A physics-based energy function allows the computational redesign of a PDZ domain

Vaitea Opuu,[†,#] Young Joo Sun,[‡,#] Titus Hou,[‡] Nicolas Panel,[†] David M. Ichikawa,[¶] Carlos Corbi-Verge,[§] Philip M. Kim,[§,‖,⊥] Marcus Noyes,[¶] Ernesto J. Fuentes,[*,‡] and Thomas Simonson[*,†]

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

‡Department of Biochemistry, Carver College of Medicine, University of Iowa, Iowa City, USA

¶Department of Biochemistry & Molecular Pharmacology and Institute for Systems Genetics, New York University School of Medecine, New York, USA

§Donnelly Centre for Cellular and Biomolecular Research, University of Toronto

‖Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada

⊥Department of Computer Science, University of Toronto, Toronto, Ontario, Canada

#Joint first authors

E-mail: ernesto-fuentes@uiowa.edu; thomas.simonson@polytechnique.fr

Abstract

A powerful approach to understand protein structure and evolution is to perform computer simulations that mimic aspects of evolution. In particular, structure-based computational protein design (CPD) can address the inverse folding problem, exploring a large space of amino acid sequences and selecting ones predicted to adopt a given fold. Previously, CPD has been used to entirely redesign several proteins: all or most of the protein sequence was allowed to mutate freely; among sampled sequences, those with low computed folding energy were selected, and a few percent of these did indeed adopt the correct fold. Those studies used an energy function that was partly or largely knowledge-based, with several empirical terms. Here, we show that a PDZ domain can be entirely redesigned using a "physics-based" energy function that combines standard molecular mechanics and a recent, continuum electrostatic solvent model. Many thousands of sequences were generated by Monte Carlo simulation. Among the lowest-energy sequences, three were chosen for experimental testing. All three could be overexpressed and had native-like circular dichroism and 1D NMR spectra. Two exhibited an upshift of their thermal denaturation curves when a peptide ligand was present, indicating they were able to bind and were most likely correctly folded. Evidently, the physical principles that govern molecular mechanics and continuum electrostatics are sufficient to perform whole-protein redesign. This is encouraging, since these methods provide physical insights, can be systematically improved, and are transferable to other biopolymers and ligands of medical or technological interest.

Introduction

Protein sequences have been selected by millions of years of evolution to fold into specific 3D structures, stabilized by a subtle balance of interactions involving protein and solvent.[1–3] In contrast, random polymers of amino acids are very unlikely to adopt a specific, stable, folded structure,[4–6] and exhibit instead a more disordered structure.[7] A powerful approach to understand the evolution of proteins and the physical origins of folding is to perform simulations that mimic evolution in a computer. This can be done with computational protein design (CPD), which explores a space of amino acid sequences and selects ones that are predicted to adopt a given fold.[8–12] A typical simulation imposes a specific geometry for the protein backbone, corresponding to the experimental conformation of a natural protein. Amino acid side chains are mutated randomly, for example through a Monte Carlo procedure. Variants that have a favorable predicted folding free energy are saved. The energy of the folded state is predicted with an energy function that can be physics-based or knowledge-based.[13–16] The unfolded state is usually described by a simple energy function that depends on the protein sequence but does not involve a detailed structural model. If most or all of the protein side chains are allowed to mutate during the simulation, we say that the protein is "completely redesigned". Indeed, the final, predicted sequences will then have little sequence identity to the natural protein whose backbone was used as a starting point.

The successful redesign of several complete proteins was first reported in 2003.[9,17] It was based on the Rosetta energy function, which contains several empirical terms and was parameterized specifically for protein design. Therefore, it can be considered to be at least partly knowledge-based. Several other successes were obtained[18,19] with updated versions of the Rosetta energy function,[16] including a recent large-scale study where 15000 miniproteins (40–43 amino acids) were redesigned.[20] 6% of the 15000 designs were shown to be successful; i.e., the designed miniproteins folded into the correct 3D conformation. The others either could not be overexpressed and purified, or did not fold as predicted. In addition to Rosetta, other knowledge-based energy functions were used to successfully redesign several proteins.[21,22]

Energy functions for the folded state can also be taken from molecular mechanics.[23,24] There are then only two energy terms for nonbonded interactions between protein atoms, which correspond to the elementary Coulomb and Lennard-Jones effects. Their parameterization relies mainly on fitting quantum chemical calculations performed on small model compounds in the gas phase. The solvent is described implicitly, using varying levels of approximation.[25] The most rigorous model used so far for CPD is a dielectric continuum model.[14,26] This requires solving a differential equation, which is technically impractical in a protein design framework. Therefore, a Generalized Born (GB) approximation is more common. GB contains much of the same physics but provides a simpler, analytical energy expression.[25,27] GB models have been studied extensively in the context of protein design but also molecular dynamics, free energy simulations, acid/base calculations, ligand binding and protein folding.[25,28–32] They reproduce the behavior of the dielectric continuum model rather accurately. Therefore, an energy function that combines molecular mechanics for the protein with a Generalized Born solvent can be considered "physics-based", even though it is not entirely constructed from first principles. A molecular mechanics energy, combined with a very simple solvent model, was used to computationally design two artificial proteins that each consisted of a four-helix bundle, where an elementary unit of 34 amino acids was replicated four times.[33,34] However, until now, there had not been a complete redesign of a natural protein using a physics-based energy function.

Here, we report the successful use of a physics-based energy function to completely redesign a PDZ domain of 83 amino acids. PDZ domains ("Postsynaptic density-95/Discs large/Zonula occludens-1") are globular domains that establish protein-protein interaction networks.[35–37] They interact specifically with target proteins, usually by recognizing a few amino acids at the target C-terminus. They have been extensively studied and used to elucidate principles of protein evolution and folding.[38–40] Our design started from the PDZ domain of the Calcium/calmodulin-dependent serine kinase (CASK) protein. It used the backbone conformation from an X-ray structure reported here. Positions occupied by glycine (seven) or proline (two) were not allowed to mutate. 13 positions involved in peptide binding also kept their wildtype identity. All 61 of the other side chains (73.5% of the sequence) were allowed to mutate freely into any amino acid type except Gly or Pro, for a total of 3.7 10^{76} possible sequences. The energy function combined the Amber ff99SB molecular mechanics force field[41] and a GB solvent model.[27,42] Computations were done with the Proteus software.[43,44] Three designs were tested experimentally and were all shown to fold. For two, binding to one or two peptides that are known CASK ligands was demonstrated. Evidently, the physical principles that govern molecular mechanics and continuum electrostatics are sufficient to allow large-scale computational protein design. This is encouraging, since these methods give physical insights, can be systematically improved, and are transferable to nucleic acids, sugars, noncanonical amino acids, biological cofactors, and many ligands of therapeutic or biotechnological interest.

Materials and methods

Computational design methods

Energy function for the folded state

We used the following effective energy function for the folded state:

$$E = E_{\rm MM} + E_{\rm GB} + E_{\rm SA} \tag{1}$$

$E_{\rm MM}$ is the protein internal energy, taken from the Amber ff99SB molecular mechanics (MM) energy function.[41] $E_{\rm GB}$ is a Generalized Born (GB) implicit solvent contribution:[27,45]

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

Here, ϵ_W and ϵ_P are the solvent and protein dielectric constants (80 and 4, respectively); r_{ij} is the distance between atoms i, j and b_i is the "solvation radius" of atom i.[27,42] The dependency of the b_i on the protein conformation corresponds to a GB variant we call GB/HCT (for "Hawkins-Cramer-Truhlar").[27,42] For some of the design calculations, an additional "Native Environment Approximation", or NEA was used for efficiency,[45,46] where the solvation radius b_i of each particular group (backbone, sidechain or ligand) was computed ahead of time, with the rest of the system having its native sequence and conformation.[46,47] For the other designs, we computed the solvation radii on the fly during the MC simulation, using a very fast implementation called "Fluctuating Dielectric Boundary," or FDB[47] that uses lookup tables.

The last term in Eq. (1) is a surface area term:

$$E_{\rm SA} = \sum_i \sigma_i A_i \tag{3}$$

A_i is the exposed solvent accessible surface area of atom i; σ_i is a parameter that reflects each atom's preference to be exposed or hidden from solvent. The solute atoms were divided into four groups with specific σ_i values. The values were -60 (nonpolar), 30 (aromatic), -120 (polar), and -110 (ionic) cal/mol/Å^2. The coefficient for hydrogens was zero. Surface areas were computed by the Lee and Richards algorithm,[48] implemented in the Proteus software,[43] using a 1.5 Å probe radius. To avoid overcounting buried surface, a scaling factor of 0.65 was applied to the contact areas involving at least one buried side chain.[42,45]

The unfolded state model

For a particular sequence S, the unfolded state energy has the form:

$$E^u = \sum_{i \in S} E^r(t_i, B_i). \tag{4}$$

The sum is over all amino acids; t_i represents the side chain type at position i; B_i represents the buried or exposed character of position i in the folded state. The quantities $E^r(t, B) \equiv E^r_t$ are referred to as "reference energies"; they can be thought of as effective chemical potentials of each amino acid type. Their values were chosen to maximize the likelihood of a set of experimental PDZ sequences. In practice, this means that a Monte Carlo simulation should give amino acid frequencies that match those in the experimental sequences.[49] We assigned different values to buried and exposed positions, because we assume residual structure is present in the unfolded state, so that amino acids partly retain their buried/exposed character. To define the target amino acid frequencies for likelihood maximization, we used a set of PDZ sequences collected earlier.[49]

Structural model and energy matrix

For CASK, we used a new X-ray structure of the apo PDB domain, reported here (PDB entry 6NH9). To carry out the Monte Carlo simulations, an energy matrix was computed using procedures described previously.[49] Briefly, 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.[50] Side chain rotamers were described by the Tuffery library,[51] expanded to include additional hydrogen orientations for OH and SH groups.[45] The energies were stored in an energy table, or "matrix" for use during MC.

Monte Carlo simulations

Sequence design was done by running long Monte Carlo (MC) simulations where 61 out of 83 positions could mutate freely: all but 7 Gly, 2 Pro and 13 positions that are directly involved in binding the peptide ligand. The MC simulations used one- and two-position moves, where either rotamers, amino acid types, or both changed. For two-position moves, the second position was near the first in space. Sampling was enhanced by Replica Exchange Monte Carlo (REMC), where eight MC simulations ("replicas") were run in parallel, at different temperatures.[52] Periodic swaps were attempted between the conformations of two replicas i, j (adjacent in temperature), subject to a Metropolis acceptance test.[52] Thermal energies ranged from 0.125 to 3 kcal/mol. Simulations were done with the Proteus software.[46,52]

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. Designed sequences were also submitted to the Superfamily library of Hidden Markov Models,[53,54] which attempts to classify sequences according to the Structural Classification Of Proteins, or SCOP.[55] The isoelectric point of each sequence was estimated by assuming each titratable side chain had its standard pK_a value.

Molecular dynamics simulations

Wildtype CASK and six sequences designed with Proteus were subjected to MD simulations with explicit solvent and no peptide ligand. The starting structures were taken from the MC trajectory or the crystal structure and slightly minimized with harmonic restraints to maintain the backbone geometry. Each protein was immersed in a solvent box using the CHARMM GUI.[56,57] The boxes had a truncated octahedral shape. The minimum distance between protein atoms and the box edge was 15 Å. The final models included about 11,000 water molecules. A few sodium or chloride ions were included to ensure overall electroneutrality. The protonation states of histidines were assigned to be neutral, based on visual inspection. MD was done at room temperature and pressure, using Langevin dynamics with a Langevin Piston Nosé-Hoover barostat.[58,59] Long-range electrostatic interactions were treated with a Particle Mesh Ewald approach.[60] The Amber ff14SB force field and the TIP3P model[61] were used for the protein and water, respectively. Simulations were run for one microsecond, using the Charmm and NAMD programs.[57,62]

Protein expression and purification

The genes encoding the Proteus PDZ designs were codon-optimized for *Escherichia coli* expression and chemically synthesized by GenScript Inc. (Piscataway, NJ). The genes were cloned into a modified pET21a vector (Novagen) that contains a His_6-tag and Tobacco etch virus protease cleavage site at the 5′-end of the multiple cloning site. The nucleotide coding sequence of the pET21a-PDZ vector was verified by automated DNA sequencing (University of Iowa, DNA Facility). Protein expression was conducted in BL21(DE3) (Invitrogen) *E. coli cells.* Typically, cells were grown at 37°C in Luria-Bertani medium supplemented with ampicillin (100 μg/mL) under vigorous agitation until an A600 of 0.6-0.8 was reached. Cultures were subsequently cooled to 18°C and protein expression was induced by the addition of isopropyl 1-thio-β-d-galactopyranoside to 1 mM final concentration. Induced cells were incubated for an additional 16-18 hrs at 18°C and harvested by centrifugation. Proteins were initially purified by nickel-chelate chromatography (GE-Healthcare). The proteins were further purified by size-exclusion chromatography (Superdex 75, GE Healthcare) using a buffer containing 20 mM phosphate, pH 6.8, 50 mM NaCl, and 0.5 mM EDTA. Samples were used immediately.

Crystal structure of the wildtype apo CASK PDZ domain

A crystal structure of the apo CASK PDZ domain was determined in this work. High-throughput hanging-drop, vapor-diffusion screens using a Mosquito drop setter (TTP LabTech) were used to determine the crystallization conditions. The CASK PDZ domain was prepared in 20 mM Tris pH 7.5 and 50 mM NaCl. 200 nL of precipitant and PDZ domain (10-30 mg/mL) was used for each screening condition. Initial screening for diffracting crystals was done with a CuK rotating anode beam. Collection of full X-ray diffraction datasets for structure determination was done at beamline 4.2.2 at the Advanced Light Source (Berkeley, CA). Proper space group handedness was verified by analysis of the electron density.

XDS was used for indexing, integration, and scaling of the diffraction data,[63,64] to 2.0 Å resolution. XSCALE was used to merge multiple datasets. We used PHASER and previously-determined PDZ structures for initial phasing.[65] We used Refmac[66,67] for the early stages of refinement and PHENIX[68,69] for the final refinement. Refinement statistics are given in Supplementary Material (Table S1). Manual model building was done based on visualized electron density in Coot.[70,70] 10% of the reflections were randomly selected to be excluded from the refinement and used to calculate R_{free} values. Alignment of structures and generation of figures were done with PyMOL (Schrodinger, LLC, The PyMOL Molecular Graphics System).

Biophysical characterization of designed proteins

Synthetic peptides

All peptides were chemically synthesized by GenScript Inc. (Piscataway, NJ) and were >95% pure as judged by analytical HPLC and mass spectrometry. Peptides were dansylated at the N-terminus and had a free carboxyl at the C-terminus. The peptides used in this study were derived from the following proteins: Neurexin (residues 1,470-1,477: NKDKEYYV$_{COOH}$), Caspr4 (residues 1,301-1,308: ENQKEYFF$_{COOH}$) and Syndecan1 (residues 303310: TKQEEFYA$_{COOH}$).

Circular dichroism

Circular dichroism signals were measured using a Jasco J-815 circular dichroism spectropolarimeter. The concentration of each protein ranged from 10 to 20 μM. All proteins were in a buffer composed of 20 mM phosphate, pH 6.8, 50 mM NaCl, and 0.5 mM EDTA. Spectra were taken from the 190 nm to 260 nm wavelength window with a 1 nm data interval at 25°C. Data integration time was 2 seconds and the scanning speed was 100 nm/min.

NMR

NMR experiments were carried out at 298 K (calibrated with methanol) on a Bruker Avance II 800 MHz spectrometer equipped with a 5 mm TCI CryoProbe. All protein samples were prepared in 20 mM phosphate, pH 6.8, 50 mM NaCl, 0.5 mM EDTA, and 10% (v/v) D_2O with a concentration of 14 μM to 22 μM.

Differential scanning fluorimetry

Standard methodology was used for differential scanning fluorimetry (DSF).[71,72] Briefly, DSF was performed using 96-well PCR plates and the Sypro Orange (Thermo Fisher) dye. Each well in the PCR plate had a 20 μL final volume containing 0.25 mg/mL of protein, 130 μM of peptide, and 5x Sypro Orange final concentration (from a 5000x stock) in a buffer containing 20 mM phosphate, pH 6.8, 50 mM NaCl, and 0.5 mM EDTA. The DSF assays were performed using a Bio-Rad CFX96 real-time polymerase chain reaction instrument equipped to read 96-well plates. The protein of interest was thermally denatured from 5°C to 95°C at a ramp rate of 1°C/min. The protein melting/unfolding curves were generated by monitoring changes in Sypro orange fluorescence (at 610 nm wavelength). Raw fluorescence data were analyzed using DMAN, and the first derivative value from the denaturation data was used to determine the apparent melting temperature[73] ($T_{1/2}$). Each peptide was assayed in triplicate. A 96-well plate containing no peptide was assayed to determine the apparent $T_{1/2}$ of each PDZ domain in the absence of any peptide. A shift of more than 1°C in $T_{1/2}$ indicates binding (based on SEM).

Results

Protein design simulations were done using the CASK backbone conformation, shown in Fig. 1. 61 out of 83 residues were allowed to mutate into all types except Gly and Pro, for a total space of $18^{61} = 3.7\ 10^{76}$ possible sequences. This space was explored using Replica Exchange Monte Carlo, without any bias towards natural sequences or any limit on the number of mutations. The 2,000 sequences with the lowest folding energies were retained for analysis. We describe their computational characterization and the selection of three sequences for experimental testing. Next, we describe experimental characterization of the selected sequences.

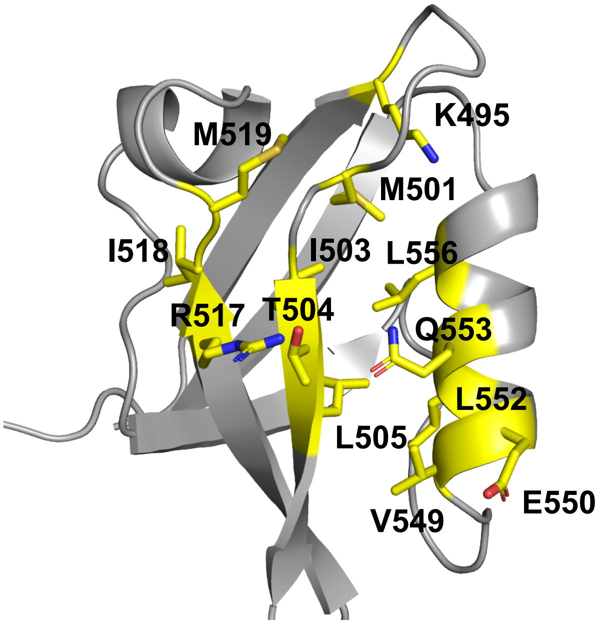


Figure 1: CASK 3D structure. In yellow: 13 amino acid positions whose types were kept fixed in the Proteus designs.

Computational characterization and sequence selection

The top 2,000 sequences spanned a folding energy interval of 1.5 kcal/mol. Since negative design was not included in our sequence generation, there is a possibility some of them would fold into a non-PDZ structure. Therefore, they were analyzed by the Superfamily fold recognition tool,[54,74] which assigns sequences to SCOP structural families. All 2,000 Proteus sequences were assigned by Superfamily to the PDZ fold; none were predicted to adopt any other fold in SCOP. We next computed the Blosum40 similarity scores between the designed sequences and natural sequences from the Pfam database. Two histograms of scores are shown in Fig. 2: one for the whole protein and one for the protein core (15 positions). The scores of the sequences designed with Proteus (with 61 out of 83 positions allowed to mutate) are high, and comparable to the scores of natural PDZ domains. The peaks in the Proteus histograms are narrow, indicating that the 2,000 lowest-energy sequences are similar to each other.

To narrow down the number of sequences, we excluded those with isoelectric points estimated to be close to the physiological pH, between 6.5 and 8.5, which might be subject to aggregation and hard to express. This reduced the number of sequences from 2000 to 1268. Next, we used the criterion of negative design, by only retaining sequences that had above-average Superfamily results. We kept sequences with above-average Superfamily match lengths (above 78) and E-values ($\log_{10}$ E $<$ -31). This left us with 692 sequences.

Since we planned to test only a few sequences experimentally, we reduced the number of candidates further using four additional criteria: (1) We excluded sequences with below-average similarity scores versus Pfam (the left part of the all-position peak in Fig. 2), leaving 215 sequences. (2) We excluded sequences that had a cavity buried in the predicted 3D structure. (3) We required a total unsigned protein charge of less than 6. (4) We allowed no more than 15 mutations that drastically changed the amino acid type

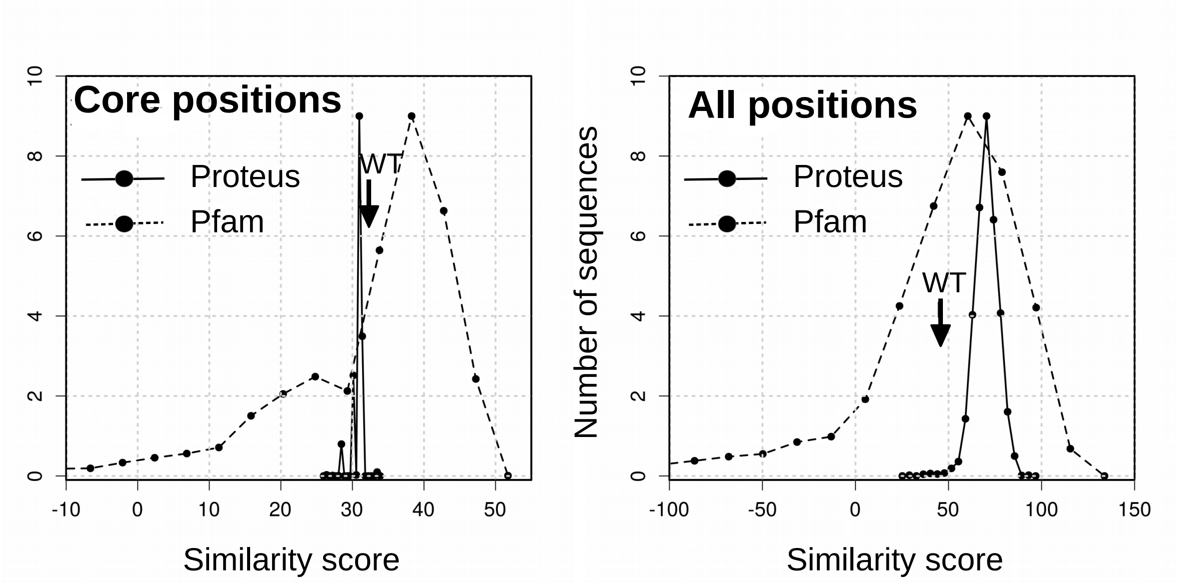


Figure 2: Blosum similarity scores compared to natural Pfam sequences. Black line: histogram of scores for the top 2000 Proteus sequences, considering only core positions (left) or all positions (right). Dashed line: scores for the Pfam sequences themselves. WT score is indicated by an arrow.

(defined by a Blosum62 similarity score between the two amino acid types of -2 or less). This left us with 16 candidate sequences, shown in Fig. 3. These were separated into four groups, based on visual inspection. Group 2 was eliminated based on its Arg494 residue, absent from CASK homologs. One candidate was selected from each of the other groups (highlighted in Fig. 3), with a preference for native or homologous residue types at positions 492 (candidate 1350), 494 (candidate 1555), and 548 (candidate 1669), all rather close to the peptide binding interface. Finally, the three candidates were each simulated by molecular dynamics with explicit solvent for one microsecond, where their stabilities appeared comparable to the wildtype (Supplementary Material, Fig. S5). Therefore, the three sequences were retained for experimental testing. The number of mutations, compared to wildtype CASK, were 50 (candidate 1350), and 51 (candidates 1555 and 1669), representing over 60% of the sequence.

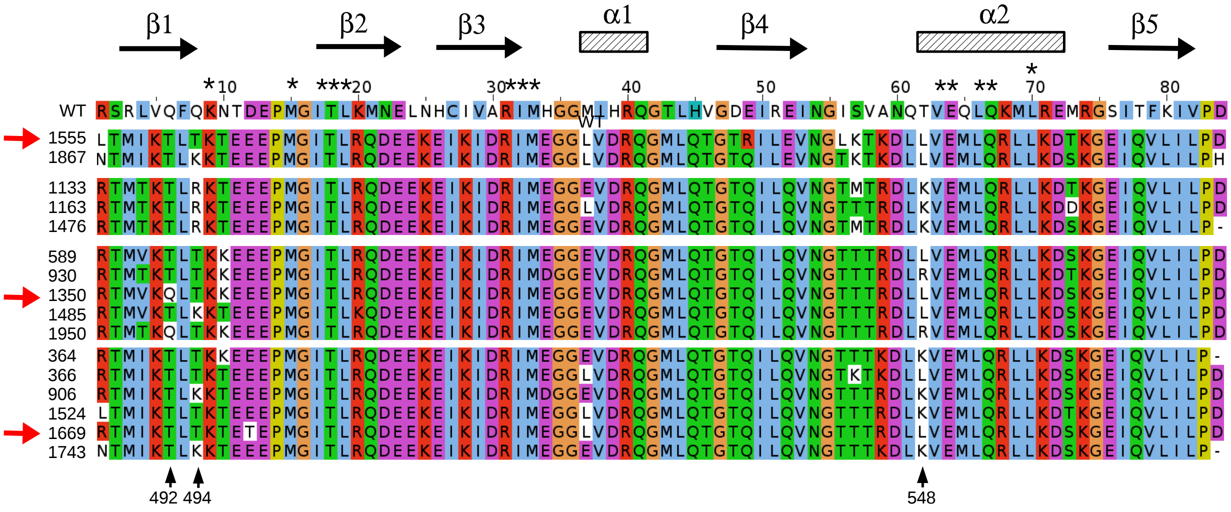

β1
β2
β3
α1
β4
α2
β5
*10 * ***20 30*** 40 50 60 ** ** * 70 80
WT RSRLVQFQKNTDEPMGITLKMNELNHCIVARIMHGGMLHRQGTLHVGDEIREINGISVANQTVEQLQKMLREMRGSITFKIVPD
WT
1555 LTMIKTLTKTEEEPMGITLRQDEEKEIKIDRIMEGGLVDRQGMLQTGTRILEVNGLKTKDLLVEMLQRLLKDTKGEIQVLILPD
1867 NTMIKTLKKTEEEPMGITLRQDEEKEIKIDRIMEGGLVDRQGMLQTGTQILEVNGTKTKDLLVEMLQRLLKDSKGEIQVLILPH
1133 RTMTKTLRKTEEEPMGITLRQDEEKEIKIDRIMEGGEVDRQGMLQTGTQILQVNGTMTRDLKVEMLQRLLKDTKGEIQVLILPD
1163 RTMTKTLRKTEEEPMGITLRQDEEKEIKIDRIMEGGLVDRQGMLQTGTQILQVNGTTTRDLKVEMLQRLLKDDKGEIQVLILPD
1476 RTMTKTLRKTEEEPMGITLRQDEEKEIKIDRIMEGGEVDRQGMLQTGTQILQVNGTMTRDLKVEMLQRLLKDSKGEIQVLILP-
589 RTMVKTLTKKEEEPMGITLRQDEEKEIKIDRIMEGGEVDRQGMLQTGTQILQVNGTTTRDLLVEMLQRLLKDSKGEIQVLILPD
930 RTMTKTLTKKEEEPMGITLRQDEEKEIKIDRIMDGGEVDRQGMLQTGTQILQVNGTTTRDLRVEMLQRLLKDTKGEIQVLILPD
1350 RTMVKQLTKKEEEPMGITLRQDEEKEIKIDRIMEGGEVDRQGMLQTGTQILQVNGTTTRDLLVEMLQRLLKDSKGEIQVLILPD
1485 RTMVKTLKKTEEEPMGITLKQDEEKEIKIDRIMEGGEVDRQGMLQTGTQILQVNGTTTRDLLVEMLQRLLKDSKGEIQVLILPD
1950 RTMTKQLTKKEEEPMGITLRQDEEKEIKIDRIMEGGEVDRQGMLQTGTQILQVNGTTTRDLRVEMLQRLLKDSKGEIQVLILPD
364 RTMIKTLTKKEEEPMGITLRQDEEKEIKIDRIMEGGEVDRQGMLQTGTQILQVNGTTTKDLKVEMLQRLLKDSKGEIQVLILP-
366 RTMIKTLTKTEEEPMGITLRQDEEKEIKIDRIMEGGLVDRQGMLQTGTQILQVNGTKTKDLLVEMLQRLLKDSKGEIQVLILPD
906 RTMIKTLKKTEEEPMGITLRQDEEKEIKIDRIMDGGEVDRQGMLQTGTQILQVNGTTTRDLKVEMLQRLLKDSKGEIQVLILPD
1524 LTMIKTLTKTEEEPMGITLRQDEEKEIKIDRIMEGGLVDRQGMLQTGTQILQVNGTTTRDLKVEMLQRLLKDTKGEIQVLILPD
1669 RTMIKTLTKTETEPMGITLRQDEEKEIKIDRIMEGGLVDRQGMLQTGTQILQVNGTTTRDLLVEMLQRLLKDSKGEIQVLILPD
1743 NTMIKTLKKTEEEPMGITLRQDEEKEIKIDRIMEGGEVDRQGMLQTGTQILQVNGTTTKDLKVEMLQRLLKDSKGEIQVLILP-
492 494
548

Figure 3: WT and designed sequences based on the CASK template forming the final group of 16 candidates. The sequences tested experimentally are shown by red arrows. Asterisks (above) indicate 13 positions not allowed to mutate during the design, in addition to Gly and Pro residues.

Experimental characterization of selected sequences

Earlier designs based on the Tiam1 template

Computational redesign of Tiam1 was described earlier.[49] It used the Tiam1 PDZ domain structure (PDB code 4GVD) and the simpler, NEA electrostatics model.[49] 8 designs were expressed and purified. The yields were low. Circular dichroism (CD) yielded spectra typical of random coil polymers, suggesting the proteins were misfolded, whereas the Tiam1 PDZ domain yielded a spectrum typical of a folded structure containing both helical and beta sheet secondary structure (Supplementary Material, Fig. S2). Similarly, 1D-NMR spectra of the amide region of the NEA designs had limited dispersion and broad resonances compared to the native Tiam1 PDZ domain (Fig. S3). Moreover, differential scanning fluorimetry (DSF) in the presence of known Tiam1 ligands did not show and binding by the Tiam1 NEA designs, while the Tiam1 PDZ domain showed robust binding (Suppl. Material, Fig. S4). Together, these data indicate that the NEA-based designs of the Tiam1 PDZ domain could be overexpressed but adopted unfolded structures thated lack the ability to bind known Tiam1 peptide ligands.

Designs based on the CASK template

Next, we characterized the three designs selected above. They were obtained using a new apo CASK PDZ domain structure (PDB code 6NH9) as template and the more rigorous FDB electrostatics model.[47] The expression yields in *E. coli* were improved over the NEA Tiam1 designs, though not to the level typically seen with native PDZ domains. In contrast to the NEA Tiam1 designs, CD spectra of FDB designs were similar to native PDZ domains, suggesting that these designs were structured (Fig 4). 1D-proton NMR of the amide region showed good dispersion and relatively sharp lines, consistent with a folded protein (Fig 5).

The design strategy included fixing the sequence of 13 amino acid positions important for peptide binding. Therefore, we tested the ability of the designs to bind CASK ligands, using DSF experiments. The CASK PDZ domain showed binding to SDC1, Caspr4 and NRXN (Fig 6 and Table 1), as expected. Strikingly, the three CASK FDB designs characterized also showed binding to some of the peptides, albeit not to the same extent. Thus, FDB-1350 had a significant thermal shift in the presence of NRXN and SDC1. FDB-1669 showed a 1.0°C change in $T_{1/2}$ in the presence of the NRXN peptide. In contrast, FDB-1555 did not show significant thermal shifts in the presence of any peptide. From these data, we conclude that the three CASK FDB designs were folded and two were capable of interacting with peptide ligands.

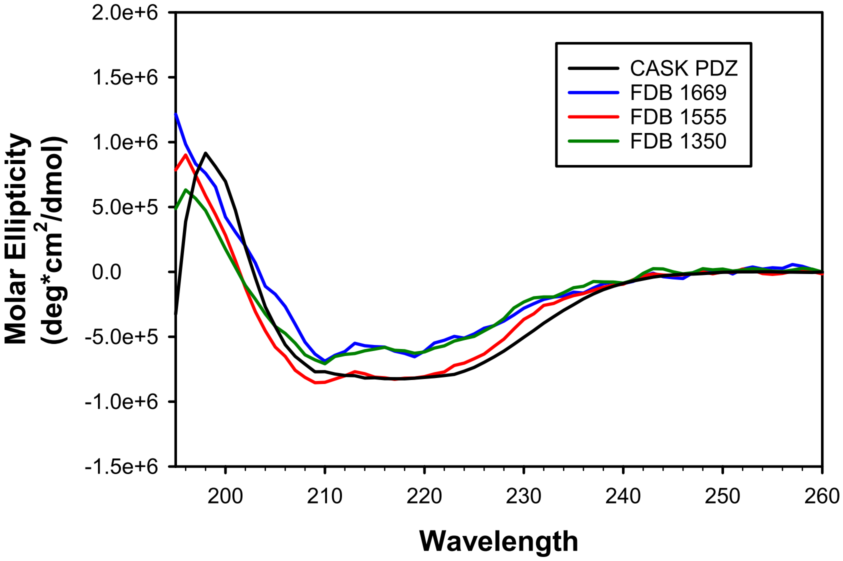


Figure 4: Circular dichroism spectra of a natural PDZ domain (Tiam1) and three selected designs based on the CASK template and the FDB electrostatic model. FDB-1350 (green), FDB-1555 (red), and FDB-1669 (blue) all have α helix and β strand signals similar to a native PDZ domain like CASK (black). The concentration of each protein ranged from 10 to 20 μM.

Table 1: DSF for wildtype CASK and three Proteus designs

	T_m (°C) and $\delta T_m = T_{1/2}^{apo}$ - $T_{1/2}$ (in parentheses)				
protein[a]	Apo	SDC1	Caspr4	NRXN	binding[b]
CASK PDZ	57.2 ± 0.2	58.4 ± 0.1 (+1.2)	58.7 ± 0.1 (+1.5)	58.1 ± 0.2 (+0.9)	**Sdc1, Caspr4** NRXN
FDB-1350	49.8 ± 0.4	50.7 ± 0.2 (+0.9)	50.4 ± 0.4 (+0.6)	51.3 ± 0.2 (+1.5)	Sdc1 **NRXN**
FDB-1669	49.1 ± 0.1	49.6 ± 0.1 (+0.4)	49.5 ± 0.0 (+0.4)	50.1 ± 0.1 (+1.0)	**NRXN**
FDB-1555	49.9 ± 0.2	50.2 ± 0.1 (+0.3)	50.3 ± 0.1 (+0.5)	50.5 ± 0.6 (+0.6)	–

[a] Protein concentration was ~25 μM (about 0.25 mg/ml). [b] When $\delta T_{1/2}$ was larger than sum of the standard deviation of apo and each PBMs, we considered the PBMs to have a significant change in $T_{1/2}$, indicating binding to the PDZ domain. ± indicates standard deviation of three biological replicates.

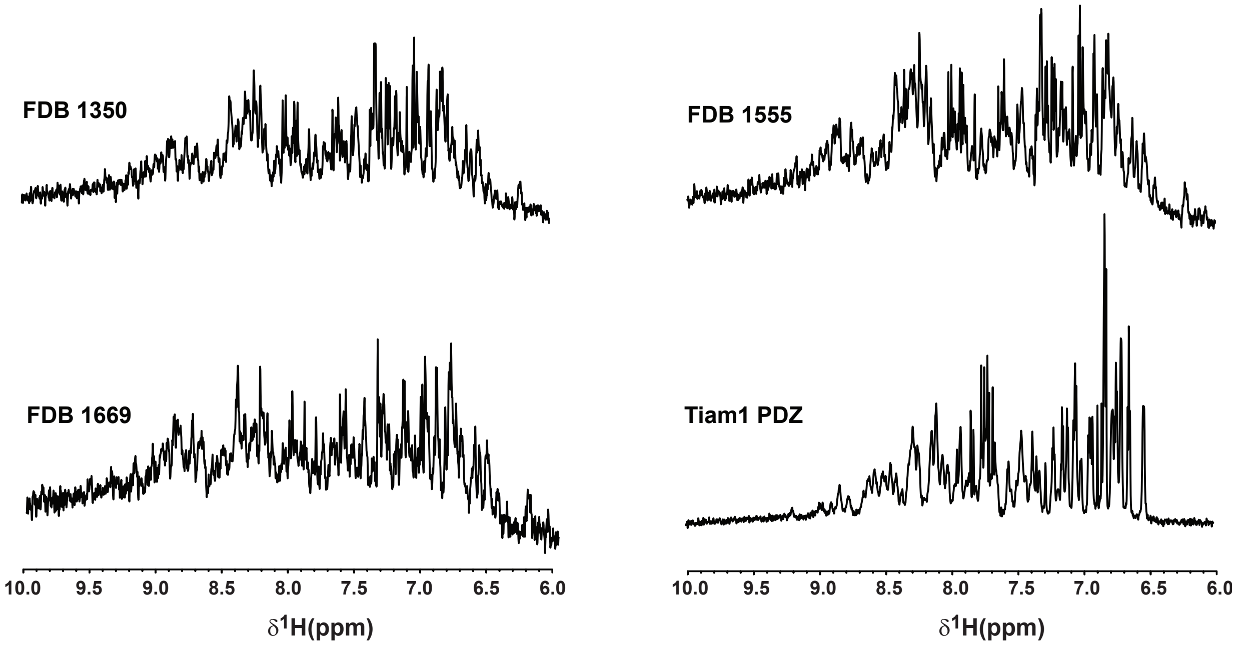


Figure 5: Proton NMR spectra of the natural Tiam1 PDZ domain and three selected designs, obtained based on the CASK template and the FDB electrostatic model.

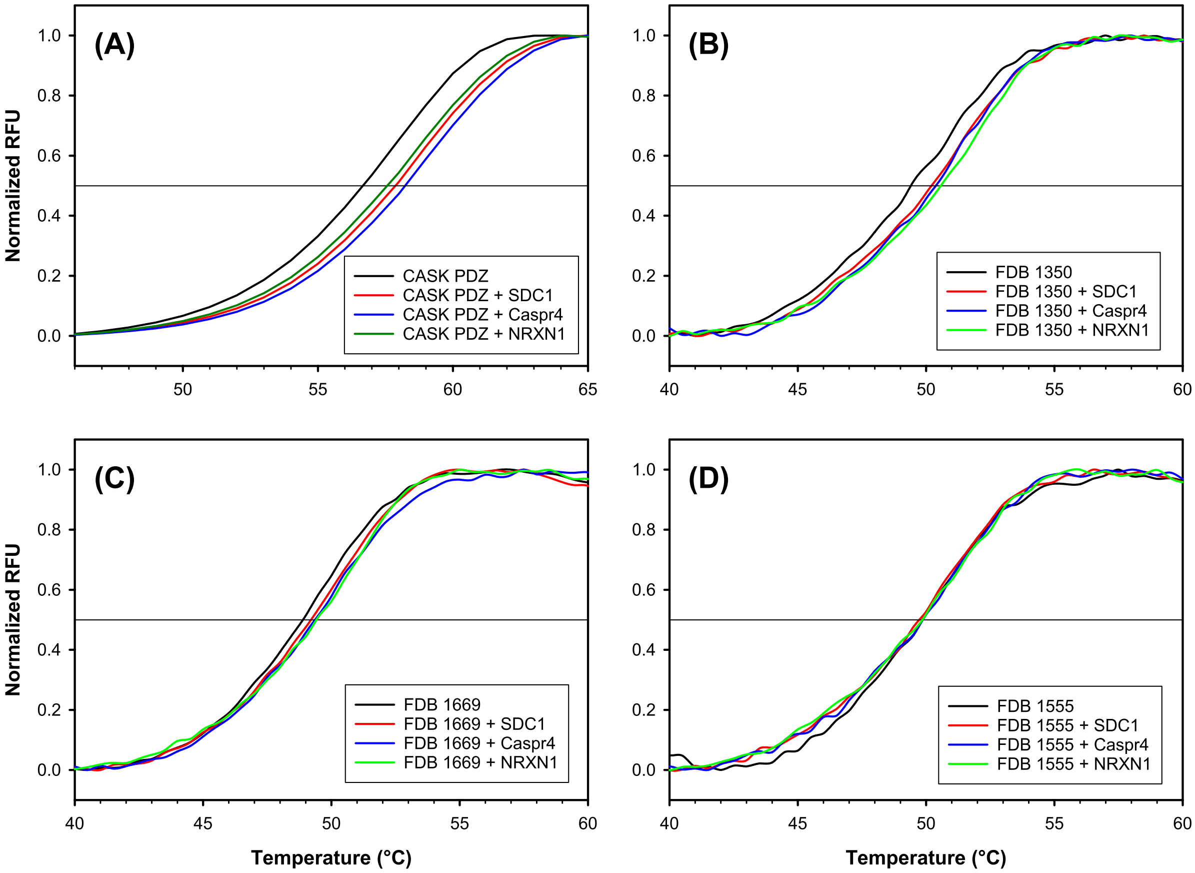


Figure 6: Differential scanning fluorimetry of (A) a natural PDZ domain (CASK) and (B–D) three selected designs based on the CASK template and the FDB electrostatic model. Signals in the absence and presence of the SDC1, Capr4 and NRXN peptides.

Discussion

Protein folding is induced partly by solvent–solvent interactions, and the solvent model is a key ingredient of our design procedure. The first solvation component in our model is nonpolar and uses accessible surface areas. Surface interactions in proteins have a many-body character,[8,42] since three or more residues can have surfaces that all overlap. Our model explicitly includes the most important 3-residue effects, while others are accounted for implicitly.[45] It relies on atomic surface tensions, a concept that is supported by atomistic simulations of small molecule solvation.[75] Their parametrization was updated recently (compared to our earlier Tiam1 designs).[76]

The largest protein solvation effects are electrostatic, and they also have a many-body character. The electrostatic, Generalized Born component of our model was carefully parameterized[42] and tested for several problems, including side chain acid/base constants, or pK_a's.[77] For Tiam1 redesign,[49] we had used an approximation where each protein residue experiences a constant, native-like, dielectric environment. This removed the many-body character of electrostatic solvation (and made the calculations very efficient). However, the Tiam1 designs were shown here to be largely unsuccessful: the proteins could be overexpressed but were only weakly structured. In contrast, preserving the many-body solvation effects was shown recently to give improved accuracy for side chain pK_a's.[47] It also led to increased similarity between CPD sequences and natural sequences of several PDZ proteins.[47] Therefore, for the present CASK redesign, we applied the many-body FDB electrostatic model and obtained much better results.

Our calculations used a CASK X-ray structure first reported here, determined at 1.85 Å resolution. In our design procedure, the protein backbone was held fixed in the X-ray conformation, while side chains mutated and explored rotamers. More precisely, the backbone motions were not ignored but were treated implicitly, through the protein dielectric constant, ϵ_P. The value used here, ϵ_P=4, is known to be physically reasonable for proteins.[78,79] Microsecond MD simulations further showed that the tested sequences, FDB-1350, -1555, and -1669, have backbone structures very similar to the wildtype protein. They also have native-like flexibilities.

Although our folded state model is arguably non-empirical, our design procedure did include several empirical elements. First, for the unfolded state, we assumed a simple, extended peptide model,[80] to which a correction was added that involved type-dependent amino acid chemical potentials.[43] These were chosen[49] to maximize the likelihood of sampling a collection of natural PDZ sequences. This was possible because our method samples sequences according to a known, Boltzmann probability distribution (known up to a constant normalization factor).

Second, we used several filters to choose a handful of sequences for experimental testing. Starting from sequences within 1.5 kcal/mol of the top folding energy, we used the (computed) isoelectric point to reduce the chances of aggregation. We also used negative design, which was not part of the sequence generation, eliminating candidates with below-average Superfamily fold recognition scores. We also eliminated sequences with below-average similarity scores, relative to natural sequences. These two criteria were actually not very stringent, in the sense that the score distributions were very narrow (Fig. 2). At this point, we were left with 215 sequences. We then eliminated sequences whose structural models included large cavities and ones with a large net charge. Finally, we eliminated sequences with more than 15 “drastic” mutations (corresponding to Blosum scores of -2 or less). This last, purely empirical filter left us with 16 sequences. Among these, we chose three that were representative. If some of the filters were omitted, for example if we had selected sequences randomly at the 215-sequence stage, we expect that there might have been more failures, instead of 2–3 successes out of 3 tests.

The three tested proteins could be overexpressed, had sharp 1D-NMR peaks and native-like CD spectra. Two exhibited a shift of their thermal denaturation in the presence of one or two peptides that are known CASK ligands. The expression yields and protein solubilities were lower than for wildtype CASK, so that it was not possible to produce large amounts of pure protein and do 2D-NMR or X-ray crystallogrphy. It may be possible to improve this behavior by testing a larger number of designs and/or using an empirical filtering of candidate sequences for solubility.

The present design method, which combines molecular mechanics, continuum electrostatics, and Boltzmann sampling, is an example of "physics-based" CPD. It is striking and encouraging that this approach allows whole protein redesign to be done successfully. Evidently, the solid physical basis of the energy function and its careful parameterization can lead to a good level of success, despite various approximations. We expect that the "physics-based" route will increasingly yield valuable physical insights and should be a valuable complement to knowledge-based CPD and experimental design.

Supplementary Material

Supplementary Material is available providing (1) experimental characterization of sequences designed using the Tiam1 template and the NEA electrostatics model; (2) statistical data on the X-ray structure refinement and (3) results from MD simulations of Tiam1, CASK and the three selected CASK-based designs.

Acknowledgements

Some of the calculations were done at the CINES supercomputer center of the French Ministry of Research. NAMD was developed by the Theoretical and Computational Biophysics Group in the Beckman Institute at the University of Illinois at Urbana. We are grateful to Dr. Jay Nix and the staff at beamline 4.2.2 at the Advanced Light Source, Lawrence Berkeley National Laboratory. We thank Lokesh Gakhar (U of Iowa Protein and Crystallography Facility) for helpful advice in crystallgraphy. The X-ray crystallography software used in the project was installed and configured by SBGrid. The Roy J. Carver Charitable Trust is acknowledged for funding of the Carver College of Medicine Medical Nuclear Magnetic Resonance Facility.

References

(1) Kauffman, S. A. *The origins of order. Self-organization and selection in evolution*; Oxford University Press, New York, 1993.

(2) Branden, C.; Tooze, J. *Introduction to Protein Structure*; Garland Publishing, NY, 1999.

(3) Onuchic, J. N.; Luthey-Schulten, Z.; Wolynes, P. G. Theory of protein folding: the energy landscape perspective. *Ann. Rev. Phys. Chem.* **1997**, *48*, 545–600.

(4) Davidson, A. R.; Sauer, R. T. Folded proteins occur frequently in libraries of random amino acid sequences. *Proc. Natl. Acad. Sci. USA* **1994**, *91*, 2146–2150.

(5) Wilson, D. S.; Keefe, A. D.; Szostak, J. W. The use of mRNA display to select high-affinity protein-binding peptides. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 3750–3755.

(6) Jackel, C.; Kast, P.; Hilvert, D. Protein design by directed evolution. *Ann. Rev. Biochem* **2008**, *37*, 153–173.

(7) Ptitsyn, O. B. Molten globule and protein folding. *Adv. Prot. Chem.* **1995**, *47*, 83–229.

(8) Dahiyat, B. I.; Mayo, S. L. De novo protein design: fully automated sequence selection. *Science* **1997**, *278*, 82–87.

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

(10) Samish, I.; MacDermaid, C. M.; Perez-Aguilar, J. M.; Saven, J. G. Theoretical and computational protein design. *Ann. Rev. Phys. Chem.* **2011**, *62*, 129–149.

(11) Feldmeier, K.; Hoecker, B. Computational protein design of ligand binding and catalysis. *Curr. Opin. Chem. Biol.* **2013**, *17*, 929–933.

(12) Au, L.; Green, D. F. Direct Calculation of Protein Fitness Landscapes through Computational Protein Design. *Biophys. J.* **2016**, *110*, 75–84.

(13) Pokala, N.; Handel, T. M. Energy functions for protein design I: Efficient and accurate continuum electrostatics and solvation. *Prot. Sci.* **2004**, *13*, 925–936.

(14) Vizcarra, C. L.; Mayo, S. L. Electrostatics in computational protein design. *Curr. Opin. Chem. Biol.* **2005**, *9*, 622–626.

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

(16) Alford, R. F. et al. The Rosetta all-atom energy function for macromolecular modeling and design. *J. Chem. Theory Comput.* **2017**, *13*, 3031–3048.

(17) Kuhlman, B.; Dantas, G.; Ireton, G.; Varani, G.; Stoddard, B.; Baker, D. Design of

a novel globular protein fold with atomic-level accuracy. *Science* **2003**, *302*, 1364–1368.

(18) Dantas, G.; Corrent, C.; Reichow, S. L.; Havranek, J. J.; Eletr, Z. M.; G.Isern, N.; Kuhlman, B.; Varani, G.; Merritt, E. A.; Baker, D. High-resolution Structural and Thermodynamic Analysis of Extreme Stabilization of Human Procarboxypeptidase by Computational Protein Design. *J. Mol. Biol.* **2007**, *366*, 1209–1221.

(19) Johansson, K. E.; Johansen, N. T.; Christensen, S.; Horowitz, S.; Bardwell, J. C. A.; Olsen, J. G.; Willemoes, M.; Lindorff-Larsen, K.; Ferkinghoff-Borg, J.; Hamelryck, T.; Winther, J. R. Computational Redesign of Thioredoxin Is Hypersensitive toward Minor Conformational Changes in the Backbone Template. *J. Mol. Biol.* **2016**, *428*, 4361–4377.

(20) Rocklin, G. J.; Chidyausiku, T. M.; Goreshnik, I.; Ford, A.; Houliston, S.; Lemak, A.; Carter, L.; Ravichandran, R.; Mulligan, V. K.; Chevalier, A.; Arrowsmith, C. H.; Baker, D. Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science* **2017**, *357*, 168–175.

(21) Xiong, P.; Wang, M.; Zhou, X.; Zhang, T.; Zhang, J.; Chen, Q.; Liu, H. Protein design with a comprehensive statistical energy function and boosted by experimental selection for foldability. *Nature Comm.* **2014**, *5*, 5330.

(22) Tian, P.; Louis, J. M.; Baber, J. L.; Aniana, A.; Best, R. B. Co-Evolutionary Fitness Landscapes for Sequence Design. *Ang. Chemie* **2018**, *57*, 5674–5678.

(23) Brooks, C. L.; Karplus, M.; Pettitt, M. Proteins: a theoretical perspective of dynamics, structure and thermodynamics. *Adv. Chem. Phys.* **1987**, *71*, 1–259.

(24) MacKerell Jr., A. D. In *Computational Biochemistry & Biophysics*; Becker, O., MacKerell Jr., A. D., Roux, B., Watanabe, M., Eds.; Marcel Dekker, N.Y., 2001; Chapter 1.

(25) Roux, B.; Simonson, T. Implicit solvent models. *Biophys. Chem.* **1999**, *78*, 1–20.

(26) Barth, P.; Alber, T.; Harbury, P. B. Accurate, conformation-dependent predictions of solvent effects on protein ionization constants. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 4898–4903.

(27) Hawkins, G. D.; Cramer, C.; Truhlar, D. Pairwise descreening of solute charges from a dielectric medium. *Chem. Phys. Lett.* **1995**, *246*, 122–129.

(28) Simmerling, C.; Strockbine, B.; Roitberg, A. E. All-atom structure prediction and folding simulations of a stable protein. *J. Am. Chem. Soc.* **2002**, *124*, 11258–11259.

(29) Simonson, T.; Carlsson, J.; Case, D. A. Proton binding to proteins: pK_a calculations with explicit and implicit solvent models. *J. Am. Chem. Soc.* **2004**, *126*, 4167–4180.

(30) Mongan, J.; Case, D. A.; McCammon, J. A. Constant pH molecular dynamics in generalized Born implicit solvent. *J. Comput. Chem.* **2004**, *25*, 2038–2048.

(31) Nguyen, H.; Roe, D. R.; Simmerling, C. Improved generalized Born solvent model parameters for protein simulations. *J. Chem. Theory Comput.* **2013**, *9*, 2020–2034.

(32) Panel, N.; Sun, Y. J.; Fuentes, E. J.; Simonson, T. A simple PB/LIE free energy function accurately predicts the peptide binding specificity of the Tiam1 PDZ domain. *Front. Molec. Biosci.* **2017**, *4*, art. 65.

(33) Cochran, F. V.; Wu, S. P.; Wang, W.; Nanda, V.; Saven, J. G.; Therien, M. J.; DeGrado, W. F. Computational de novo design and characterization of a four-helix

bundle that selectively binds a nonbiological cofactor. *J. Am. Chem. Soc.* **2005**, *127*, 1346–1347.

(34) Fry, H. C.; Lehmann, A.; Sinks, L. E.; Asselberghs, I.; Tronin, A.; Krishnan, V.; Blasie, J. K.; Clays, K.; DeGrado, W. F.; Saven, J. G.; Therien, M. J. Computational de Novo Design and Characterization of a Protein That Selectively Binds a Highly Hyperpolarizable Abiological Chromophore. *J. Am. Chem. Soc.* **2013**, *135*, 13914–13926.

(35) Hung, A. Y.; Sheng, M. PDZ domains: structural modules for protein complex assembly. *J. Biol. Chem.* **2002**, *277*, 5699–5702.

(36) Subbaiah, V. K.; Kranjec, C.; Thomas, M.; Ban, L. PDZ domains: the building blocks regulating tumorigenesis. *Biochem. J.* **2011**, *439*, 195–205.

(37) Shepherd, T. R.; Fuentes, E. J. Structural and thermodynamic analysis of PDZ-ligand interactions. *Methods in Enzymology* **2011**, *488*, 81–100.

(38) Lockless, W.; Ranganathan, R. Evolutionary Conserved Pathways of Energetic Connectivity in Protein Families. *Science* **1999**, *295*, 295–299.

(39) McLaughlin Jr, R. N.; Poelwijk, F. J.; Raman, A.; Gosal, W. S.; Ranganathan, R. The spatial architecture of protein function and adaptation. *Nature* **2012**, *458*, 859–864.

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

(41) Cornell, W.; Cieplak, P.; Bayly, C.; Gould, I.; Merz, K.; Ferguson, D.; Spellmeyer, D.; Fox, T.; Caldwell, J.; Kollman, P. A Second Generation Force Field for the Simula-

tion of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.

(42) Lopes, A.; Aleksandrov, A.; Bathelt, C.; Archontis, G.; Simonson, T. Computational sidechain placement and protein mutagenesis with implicit solvent models. *Proteins* **2007**, *67*, 853–867.

(43) Simonson, T.; Gaillard, T.; Mignon, D.; Schmidt am Busch, M.; Lopes, A.; Amara, N.; Polydorides, S.; Sedano, A.; Druart, K.; Archontis, G. Computational protein design: the Proteus software and selected applications. *J. Comput. Chem.* **2013**, *34*, 2472–2484.

(44) Simonson, T. *The Proteus software for computational protein design*; https://proteus.polytechnique.fr: Ecole Polytechnique, Paris, 2019.

(45) Gaillard, T.; Simonson, T. Pairwise Decomposition of an MMGBSA Energy Function for Computational Protein Design. *J. Comput. Chem.* **2014**, *35*, 1371–1387.

(46) Simonson, T. Protein:ligand recognition: simple models for electrostatic effects. *Curr. Pharma. Design* **2013**, *19*, 4241–4256.

(47) Villa, F.; Mignon, D.; Polydorides, S.; Simonson, T. Comparing pairwise-additive and many-body Generalized Born models for acid/base calculations and protein design. *J. Comput. Chem.* **2017**, *38*, 2396–2410.

(48) Lee, B.; Richards, F. The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **1971**, *55*, 379–400.

(49) Mignon, D.; Panel, N.; Chen, X.; Fuentes, E. J.; Simonson, T. Computational design of the Tiam1 PDZ domain and its ligand binding. *J. Chem. Theory Comput.* **2017**, *13*, 2271–2289.

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

(51) Tuffery, P.; Etchebest, C.; Hazout, S.; Lavery, R. A New Approach to the Rapid Determination of Protein Side Chain Conformations. *J. Biomol. Struct. Dyn.* **1991**, *8*, 1267–1289.

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

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

(54) Wilson, D.; Madera, M.; Vogel, C.; Chothia, C.; Gough, J. The SUPERFAMILY database in 2007: families and functions. *Nucl. Acids Res.* **2007**, *35*, D308–D313.

(55) Andreeva, A.; Howorth, D.; Brenner, S. E.; Hubbard, J. J.; Chothia, C.; Murzin, A. G. SCOP database in 2004: refinements integrate structure and sequence family data. *Nucl. Acids Res.* **2004**, *32*, D226–229.

(56) Jo, S.; Kim, T.; Iyer, V. G.; Im, W. CHARMM-GUI: a web-based graphical user interface for CHARMM. *J. Comput. Chem.* **2008**, *29*, 1859–1865.

(57) Brooks, B. et al. CHARMM: The biomolecular simulation program. *J. Comput. Chem.* **2009**, *30*, 1545–1614.

(58) Martyna, G. J.; Tobias, D. J.; Klein, M. L. Constant pressure molecular dynamics algorithms. *J. Chem. Phys.* **1994**, *101*, 4177–4189.

(59) Feller, S. E.; Zhang, Y.; Pastor, R. W.; Brooks, B. R. Constant pressure molecular dynamics simulation: the Langevin piston method. *J. Chem. Phys.* **1995**, *103*, 4613–4622.

(60) Darden, T. In *Computational Biochemistry & Biophysics*; Becker, O., MacKerrell Jr., A. D., Roux, B., Watanabe, M., Eds.; Marcel Dekker, N.Y., 2001; Chapter 4.

(61) Jorgensen, W. L.; Chandrasekar, J.; Madura, J.; Impey, R.; Klein, M. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926–935.

(62) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **2005**, *26*, 1781–1802.

(63) Kabsch, W. Integration, scaling, space-group assignment and post-refinement. *Acta Cryst. D* **2010**, *66*, 133–144.

(64) Kabsch, W. *Acta Cryst. D* **2010**, *66*, 125–132.

(65) McCoy, A. J.; Grosse-Kunstleve, R. W.; Adams, P. D.; Winn, M. D.; Storoni, L. C.; Read, R. J. Phaser crystallographic software. *J. Appl. Cryst.* **2007**, *40*, 658–674.

(66) Murshudov, G.; Vagin, A.; Dodson, E. Refinement of macromolecular structures by the maximum likelihood method. *Acta Cryst.* **1997**, *D53*, 240–255.

(67) Vagin, A. A.; Steiner, R. A.; Lebedev, A. A.; Potterton, L.; McNicholas, S.; F, F. L.; et al, REFMAC5 dictionary: organization of prior chemical knowledge and guidelines for its use. *Acta Cryst. D* **2004**, *60*, 2184–2195.

(68) Adams, P. D.; Grosse-Kunstleve, R. W.; Hung, L. W.; Ioerger, T. R.; McCoy, A. J.; Moriarty, N. W.; et al, PHENIX: building new software for automated crystallographic structure determination. *Acta Cryst. D* **2002**, *58*, 1948–1954.

(69) Adams, P. D.; Afonine, P. V.; Bunkoczi, G.; Chen, V. B.; Davis, I. W.; N, N. E.; et al, PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Cryst. D* **2010**, *66*, 213–221.

(70) Emsley, P.; Lohkamp, B.; Scott, W. G.; Cowtan, K. Features and development of Coot. *Acta Cryst. D* **2010**, *66*, 486–501.

(71) Ehrhardt, M. K. G.; Warring, S. L.; Gerth, M. L. Screening chemoreceptor-ligand interactions by high-throughput thermal-shift assays. Methods. *Methods Molec. Biol.* **2018**, *1729*, 281–290.

(72) Kranz, K. K.; Schalk-Hihi, C. Protein thermal shifts to identify low molecular weight fragments. *Methods Enzym.* **2011**, *493*, 277–298.

(73) Wang, C. K.; Weeratunga, S. K.; Pacheco, C. M.; Hofmann, A. DMAN: a Java tool for analysis of multi-well differential scanning fluorimetry experiments. *Bioinf.* **2012**, *28*, 439–440.

(74) Madera, M.; Vogel, C.; Kummerfeld, S. K.; Chothia, C.; Gough, J. The SUPERFAMILY database in 2004: additions and improvements. *Nucl. Acids Res.* **2004**, *32*, D235–D239.

(75) Mobley, D. L.; Bayly, C. I.; Cooper, M. D.; Shirts, M. R.; Dill, K. A. Small Molecule Hydration Free Energies in Explicit Solvent: An Extensive Test of Fixed-Charge Atomistic Simulations. *J. Chem. Theory Comput.* **2009**, *5*, 350–358.

(76) Gaillard, T.; Simonson, T. Full protein sequence redesign with an MMGBSA energy function. *J. Chem. Theory Comput.* **2017**, *13*, 4932–4943.

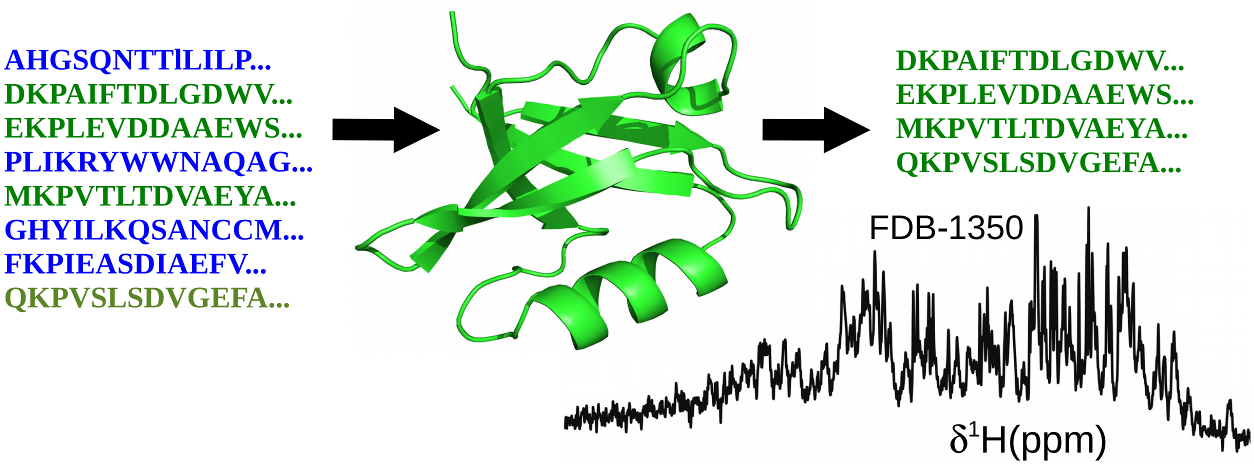
(77) Polydorides, S.; Simonson, T. Monte Carlo simulations of proteins at constant pH with generalized Born solvent, flexible sidechains, and an effective dielectric boundary. *J. Comput. Chem.* **2013**, *34*, 2742–2756.

(78) Simonson, T. Electrostatics and dynamics of proteins. *Rep. Prog. Phys.* **2003**, *66*, 737–787.

(79) Simonson, T. What is the dielectric constant of a protein when its backbone is fixed? *J. Chem. Theory Comput.* **2013**, *9*, 4603–4608.

(80) Pokala, N.; Handel, T. M. Energy functions for protein design: adjustment with protein-protein complex affinities, models for the unfolded state, and negative design of solubility and specificity. *J. Mol. Biol.* **2005**, *347*, 203–227.

Table of Contents Graphic



Supplementary Material: A physics-based energy function allows the computational redesign of a PDZ domain

Vaitea Opuu[a,†], Young Joo Sun[b,†], Titus Hou[b], Nicolas Panel[a], David M. Ichikawa[c], Carlos Corbi-Verge[d], Philip M. Kim[d,e,f], Marcus Noyes[c], Ernesto J. Fuentes[b,*] & Thomas Simonson[a,*]

[a]Laboratoire de Biochimie (CNRS UMR7654), Ecole Polytechnique, Palaiseau, France
[b]Dept. of Biochemistry, Carver College of Medicine, University of Iowa, Iowa City, USA
[c]Dept. of Biochemistry & Molecular Pharmacology and Institute for Systems Genetics, New York University School of Medecine, New York, USA
[d]Donnelly Centre for Cellular & Biomolecular Research, University of Toronto
[e]Dept. of Molecular Genetics and [f]Dept. of Computer Science, University of Toronto, Toronto, Canada

Below, we provide data on the experimental characterization of PDZ sequences designed using the Tiam1 template structure and the NEA electrostatics model. Next, we report the crystallographic structure statistics for the apo CASK PDZ domain. Finally, we report information on the stability and flexibility of the selected CASK-based designs in microsecond molecular dynamics (MD) simulations.

Experimental characterization of Proteus designs obtained with the Tiam1 template and the NEA electrostatic model

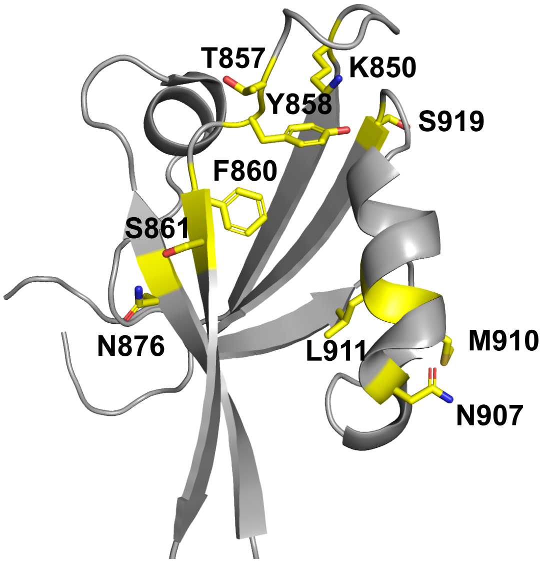


Figure S1: Tiam1 structure. Yellow: 13 positions whose types were fixed in the Proteus designs.

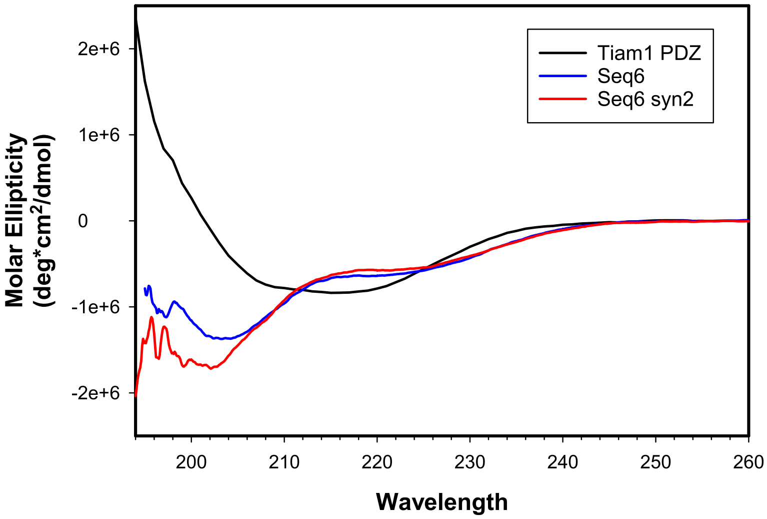


Figure S2: CD spectra of Tiam1 and two designs based on the Tiam1 template and NEA electrostatics.

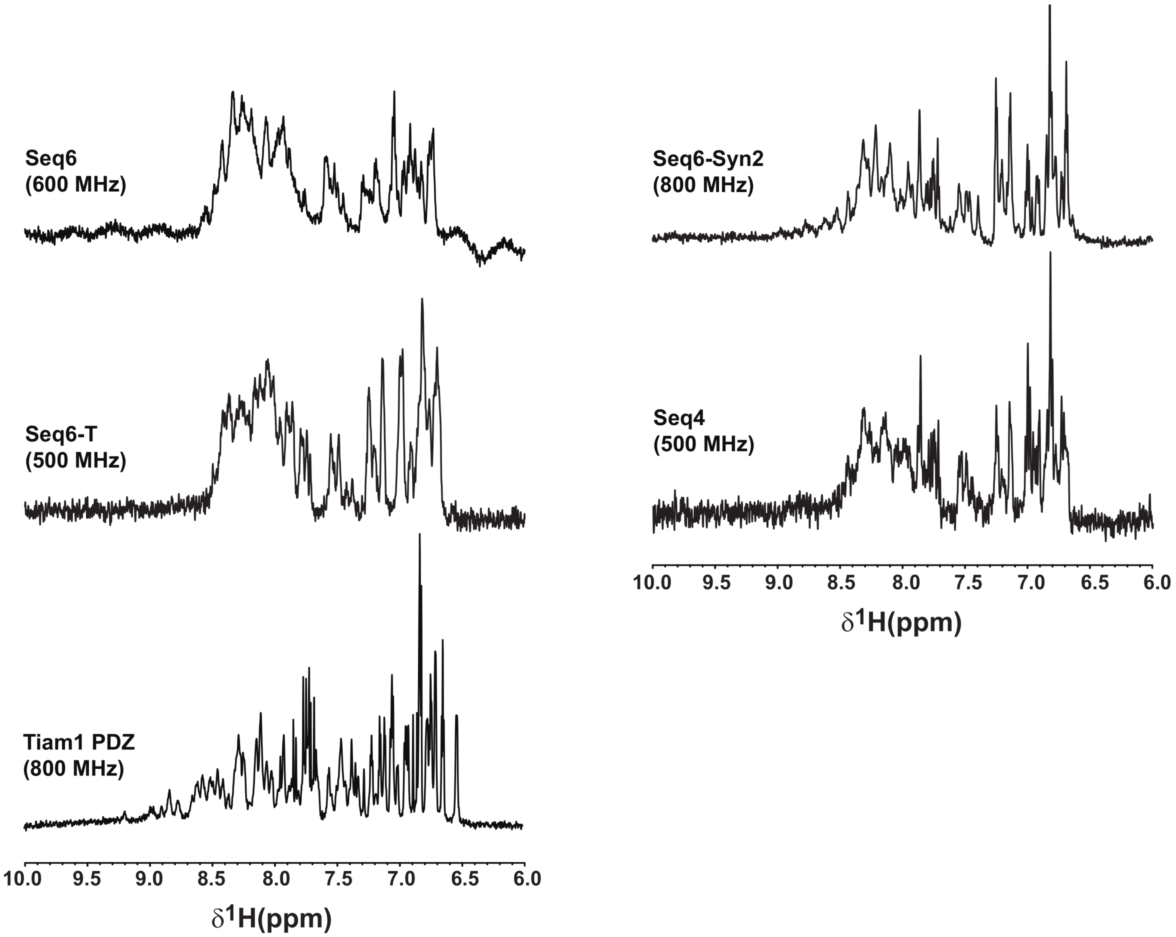


Figure S3: Proton NMR spectra of the Tiam1 PDZ domain and four designs obtained with the Tiam1 template and NEA electrostatics.

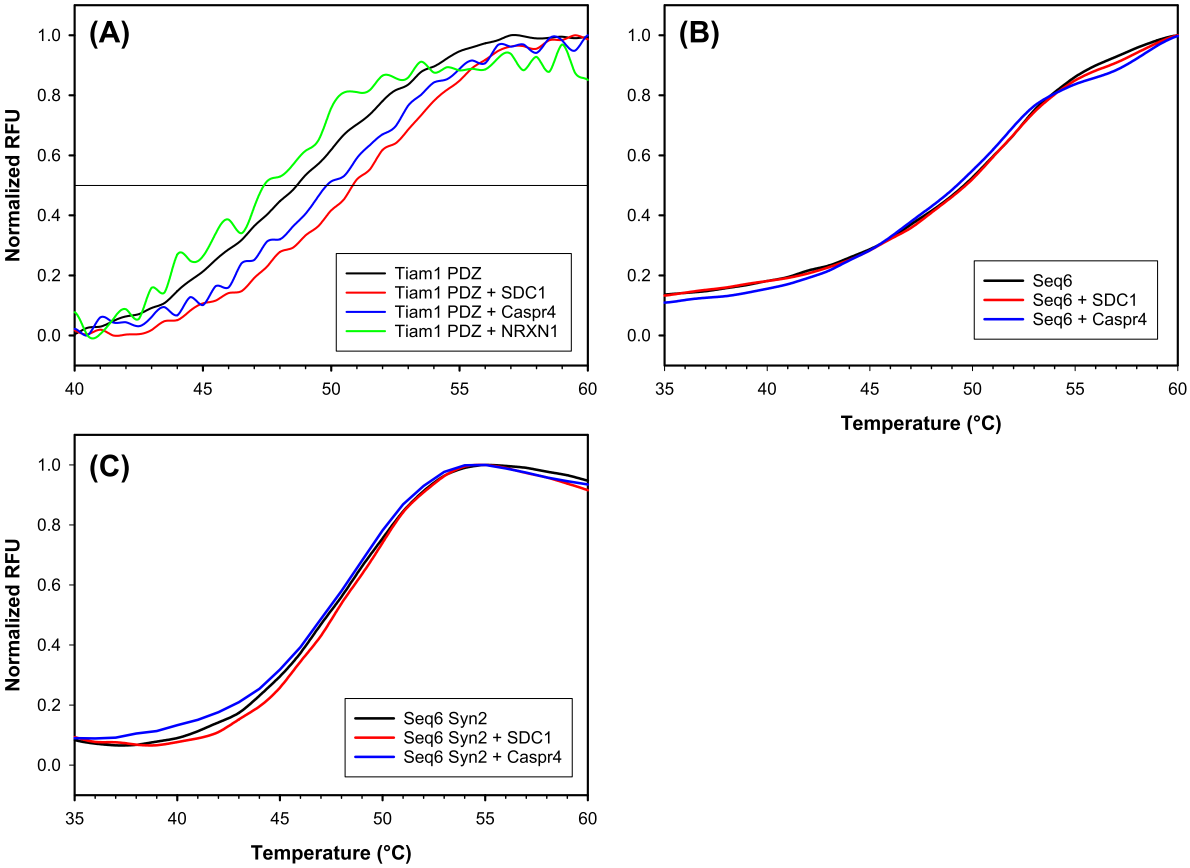


Figure S4: Differential scanning fluorimetry of a natural PDZ domain (Tiam1) and two designs based on the Tiam1 template and the NEA electrostatic model. Signals in the absence and presence of the SDC1, Capr4 and NRXN peptides.

Human apo CASK PDZ domain X-ray structure statistics

Table S1: Crystallographic statistics for the human apo CASK PDZ domain

Data collection statistics	
Beam line	ALS 4.2.2
Wavelength (Å)	1.0003
Space group	C 1 2 1
Unit cell dimensions (a, b, c) (Å)	61.1, 35.4, 119.5
Unit cell dimensions (α, β, γ)	90°, 90.3°, 90°
Resolution range (Å)	59.8—1.85
Total reflections	37,385 (7,461)
Unique reflections	20,769 (1,910)
Multiplicity	1.8 (1.7)
Completeness (%)	93.7 (93.7)
I/σ (I)	10.4 (2.1)
Wilson B-factor (Å^2)	50.7
Rmeas	0.030 (0.402)
$CC_{1/2}$	99.8 (91.1)
Refinement statistics	
Resolution (Å)	1.85
No. of reflections used in refinement	20,739 (2,705)
No. of reflections used for R_{free}	964 (133)
R_{work}/R_{free}	0.226/0.263
No. of atoms	4,188
Protein	4,037
Water	151
B-factors (Å^2)	53.0
R.M.S.D.[a]	
Bond length (Å)	0.29
Bond angle (degrees)	0.46
Ramachandran plot statistics (%)	
In preferred regions	98.0
In allowed regions	2.0
Outliers	0.0
PDB accession code	6NH9

The numbers in parentheses are for the highest-resolution shell. [a]Root mean square deviation from ideal values.

Stability of the three selected CASK-based designs in MD

As a first test of the three selected sequences, FDB1350, FDB1555, and FDB1669, they were subjected to MD simulations using an explicit solvent environment, for 1000 ns. Wildtype CASK (WT) was also simulated. Convergence of the simulations was good (based on a principal component analysis, not shown). The WT protein was quite stable, with rms deviations from the starting, X-ray structure of 1–1.5 Å (excluding 3–4 residues at each terminus and one very flexible loop, residues 495–502; see Fig. S5). Deviations from its own mean MD structure were similar (Fig. S5). The designed proteins exhibited only slightly larger deviations from the WT X-ray structure (1.2–1.8 Å) and similar, small deviations from their respective mean MD structures, with no visible drift (Fig. S5).

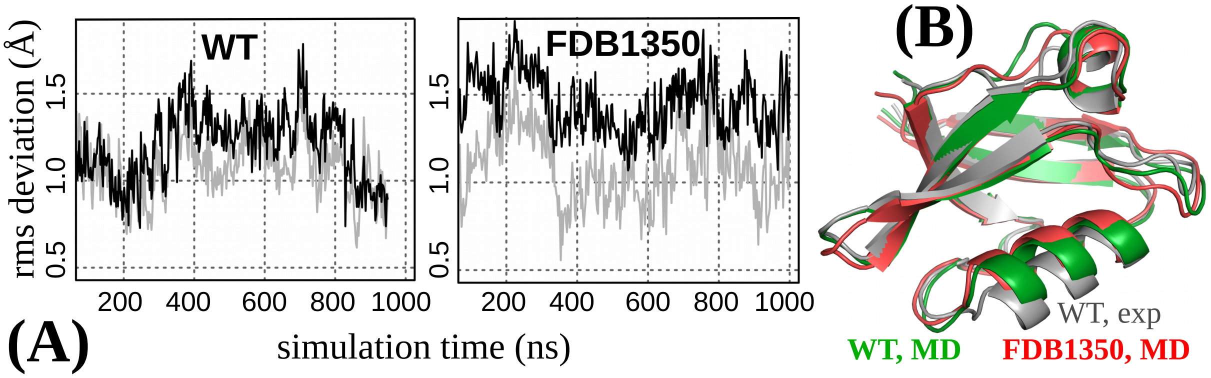


Figure S5: MD simulations of CASK-based designs. **A)** Backbone rms deviations for WT and the FDB1350 designed variant relative to the starting structure (black) and the mean MD structure (grey). **B)** Mean MD structures of WT and designed variant FDB1350.

We also characterized the backbone flexibility of the designed proteins by computing NMR order parameters for the backbone amide groups (Fig. S6). Experimental values were not available for WT CASK, but were available for Tiam1 and a quadruple mutant of Tiam1 (Liu et al, Structure, 12:342, 2016). These proteins were also simulated by MD for one microsecond, with and without the peptide ligands Sdc1 and Caspr4, respectively. In Fig. S6, we show the order parameters for both proteins in the apo and holo states, from experiment (circles) and MD (continuous lines) (top two panels). The agreement is very good. Next, we show (Fig. S6, bottom panel) the order parameters for WT CASK and the three selected CASK-based designs, FDB1350, FDB1555, and FDB1669 (apo proteins). Comparing the designed proteins to WT CASK, the results were similar, with some differences in loop regions. Two designs were slightly less flexible than WT (see positions 492-502 in β_1-β_2, 521-524 in β_3-α_1), while FDB1350 was slightly more flexible (see 492–496 in β_1-β_2 and 559-561 in α_2-β_5). Evidently, the design calculations do not produce overly-rigid or overly-flexible proteins in a systematic way.

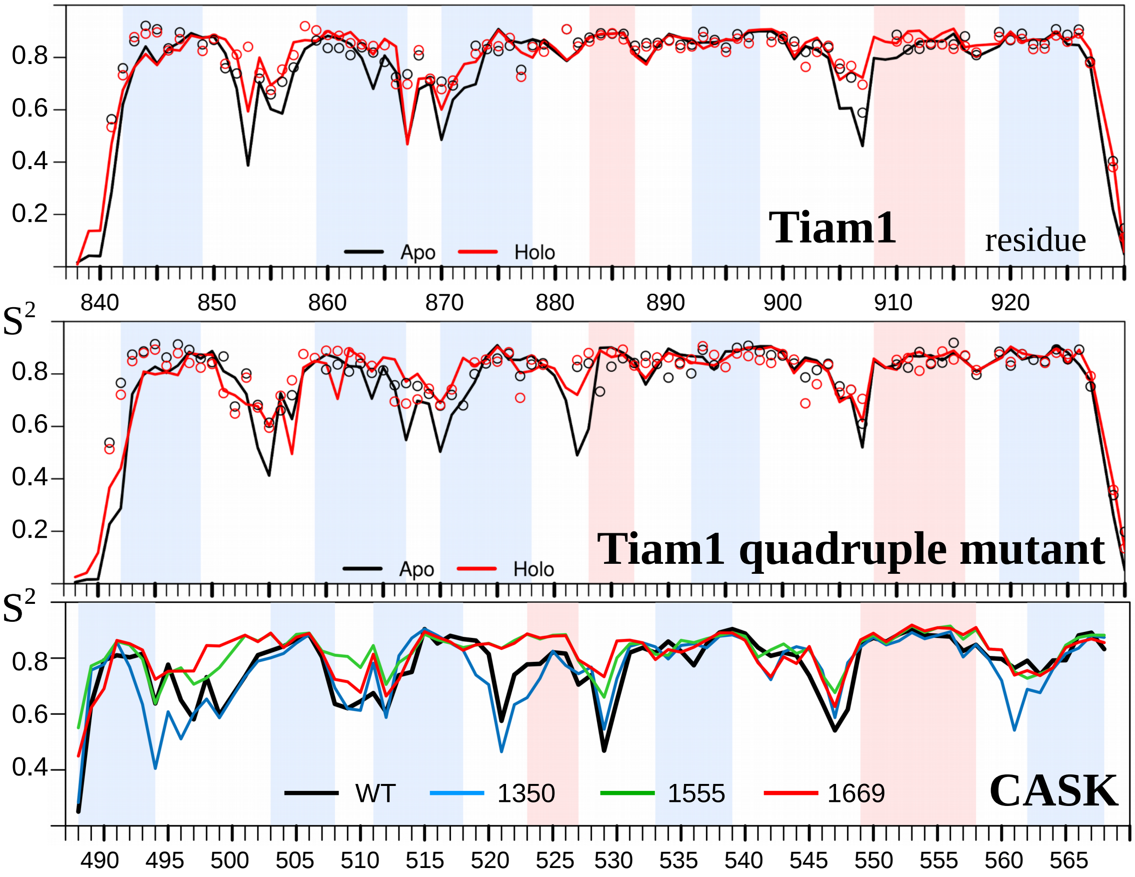


Figure S6: Backbone amide NMR order parameters for natural and designed proteins. **Top panel:** Tiam1 with and without the Sdc1 peptide ligand. Circles are experimental values; lines are from μsec MD simulations. **Middle panel:** analogous data for the Tiam1 quadruple mutant and the Caspr4 peptide. **Bottom panel:** Apo WT CASK and the three designed variants; values from MD.