 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 δ , we effectively “titrate in” hydrophobic residues.

The results for Tiam1 are illustrated in Figure 7. At the largest δ 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 δ values, the hydrophobic core remained native-like. At the most negative δ value, the hydrophobic core became larger, expanding out toward surface regions, with 14 polar positions changed to hydrophobic types. Thus, the numbers of positions changed were approximately symmetric (around ± 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 δ 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

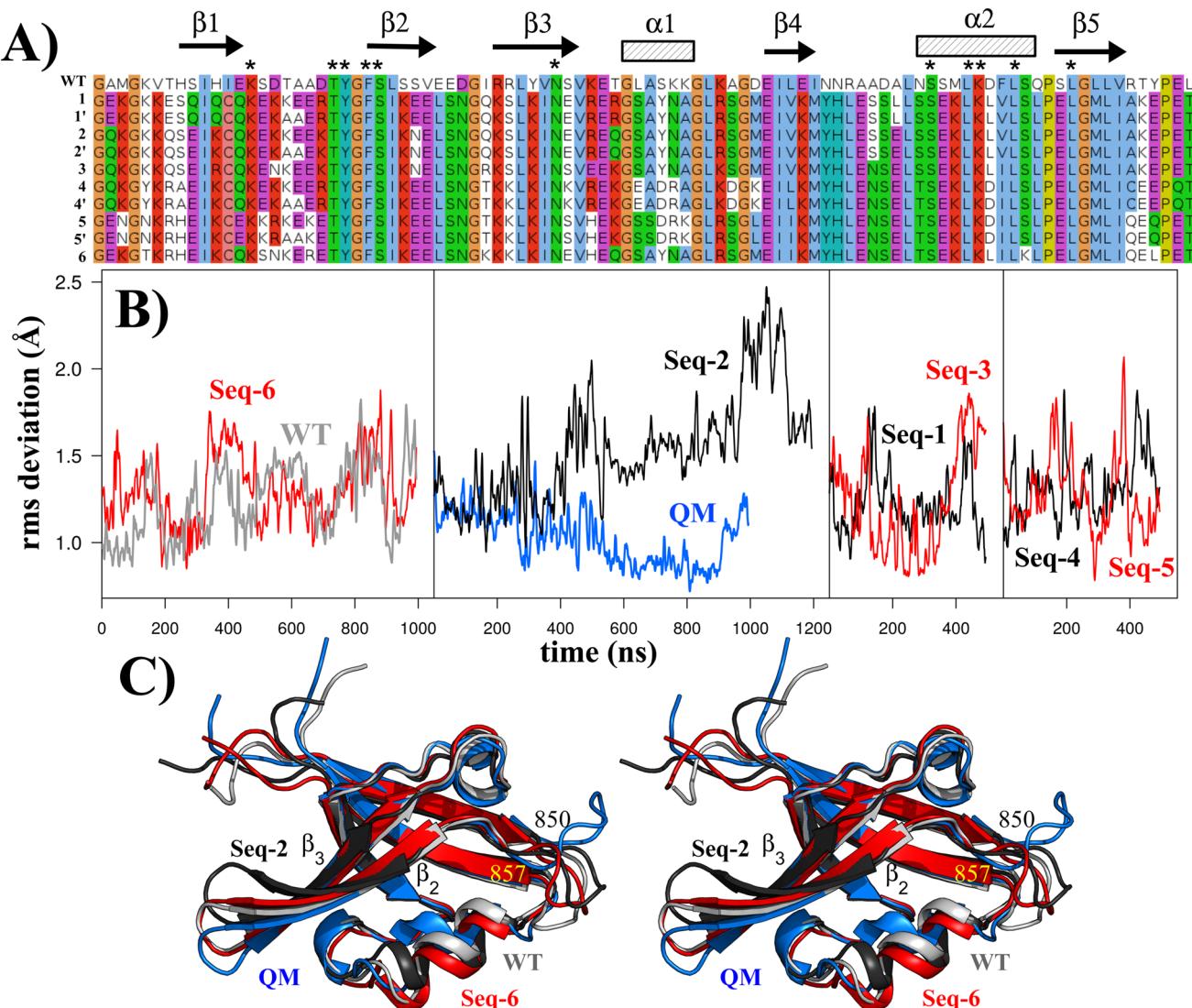


Figure 6. MD simulations of Tiam1 variants designed with Proteus. (A) Sequences of the wildtype (WT) PDZ domain and the 10 designed variants simulated by MD. Asterisks indicate peptide binding residues held fixed during the design simulations. (B) Backbone rms deviations over the course of an MD simulation for WT, QM, and six designed variants relative to the corresponding mean MD structure. (C) Mean MD structures of WT, QM, seq-6, seq-2.

810 hydrophobic at the highest bias level, indicating that these
811 positions also have a certain hydrophobic propensity. Results
812 for Cask were similar, with 11 positions changed to polar at the
813 highest polar bias and 9 changed to hydrophobic at the highest
814 hydrophobic bias.

815 We also derived a parameter to describe the relative number
816 of amino acid type changes per unit bias energy. This parameter
817 was defined as the number δN of residue positions changed
818 from nonpolar to polar, divided by the product of the change
819 δE in bias energy and the mean number N of nonpolar
820 positions at zero bias. We call it the hydrophobic susceptibility,
821 χ_h . For the Tiam1 PDZ domain, this calculation amounts to

$$\chi_h = \frac{\frac{1}{N} \delta N}{\delta E} = 0.88 \text{ changes (per position) per kcal/mol.}$$

For Cask, the susceptibility was $\chi_h = 0.71$ changes per kcal/mol.

4.7. 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 828 peptide relative to a syndecan-1 (Sdc1) peptide.^{70,73} The four 829 mutations in the QM PDZ domain were L911M, K912E, 830 L915F, and L920 V. All four positions but Lys912 are part of 831 the conserved hydrophobic core. The four single and two 832 double mutants were also characterized experimentally.^{70,73} For 833 simplicity, we denote the native (WT) sequence as LKLL and 834 the quadruple mutant (QM) as MEFV. Other variants are 835 denoted similarly. Replica Exchange MC simulations were 836 conducted on several structural templates, where all four 837 positions could mutate simultaneously, into all amino acid types 838 except Gly and Pro. We used the Proteus model with the lower 839 dielectric constant, $\epsilon_p = 4$, which gave similarity scores 840 equivalent to the $\epsilon_p = 8$ model but had a reduced tendency 841 to bury polar side chains, thanks to its lower dielectric constant. 842 In addition, no bias energy term was used, only the MMGBSA 843 energy function. The CPD model used either the wildtype or 844 the quadruple mutant crystal structure as backbone template 845 for the design (PDB codes 4GVD and 4NXQ, respectively), 846 shown in Figure 8. Although these two structures were 847 88

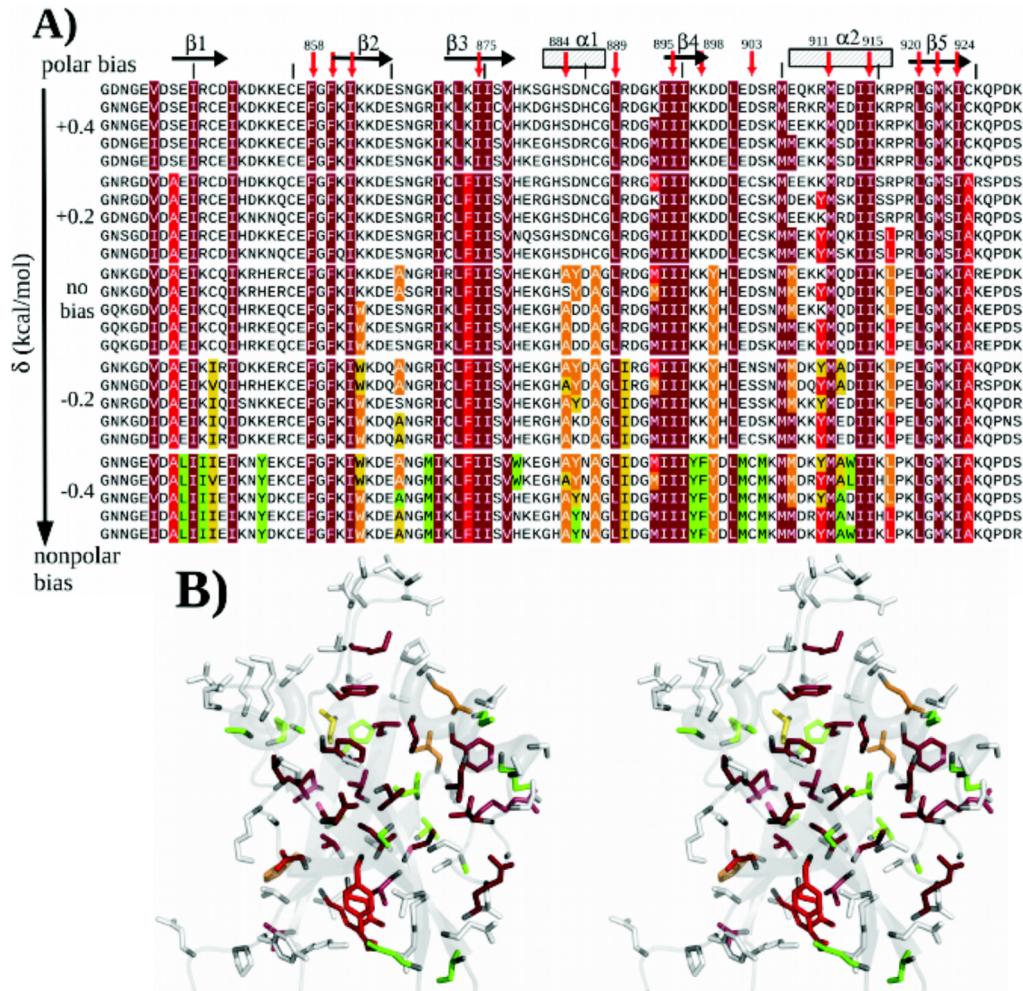


Figure 7. (A) Tiam1 sequences designed with different levels of hydrophobic bias. The top (respectively, bottom) sequences were obtained in the presence of a bias δ opposing (favoring) hydrophobic types. The middle sequences were obtained from Proteus simulations without any bias ($\delta = 0$). For each 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 residue colors as in (A). The backbone is shown in light gray.

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 α_2 . This helix is pushed slightly outward in the mutant complex, to accommodate 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 side chain.

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 Figure 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 homologue was the second best sequence overall, and the homologues IKLL and LKLV were just 1–2 kcal/mol higher in energy. The mutant sequence MEFV did not appear, nor did any close homologues, probably because the wildtype backbone structure cannot accommodate 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

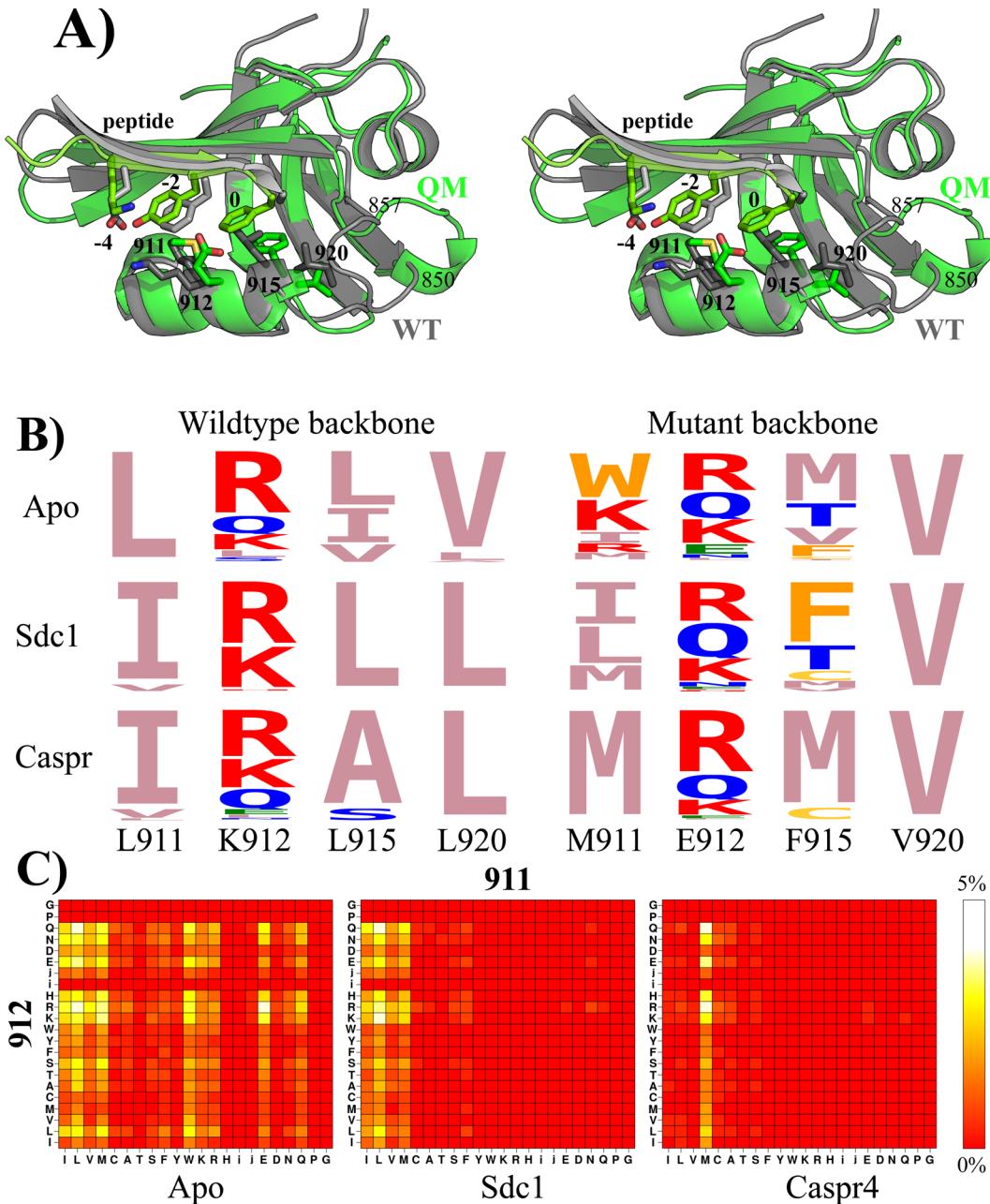


Figure 8. Design of four Tiam1 specificity positions. (A) X-ray structures (stereo) with the wildtype sequence (LKLL; labeled WT; PDB code 4GVD) or the quadruple mutant sequence (MEFV; labeled QM; PDB code 4NXQ), with bound Sdc1 and Caspr4, respectively. The four designed side chains are shown and labeled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

energy for Caspr4, and stabilities just 2 kcal/mol apart, suggesting that both should be sampled. Instead, neither was sampled. The closest homologue 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 homologue 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 homologues MQMV and MEMV appeared at +5 kcal/mol. The wildtype LKLL and its close homologues 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 ∈ {I,L,V,M}), the mean apo and Sdc1-bound energies are 0.9 ± 0.6 and 1.1 ± 0.5 kcal/mol, respectively, which leads to a mean affinity of 0.2 ± 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 → 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 → 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 → 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 → MEFV mutation favors Caspr4, relative to Sdc1, by 0.2 kcal/mol, compared to 0.5 kcal/mol experimentally for the homologous LKLL → LELL mutation.

The simulations also gave information on correlations between the four mutating positions. Figure 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 among 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 parametrized 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 σ_i and two protein dielectric constants ϵ_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 side chains, we used a simple, discrete rotamer library and a short minimization of each side chain 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 side chain reconstruction tests⁵⁴ (comparable to the popular Scwrl4 program⁵³) and good performance for a large set of protein acid/base constants⁷⁴ (superior to the Rosetta software,⁷⁵ despite extensive *ad hoc* parameter tuning).⁹⁸³

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 homologues. 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 parametrization more robust and less sensitive to the size and structure of the natural homologues 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 ϵ_p . The performance levels were similar to 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.¹⁰²⁶

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 1 ($\epsilon_p = 4$ 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 side chain 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

1041 Tiam1 specificity positions, where the designed sequences were
1042 sensitive to the particular backbone conformation of the protein
1043 and peptide. Specifically, with the wildtype backbone structure,
1044 there was no room to insert a Phe side chain at position 915,
1045 even though Phe915 is present in the experimental quadruple
1046 mutant (which has a slightly different backbone structure).
1047 Therefore, the choice of the initial X-ray structural model is
1048 important, and several strategies are possible. Here, to
1049 parametrize the CPD model, we used X-ray structures solved
1050 with a peptide ligand, even though the parametrization
1051 simulations and most of the testing were done for the apo
1052 proteins. This choice was made partly because the apo/holo
1053 PDZ structures are quite similar and partly to make the model
1054 more transferable and facilitate applications to peptide binding.
1055 Another strategy could have been to parametrize the model
1056 using all apo structures, then switch to holo structures for the
1057 Tiam1 application.

1058 For whole protein design (such as the hydrophobic titration
1059 application), the use of a fixed backbone can be partly
1060 counterbalanced by designing two or more PDZ structures. For
1061 example, pooling the designed Tiam1 and Cask sequences gave
1062 a mean sequence entropy comparable to the experimental Pfam
1063 set, and allowed us to recapitulate more sequences than design
1064 with just one backbone. In the application to Tiam1 4-position
1065 design, the fixed backbone was also counterbalanced by doing
1066 calculations separately with two different backbone structures, a
1067 holo wildtype and a holo mutant structure. Simulations with the
1068 mutant backbone allowed us to obtain mutants having Phe at
1069 position 915. Notice that a new method for multibackbone
1070 design was recently developed in Proteus, based on a novel,
1071 nonheuristic hybrid Monte Carlo method that preserves
1072 Boltzmann sampling.³⁵ This method could be applied in the
1073 future.

1074 A third limitation of our model is the need, for optimal
1075 results, to parametrize the reference energies specifically for a
1076 given set of proteins. This step is well-automated and highly
1077 parallel. However, it involves several choices that are partly
1078 arbitrary. These include the choice of a set of protein domains
1079 to represent the protein or family of interest. We also need to
1080 choose a similarity threshold to define the target homologues
1081 from which we compute the experimental amino acid
1082 compositions. Here, we chose to use close homologues of
1083 each family member, compute their compositions, then average
1084 over the two family representatives. This method worked well
1085 but other choices are possible, and more work is needed to
1086 draw definitive conclusions. Also, the monoposition design of
1087 Tiam1 showed evidence of some systematic error (Figure 4),
1088 with a large fraction of Arg, Lys, and Gln residues types on the
1089 protein surface, despite the optimized reference energies. In the
1090 future, it may be necessary to relax the intragroup constraints
1091 toward the end of the reference energy optimization and/or
1092 target smaller numbers of mutating positions, instead of one-
1093 half of the protein at a time.

1094 A fourth limitation of our model is the discrete rotamer
1095 approximation, which requires some adaptation of the energy
1096 function to avoid exaggerated steric clashes; the method used
1097 here is the residue-pair minimization method described
1098 earlier.^{26,76} A fifth limitation is the use of a pairwise additive
1099 solvation model (as in most CPD models). Specifically, the
1100 dielectric environment of each residue pair is assumed here to
1101 be native-like (so-called “Native Environment approximation”
1102 or NEA^{74,76}). This leads to an energy function that has the
1103 form of a sum over pairs of residues and that can be

1104 precalculated and stored in an energy matrix, which then serves
1105 as a lookup table during the subsequent Monte Carlo
1106 simulations. Despite this approximation, the model gave good
1107 results for a large acid/base benchmark,⁷⁴ a problem that is very
1108 sensitive to the electrostatic treatment.¹¹⁰⁸

1109 Some of these limitations could be removed in a future work.
1110 In particular, since the energy function is mostly physics-based,
1111 it can benefit rapidly from ongoing improvements in protein
1112 force fields and solvation models. Thus, the NEA approx-
1113 imation could be removed in the future due to the recent
1114 implementation (manuscript in preparation) of a more exact
1115 Generalized Born calculation, whose efficiency is comparable to
1116 the pairwise approximation.⁷⁹ We have also implemented an
1117 improved model for hydrophobic solvation,⁸⁰ which is faster
1118 and more accurate than our current surface area energy term
1119 (manuscript in preparation).

1120 **5.2. Model Testing and Applications.** Designed
1121 sequences were extensively compared to natural sequences,
1122 through fold recognition tests, similarity calculations, and
1123 entropy calculations. In the test simulations, we designed the
1124 entire protein sequence, so that all positions (except Gly and
1125 Pro) could mutate freely, subject only to an overall bias toward
1126 the mean, experimental amino acid composition (through the
1127 reference energies). Despite the lack of experimental bias or
1128 constraints, the resulting sequences had a high overall similarity
1129 to the natural, Pfam sequences, as measured by the Blosum40
1130 similarity scores. The scores obtained were mostly comparable
1131 to the similarity scores between pairs of Pfam sequences
1132 themselves. Thus, the sequences designed with Proteus
1133 resemble moderately distant natural homologues. The similarity
1134 was very strong for residues in the core of the protein, as
1135 observed in previous CPD studies.^{30,72} In contrast, for residues
1136 at the protein surface, similarity scores were close to zero, the
1137 score one would obtain if one picked amino acid types
1138 randomly. Notice that many surface residues are involved in
1139 functional interactions, such as the 11 peptide-binding residues
1140 in PDZ domains. Surface residues are also selected by evolution
1141 to avoid aggregation or unwanted adhesion. Most of these
1142 functional constraints are not explicitly accounted for in our
1143 design protocol (although the energy function indirectly favors
1144 protein solubility). Despite the limited similarity scores for
1145 surface regions, fold recognition with the superfamily tool and
1146 the best design models was almost perfect. Earlier fold
1147 recognition tests that used a simpler energy function gave a
1148 lower fold recognition rate of about 85% (for a larger and more
1149 diverse test set) and lower similarities.^{15,50} Evidently, the
1150 combined use of an improved protein force field, Generalized
1151 Born solvent, and family specific reference energies leads to
1152 designed sequences that are more native-like and presumably
1153 better.

1154 The Proteus sequences were also compared to sequences
1155 designed with the Rosetta software (using default parameters),
1156 which has itself been extensively tested. On the basis of the
1157 Blosum similarity scores (vs natural sequences in Pfam) and the
1158 fold recognition tests, the Proteus and Rosetta sequences
1159 appear to be of about the same quality. However, Rosetta
1160 makes fewer mutations than Proteus, so that the identity scores,
1161 compared to the corresponding wildtype protein, are about 6%
1162 points higher. This means that Proteus mutates about five more
1163 positions, on average, per PDZ domain. This number could
1164 easily be reduced, by adding to the Proteus energy function an
1165 explicit bias energy term that increases with the number of
1166 mutations (away from the wildtype sequence). An equivalent

1167 bias energy was used above for just two simulations of Tiam1
1168 and Cask (see the “biased Proteus” results in the two upper
1169 panels of [Figure 5](#)), to illustrate the possibility of using
1170 experimentally restrained sampling. It remains to be seen
1171 whether a restraint based on the identity score would lead to
1172 more stable and realistic designed sequences.

1173 Another attractive route for testing designed sequences is
1174 through high-level MD simulations. Here, 10 designed Tiam1
1175 sequences were tested in rather long MD simulations, in the
1176 apo form, using the same protein force field as in the CPD
1177 model (Amber force field) and an established explicit water
1178 model. These sequences were designed using Proteus, with Gly,
1179 Pro, and 11 peptide-binding positions held fixed but all others
1180 free to mutate. Six of the sequences remained stable over 200
1181 ns simulation lengths and two were extended up to a
1182 microsecond, which represents a very stringent test of the
1183 designs. Sequence 6 was stable over the whole microsecond.
1184 The mean deviation of seq-6 from the starting, experimental
1185 wildtype structure was 1 Å, which is the same as the mean
1186 deviation in the MD simulation of the wildtype sequence itself.
1187 The deviation between the mean sequence 6 MD structure and
1188 the mean wildtype MD structure was also small, just 1.2 Å.
1189 Sequence 2 was also simulated and remained stable until just
1190 before the end of the microsecond simulation, at which point it
1191 underwent a larger fluctuation. The fluctuation regressed 100
1192 ns later. An even longer simulation would be needed to
1193 determine if this fluctuation is harmless or signals the beginning
1194 of domain unfolding. Note that in the presence of a peptide
1195 ligand, we expect the structural stability of the designed
1196 domains would increase further. MD simulations of additional
1197 designed sequences would also be of interest. Direct
1198 experimental testing of the designed proteins remains to be
1199 done.

1200 The CPD model was used for two applications. “Hydro-
1201 phobic titration” of two PDZ domains illustrated a novel way to
1202 characterize protein designability. The cost or availability of
1203 hydrophobic side chain types was controlled by a bias energy
1204 term that was gradually varied. The mean overall hydrophobic
1205 “susceptibility”, the number of type of changes per kcal/mol
1206 and per position, differed by about 20% between Tiam1 and
1207 Cask. In Tiam1, 11 of the core positions remained hydrophobic
1208 even with the largest bias value favoring polar types, while 14
1209 other positions changed to nonpolar types at the largest
1210 nonpolar bias energy value. A comparison to other domain
1211 families would be of interest, and is left for future work.

1212 Redesign of four specificity positions in Tiam1 allowed us to
1213 test the design model in a different way. It revealed some of the
1214 limitations of fixed backbone design, but also gave semi-
1215 quantitative agreement with available binding free energies.
1216 This agreement predicts new mutations that could be of
1217 interest for obtaining new specificity properties. They remain to
1218 be studied further and tested experimentally. Here, the apo and
1219 two holo states were studied, and designed separately.
1220 Information about binding affinities and binding specificity
1221 were then obtained by comparing the energies sampled in the
1222 different simulations. In the future, we would like to include
1223 binding affinity and/or specificity directly in the design
1224 calculations, as a property to be designed for or against within
1225 a single simulation. In addition, we should allow different
1226 backbone structures for the apo and each holo system. This
1227 could be done in the future, since recent hybrid Monte Carlo
1228 schemes^{35,81} can be used for multibackbone design, and can be
1229 extended to the problem of designing ligand binding specificity.

We also note that since our energy function is physics-based, it
1230 has transferability to a range of molecule types, such as
1231 nonnatural amino acids (considered in an earlier protein–
1232 ligand design study³⁴). Such amino acids could be of interest
1233 for designing PDZ peptide ligands, to provide additional
1234 diversity and perhaps enhanced resistance to proteolysis.
1235

■ ASSOCIATED CONTENT

S Supporting Information

The Supporting Information is available free of charge on the
1238 ACS Publications website at DOI: [10.1021/acs.jctc.6b01255](https://doi.org/10.1021/acs.jctc.6b01255).
1239

Detailed description of the model cross validation and
1240 results obtained with a protein dielectric $\epsilon_p = 4$ ([PDF](#))
1241

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: thomas.simonson@polytechnique.fr.

ORCID

Nicolas Panel: [0000-0001-8782-0586](https://orcid.org/0000-0001-8782-0586)

Thomas Simonson: [0000-0002-5117-7338](https://orcid.org/0000-0002-5117-7338)

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We are grateful for discussions with Michael Schnieders and
1251 Young Joo Sun (University of Iowa). Some of the calculations
1252 were done at the CINES supercomputer center of the French
1253 Ministry of Research. NAMD was developed by the Theoretical
1254 and Computational Biophysics Group in the Beckman Institute
1255 at the University of Illinois at Urbana.
1256

■ REFERENCES

- (1) Harris, B. Z.; Lim, W. A. Mechanism and role of PDZ domains in signaling complex assembly. *J. Cell Sci.* **2001**, *114*, 3219–3231.
- (2) Hung, A. Y.; Sheng, M. PDZ Domains: Structural Modules for Protein Complex Assembly. *J. Biol. Chem.* **2002**, *277*, 5699–5702.
- (3) Tonikian, R.; Zhang, Y. N.; Sazinsky, S. L.; Currell, B.; Yeh, J. H.; Reva, B.; Held, H. A.; Appleton, B. A.; Evangelista, M.; Wu, Y.; Xin, X.; Chan, A. C.; Seshagiri, S.; Lasky, L. A.; Sander, C.; Boone, C.; Bader, G. D.; Sidhu, S. S. A Specificity Map for the PDZ Domain Family. *PLoS Biol.* **2008**, *6*, 2043–2059.
- (4) Gfeller, D.; Butty, F.; Wierzbicka, M.; Verschueren, E.; Vanhee, P.; Huang, H.; Ernst, A.; Darand, N.; Stagliar, I.; Serrano, L.; Sidhu, S.; Bader, G. D.; Kim, P. M. The multiple-specificity landscape of modular peptide recognition domains. *Mol. Syst. Biol.* **2011**, *7*, art484.
- (5) Subbaiah, V. K.; Kranjec, C.; Thomas, M.; Banks, L. PDZ domains: the building blocks regulating tumorigenesis. *Biochem. J.* **2011**, *439*, 195–205.
- (6) Shepherd, T. R.; Fuentes, E. J. Structural and thermodynamic analysis of PDZ-ligand interactions. *Methods Enzymol.* **2011**, *488*, 81–100.
- (7) Bacha, A.; Clausen, B. H.; Moller, M.; Vestergaard, B.; Chic, C. N.; Round, A.; Sørensen, P. L.; Nissen, K. B.; Kastrup, J. S.; Gajhede, M.; Jemth, P.; Kristensen, A. S.; Lundstrom, P.; Lambertsen, K. L.; Stromgaard, K. A high-affinity, dimeric inhibitor of PSD-95 bivalently interacts with PDZ1–2 and protects against ischemic brain damage. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 3317–3322.
- (8) Roberts, K. E.; Cushing, P. R.; Boisguerin, P.; Madden, D. R.; Donald, B. R. Computational Design of a PDZ Domain Peptide Inhibitor that Rescues CFTR Activity. *PLoS Comput. Biol.* **2012**, *8*, e1002477.
- (9) Zheng, F.; Jewell, H.; Fitzpatrick, J.; Zhang, J.; Mierke, D. F.; Grigoryan, G. Computational Design of Selective Peptides to

- 1289 Discriminate between Similar PDZ Domains in an Oncogenic
1290 Pathway. *J. Mol. Biol.* **2015**, *427*, 491–510.
- 1291 (10) Lockless, W.; Ranganathan, R. Evolutionary Conserved
1292 Pathways of Energetic Connectivity in Protein Families. *Science*
1293 **1999**, *285*, 295–299.
- 1294 (11) Kong, Y.; Karplus, M. Signaling pathways of PDZ2 domain: A
1295 molecular dynamics interaction correlation analysis. *Proteins: Struct.,
1296 Funct., Genet.* **2009**, *74*, 145–154.
- 1297 (12) McLaughlin, R. N., Jr.; Poelwijk, F. J.; Raman, A.; Gosal, W. S.;
1298 Ranganathan, R. The spatial architecture of protein function and
1299 adaptation. *Nature* **2012**, *491*, 138–142.
- 1300 (13) Melero, C.; Ollikainen, N.; Harwood, I.; Karpik, J.; Kortemme,
1301 T. Quantification of the transferability of a designed protein specificity
1302 switch reveals extensive epistasis in molecular recognition. *Proc. Natl.
1303 Acad. Sci. U. S. A.* **2014**, *111*, 15426–15431.
- 1304 (14) Reina, J.; Lacroix, E.; Hobson, S. D.; Fernandez-Ballester, G.;
1305 Rybin, V.; Schwab, M. S.; Serrano, L.; Gonzalez, C. Computer-aided
1306 design of a PDZ domain to recognize new target sequences. *Nat.
1307 Struct. Biol.* **2002**, *9*, 621–627.
- 1308 (15) Schmidt am Busch, M.; Sedano, A.; Simonson, T. Computational
1309 protein design: validation and possible relevance as a tool for
1310 homology searching and fold recognition. *PLoS One* **2010**, *5* (5),
1311 e10410.
- 1312 (16) Smith, C. A.; Kortemme, T. Structure-Based Prediction of the
1313 Peptide Sequence Space Recognized by Natural and Synthetic PDZ
1314 Domains. *J. Mol. Biol.* **2010**, *402*, 460–474.
- 1315 (17) Butterfoss, G. L.; Kuhlman, B. Computer-based design of novel
1316 protein structures. *Annu. Rev. Biophys. Biomol. Struct.* **2006**, *35*, 49–65.
- 1317 (18) Lippow, S. M.; Tidor, B. Progress in computational Protein
1318 Design. *Curr. Opin. Biotechnol.* **2007**, *18*, 305–311.
- 1319 (19) Dai, L.; Yang, Y.; Kim, H. R.; Zhou, Y. Improving computational
1320 protein design by using structure-derived sequence profile. *Proteins:
1321 Struct., Funct., Genet.* **2010**, *78*, 2338–2348.
- 1322 (20) Feldmeier, K.; Hoecker, B. Computational protein design of
1323 ligand binding and catalysis. *Curr. Opin. Chem. Biol.* **2013**, *17*, 929–
1324 933.
- 1325 (21) Tinberg, C. E.; Khare, S. D.; Dou, J.; Doyle, L.; Nelson, J. W.;
1326 Schena, A.; Jankowski, W.; Kalodimos, C. G.; Johnsson, K.; Stoddard,
1327 B. L.; Baker, D. Computational design of ligand-binding proteins with
1328 high affinity and selectivity. *Nature* **2013**, *501*, 212–218.
- 1329 (22) Au, L.; Green, D. F. Direct Calculation of Protein Fitness
1330 Landscapes through Computational Protein Design. *Biophys. J.* **2016**,
1331 *110*, 75–84.
- 1332 (23) Pokala, N.; Handel, T. Energy functions for protein design I:
1333 Efficient and accurate continuum electrostatics and solvation. *Protein
1334 Sci.* **2004**, *13*, 925–936.
- 1335 (24) Samish, I.; MacDermaid, C. M.; Perez-Aguilar, J. M.; Saven, J. G.
1336 Theoretical and computational protein design. *Annu. Rev. Phys. Chem.*
1337 **2011**, *62*, 129–149.
- 1338 (25) Li, L.; Francklyn, C.; Carter, C. W. Aminoacylating Urzymes
1339 Challenge the RNA World Hypothesis. *J. Biol. Chem.* **2013**, *288*,
1340 26856–26863.
- 1341 (26) Schmidt am Busch, M.; Lopes, A.; Mignon, D.; Simonson, T.
1342 Computational protein design: software implementation, parameter
1343 optimization, and performance of a simple model. *J. Comput. Chem.*
1344 **2008**, *29*, 1092–1102.
- 1345 (27) Simonson, T. Protein:ligand recognition: simple models for
1346 electrostatic effects. *Curr. Pharm. Des.* **2013**, *19*, 4241–4256.
- 1347 (28) Polydorides, S.; Michael, E.; Mignon, D.; Druart, K.; Archontis,
1348 G.; Simonson, T. In *Methods in Molecular Biology: Design and Creation
1349 of Protein Ligand Binding Proteins*; Stoddard, B., Ed.; Springer Verlag:
1350 New York, 2016; Vol. 1414; p 0000.
- 1351 (29) Kuhlman, B.; Baker, D. Native protein sequences are close to
1352 optimal for their structures. *Proc. Natl. Acad. Sci. U. S. A.* **2000**, *97*,
1353 10383–10388.
- 1354 (30) Dantas, G.; Kuhlman, B.; Callender, D.; Wong, M.; Baker, D. A
1355 Large Test of Computational Protein Design: Folding and Stability of
1356 Nine Completely Redesigned Globular Proteins. *J. Mol. Biol.* **2003**,
1357 *332*, 449–460.
- 1358 (31) Rohl, C. A.; Strauss, C. E. M.; S, M. K. M.; Baker, D. Protein
1359 structure prediction using Rosetta. *Methods Enzym.* **2004**, *383*, 10383–
1360 103866–93.
- 1361 (32) Lazaridis, T.; Karplus, M. Effective energy function for proteins
1362 in solution. *Proteins: Struct., Funct., Genet.* **1999**, *35*, 133–152.
- 1363 (33) Mignon, D.; Simonson, T. Comparing three stochastic search
1364 algorithms for computational protein design: Monte Carlo, Replica
1365 Exchange Monte Carlo, and a multistart, steepest-descent heuristic. *J.
1366 Comput. Chem.* **2016**, *37*, 1781–1793.
- 1367 (34) Druart, K.; Palmai, Z.; Omarjee, E.; Simonson, T. Protein:ligand
1368 binding free energies: a stringent test for computational protein design.
1369 *J. Comput. Chem.* **2016**, *37*, 404–415.
- 1370 (35) Druart, K.; Bigot, J.; Audit, E.; Simonson, T. A hybrid Monte
1371 Carlo method for multibackbone protein design. *J. Chem. Theory
1372 Comput.* **2016**, *12*, 6035–6048.
- 1373 (36) Gaillard, T.; Simonson, T. Full protein sequence redesign with
1374 an MMGBSA energy function. *J. Comput. Chem.* **2017**.
- 1375 (37) Baker, D. Prediction and design of macromolecular structures
1376 and interactions. *Philos. Trans. R. Soc., B* **2006**, *361*, 459–463.
- 1377 (38) Frenkel, D.; Smit, B. *Understanding molecular simulation*; Academic
1378 Press: New York, 1996; Chapter 3.
- 1379 (39) Grimmett, G. R.; Stirzaker, D. R. *Probability and random
processes*; Oxford University Press: Oxford, United Kingdom, 2001.
- 1380 (40) Kleinman, C. L.; Rodrigue, N.; Bonnard, C.; Philippe, H.;
Lartillot, N. A maximum likelihood framework for protein design. *BMC Bioinf.*
1382 **2006**, *7*, Art326.
- 1383 (41) Fowler, R. H.; Guggenheim, E. A. *Statistical Thermodynamics*; Cambridge
1384 University Press: Cambridge, United Kingdom, 1939.
- 1385 (42) Cornell, W.; Cieplak, P.; Bayly, C.; Gould, I.; Merz, K.; Ferguson,
1386 D.; Spellmeyer, D.; Fox, T.; Caldwell, J.; Kollman, P. A. Second Generation Force Field for the Simulation of Proteins, Nucleic
1388 Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179–
1389 S197.
- 1390 (43) Aleksandrov, A.; Polydorides, S.; Archontis, G.; Simonson, T.
1391 Predicting the Acid/Base Behavior of Proteins: A Constant-pH Monte
1392 Carlo Approach with Generalized Born Solvent. *J. Phys. Chem. B* **2010**,
1393 *114*, 10634–10648.
- 1394 (44) Hawkins, G. D.; Cramer, C.; Truhlar, D. Pairwise descreening of
1395 solute charges from a dielectric medium. *Chem. Phys. Lett.* **1995**, *246*,
1396 122–129.
- 1397 (45) Lopes, A.; Alexandrov, A.; Bathelt, C.; Archontis, G.; Simonson, T.
1398 Computational sidechain placement and protein mutagenesis with
1399 implicit solvent models. *Proteins: Struct., Funct., Genet.* **2007**, *67*, 853–
1400 867.
- 1401 (46) Lee, B.; Richards, F. The interpretation of protein structures:
1402 estimation of static accessibility. *J. Mol. Biol.* **1971**, *55*, 379–400.
- 1403 (47) Brünger, A. T. *X-PLOR version 3.1, A System for X-ray
1404 crystallography and NMR*; Yale University Press: New Haven, 1992.
- 1405 (48) Gaillard, T.; Simonson, T. Pairwise Decomposition of an
1406 MMGBSA Energy Function for Computational Protein Design. *J. Comput. Chem.* **2014**, *35*, 1371–1387.
- 1408 (49) Street, A. G.; Mayo, S. Pairwise calculation of protein solvent-
1409 accessible surface areas. *Folding Des.* **1998**, *3*, 253–258.
- 1410 (50) Schmidt am Busch, M.; Mignon, D.; Simonson, T. Computational
1411 protein design as a tool for fold recognition. *Proteins: Struct.,
1412 Funct., Genet.* **2009**, *77*, 139–158.
- 1413 (51) Eswar, N.; Marti-Renom, M. A.; Webb, B.; Madhusudhan, M. S.;
1414 Eramian, D.; Shen, M.; Pieper, U.; Sali, A. Comparative Protein
1415 Structure Modeling With MODELLER. *Curr. Prot. Bioinf.* **2006**, Suppl.
1416 *15*, S.6.1–S.6.30.
- 1417 (52) Tuffery, P.; Etchebest, C.; Hazout, S.; Lavery, R. A New
1418 Approach to the Rapid Determination of Protein Side Chain
1419 Conformations. *J. Biomol. Struct. Dyn.* **1991**, *8*, 1267–1289.
- 1420 (53) Krivov, G. G.; Shapalov, M. V.; Dunbrack, R. L. Improved
1421 prediction of protein side-chain conformations with SCWRL4.
1422 *Proteins: Struct., Funct., Genet.* **2009**, *77*, 778–795.
- 1423 (54) Gaillard, T.; Panel, N.; Simonson, T. Protein sidechain
1424 conformation predictions with an MMGBSA energy function. *Proteins:
1425 Struct., Funct., Genet.* **2016**, *84*, 803–819.

- 1427 (55) Kofke, D. A. On the acceptance probability of replica-exchange
1428 Monte Carlo trials. *J. Chem. Phys.* **2002**, *117*, 6911–6914.
- 1429 (56) Earl, D.; Deem, M. W. Parallel tempering: theory, applications,
1430 and new perspectives. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3910–3916.
- 1431 (57) Gough, J.; Karplus, K.; Hughey, R.; Chothia, C. Assignment of
1432 homology to genome sequences using a library of hidden Markov
1433 models that represent all proteins of known structure. *J. Mol. Biol.*
1434 **2001**, *313*, 903–919.
- 1435 (58) Wilson, D.; Madera, M.; Vogel, C.; Chothia, C.; Gough, J. The
1436 SUPERFAMILY database in 2007: families and functions. *Nucleic
1437 Acids Res.* **2007**, *35*, D308–D313.
- 1438 (59) Andreeva, A.; Howorth, D.; Brenner, S. E.; Hubbard, J. J.;
1439 Chothia, C.; Murzin, A. G. SCOP database in 2004: refinements
1440 integrate structure and sequence family data. *Nucleic Acids Res.* **2004**,
1441 *32*, D226–229.
- 1442 (60) Durbin, R.; Eddy, S. R.; Krogh, A.; Mitchison, G. *Biological
1443 sequence analysis*; Cambridge University Press: Cambridge, United
1444 Kingdom, 2002.
- 1445 (61) Murphy, L. R.; Wallqvist, A.; Levy, R. M. Simplified amino acid
1446 alphabets for protein fold recognition and implications for folding.
1447 *Protein Eng., Des. Sel.* **2000**, *13*, 149–152.
- 1448 (62) Launay, G.; Mendez, R.; Wodak, S. J.; Simonson, T.
1449 Recognizing protein-protein interfaces with empirical potentials and
1450 reduced amino acid alphabets. *BMC Bioinf.* **2007**, *8*, 270–291.
- 1451 (63) Jo, S.; Kim, T.; Iyer, V. G.; Im, W. CHARMM-GUI: a web-
1452 based graphical user interface for CHARMM. *J. Comput. Chem.* **2008**,
1453 *29*, 1859–1865.
- 1454 (64) Nose, S. A unified formulation of the constant temperature
1455 molecular dynamics method. *J. Chem. Phys.* **1984**, *81*, 511–519.
- 1456 (65) Hoover, W. G. Canonical dynamics: equilibrium phase-space
1457 distributions. *Phys. Rev. A: At., Mol., Opt. Phys.* **1985**, *31*, 1695–1697.
- 1458 (66) Darden, T. In *Computational Biochemistry & Biophysics*; Becker,
1459 O., Mackerell, A., Jr., Roux, B., Watanabe, M., Eds.; Marcel Dekker:
1460 N.Y., 2001; Chapter 4.
- 1461 (67) Jorgensen, W.; Chandrasekhar, J.; Madura, J.; Impey, R.; Klein,
1462 M. Comparison of simple potential functions for simulating liquid
1463 water. *J. Chem. Phys.* **1983**, *79*, 926–935.
- 1464 (68) Brooks, B.; Brooks, C. L., III; Mackerell, A. D., Jr.; Nilsson, L.;
1465 Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch,
1466 S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.;
1467 Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.;
1468 Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer,
1469 M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.;
1470 York, D. M.; Karplus, M. CHARMM: The biomolecular simulation
1471 program. *J. Comput. Chem.* **2009**, *30*, 1545–1614.
- 1472 (69) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid,
1473 E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. Scalable
1474 molecular dynamics with NAMD. *J. Comput. Chem.* **2005**, *26*, 1781–
1475 1802.
- 1476 (70) Liu, X.; Speckhard, D. C.; Shepherd, T. R.; Sun, Y. J.; Hengel, S.
1477 R.; Yu, L.; Fowler, C. A.; Gakhar, L.; Fuentes, E. J. Distinct roles for
1478 conformational dynamics in protein-ligand interactions. *Structure*
1479 **2016**, *24*, 2053–2066.
- 1480 (71) Madera, M.; Vogel, C.; Kummerfeld, S. K.; Chothia, C.; Gough,
1481 J. The SUPERFAMILY database in 2004: additions and improve-
1482 ments. *Nucleic Acids Res.* **2004**, *32*, D235–D239.
- 1483 (72) Jaramillo, A.; Wernisch, L.; Héry, S.; Wodak, S. Folding free
1484 energy function selects native-like protein sequences in the core but
1485 not on the surface. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 13554–
1486 13559.
- 1487 (73) Shepherd, T. R.; Hard, R. L.; Murray, A. M.; Pei, D.; Fuentes, E.
1488 J. Distinct Ligand Specificity of the Tiam1 and Tiam2 PDZ Domains.
1489 *Biochemistry* **2011**, *50*, 1296–1308.
- 1490 (74) Polydorides, S.; Simonson, T. Monte Carlo simulations of
1491 proteins at constant pH with generalized Born solvent, flexible
1492 sidechains, and an effective dielectric boundary. *J. Comput. Chem.*
1493 **2013**, *34*, 2742–2756.
- 1494 (75) Kilambi, K.; Gray, J. J. Rapid calculation of protein pK_a values
1495 using Rosetta. *Biophys. J.* **2012**, *103*, 587–595.
- (76) Simonson, T.; Gaillard, T.; Mignon, D.; Schmidt am Busch, M.; Lopes, A.; Amara, N.; Polydorides, S.; Sedano, A.; Archontis, K. D. G. 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: Struct., Funct., Genet.* **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, A. V. 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, R. M.; Kortemme, T. Coupling Protein Side-Chain and Backbone Flexibility Improves the Re-design of Protein-Ligand Specificity. *PLoS Comput. Biol.* **2015**, *1*, e1004335.