Title

Improving the Efficiency of Ligand-Binding Protein Design with Molecular Dynamics Simulations.

Journal

Journal of Chemical Theory and Computation, 15(10)

ISSN

1549-9618

Authors

Barros, Emilia P
Schiffer, Jamie M
Vorobieva, Anastassia
et al.

Publication Date

2019-10-08

DOI

10.1021/acs.jctc.9b00483

Peer reviewed

HHS Public Access

Author manuscript

J Chem Theory Comput. Author manuscript; available in PMC 2020 October 03.

Published in final edited form as:

J Chem Theory Comput. 2019 October 08; 15(10): 5703–5715. doi:10.1021/acs.jctc.9b00483.

Improving the Efficiency of Ligand-Binding Protein Design with Molecular Dynamics

Emilia P. Barros[1], **Jamie M. Schiffer**[2], **Anastassia Vorobieva**[3,4], **Jiayi Dou**[4,5], **David Baker**[3,4], **Rommie E. Amaro**[1,6]

[1]–Department of Chemistry and Biochemistry, University of California, San Diego, La Jolla, CA, USA

[2]–Janssen Pharmaceuticals, Inc, San Diego, CA, USA

[3]-Department of Biochemistry, University of Washington, Seattle, WA, USA

[4]–Institute for Protein Design, University of Washington, Seattle, WA, USA

[5]–Current address: Department of Bioengineering, Stanford University, Stanford, CA, USA

[6]-National Biomedical Computation Resource, University of California, San Diego, La Jolla, CA, USA

Abstract

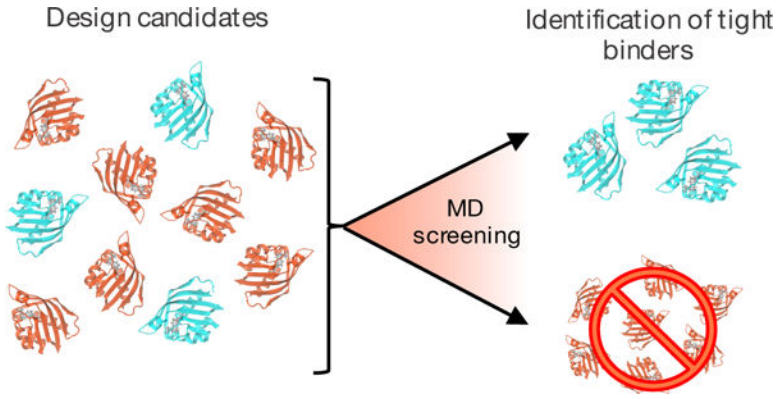
Custom-designed ligand-binding proteins represent a promising class of macromolecules with exciting applications towards the design of new enzymes or the engineering of antibodies and small-molecule recruited proteins for therapeutic interventions. However, several challenges remain in designing a protein sequence such that the binding site organization results in high affinity interaction with a bound ligand. Here, we study the dynamics of explicitly-solvated designed proteins through all-atom molecular dynamics (MD) simulations to gain insight into the causes that lead to the low affinity or instability of most of these designs, despite the prediction of their success by the computational design methodology. Simulations ranging from 500 to 1000 ns per replicate were conducted on 37 designed protein variants encompassing two distinct folds and a range of ligand affinities, resulting in more than 180 μs of combined sampling. The simulations provide retrospective insights into the properties affecting ligand affinity that can prove useful in guiding further steps of design optimization. Features indicate that entropic components are particularly important for affinity, which are not easily incorporated in the empirical models often used in design protocols. Additionally, we demonstrate that the application of machine learning approaches built upon the output from the simulations can help discriminate between successful and failed binders, such that MD could act as a screening step in protein design, resulting in a more efficient process.

SUPPORTING INFORMATION

The Supporting Information is available free of charge at http://pubs.acs.org.

DIG designs RMSF (Figure S1), POVME (Figure S2), dihedral (Figure S3) and ligand RMSD analysis (Figure S4), simulation length and replicate convergence analysis (Figures S5 and S6), static models features (Figure S7), RMSF and structural deformation of selected β-barrel designs (Figure S8), DIG and β-barrel designs joint unsupervised classification model (Figure S9), logistic regression model feature weights (Figure S10), evaluation of supervised learning classifiers using 33% or 50% of the data in the training set (Table S1).

Graphical Abstract



INTRODUCTION

Protein design is a young and ambitious field that aims to expand beyond naturally-occurring proteins to explore the massive protein sequence- and fold-spaces in the search for novel and customized structures[1,2]. Successes in the design of novel folds[3,4], ligand-binding proteins[5–7], enzymes[8,9], antibodies[10–12] and self-assembling supra-molecular structures[13–16] underscore this field's progress and growing potential. However, despite an increasing number of achievements, the protein design process remains very challenging and time consuming, with usually low success rates in initial design rounds[11,17].

Molecular recognition and protein-ligand binding are universally important processes that are however not yet fully understood or emulated. The development of novel molecules to treat diseases rely on the understanding of these interactions, and the improvement of protein-ligand affinity is far from being a negligible task[18]. In this context, the design of ligand-binding proteins offers the opportunity to better investigate the fundamentals affecting high affinity binding and selectivity[1,5], as well as lay out the foundations for custom design of *de novo* enzymes[19], biosensors[20,21] and antibody engineering[22,23]. Designing ligand-binding proteins poses the extra challenge that protein scaffolds not only need to be structurally stable and fold in the intended conformation, but also include residues lining up the binding cavity that result in high-affinity interactions with the ligand. Thus, the functionalization of the binding site, generally with polar residues for the establishment of hydrogen bonds with the ligand, has to be balanced with the hydrophobicity of the protein core to maintain an energetically favorable folded state[7] and the desolvation cost of the polar cavity upon ligand binding[24].

The general ligand-binding design protocol involves initial sampling of disembodied amino acids to create a binding site with specific protein-ligand interactions. The binding site is then positioned in a protein scaffold, and surrounding residues are further optimized to generate the desired interactions or to buttress the interactions in the secondary shell[5]. While tight ligand binders have been successfully generated by computational design, in a recent study 17 pre-selected designs of the nuclear transport factor 2 (NTF2) scaffold had to be expressed and tested to yield two successful μM-binders[5], while the pool of tested

candidates for the more hydrophobic fentanyl ligand involved 62 candidates, among which only three were successful first generation μM to nM binders[21]. More recently, the first completely de-novo designed β-barrel binding proteins required the generation of thousands of designs and experimental characterization of 56 high-scoring sequences to yield two successful binders in the first round of design generation[7]. This constitutes an expensive and lengthy process, as the computational design generation needs to be followed by expression of the highest-ranking candidates and experimental characterization, which includes assays to test proper protein folding and stability (such as circular dichroism and yeast-surface display), and ligand binding (e.g. fluorescence activation or polarization and isothermal titration calorimetry). Several challenges, including the evaluation of desolvation energies and sampling of alternate backbone conformations[25] affect the design accuracy, and thus the majority of proteins in the initial rounds of computational design end up failing the experimental validation. The most common sources of failure are due to improper protein folding (leading to aggregation and insolubility in many cases), or absence of high-affinity interactions with the ligand. The few promising candidates from the first round of design can then be optimized by techniques such as site-saturation mutagenesis to yield tighter binding proteins, further lengthening the design process.

Protein function is directly determined by the macromolecule's structure and dynamics, and in this sense molecular dynamics (MD) simulations are uniquely poised to assist in the protein design process as the simulations can inform on the designs' dynamics beyond the static models generated by the empirical design protocols[26,27]. MD simulations have been successfully applied at several different steps in the protein design methodology, such as in the refinement of predicted protein structures[28–30] or in the identification of flexible regions in designed proteins. The enumeration of mutations for the rigidification of these flexible sites[31] led to verified increase in thermal stability experimentally[32–34]. Additionally, investigation of designed enzymes using short simulations provided indication of key disrupted catalytic interactions in unsuccessful designs, and was proposed as a screening method for enzyme design[35].

Here, we investigate the dynamics of 37 ligand-binding designs from two different scaffolds and designed to bind to distinct ligands using MD simulations (Figure 1a). We survey the dynamical properties that reveal rational explanations for the unexpected failure of some of these designs, and explore the applicability of the simulations as a screening tool for early identification of the most promising designs prior to experimental validation (Figure 1b). Differences in protein structural flexibility, ligand dynamics, pocket pre-organization and water dynamics inside the cavity provide evidence of the predictive power of MD simulations when used concomitantly with the protein design process. We find that the application of machine learning approaches to the descriptors generated from simulations of the 37 designs in a retrospective analysis allows for accurate classification of the models and identification of the tight binder designs. This reinforces the potential of using MD simulations in the protein design pipeline to achieve higher efficiency and success rates in the design of novel functional proteins.

METHODS

System selection and preparation

The starting structures for the simulations consisted of Rosetta-modeled ligand-binding proteins published previously[5,7]. Thorough descriptions of the design methodologies and experimental characterization assays employed, including ligand affinity and selectivity measurements, can be found in references 5 and 7. Four designs of the digoxigenin-binders based on Nuclear Transport Factor 2 (NTF2) folds were selected (DIG10.2, DIG10.3, DIG12 and DIG16, here referred to as DIG designs)[5], as well as 33 designs of the fluorogenic 3,5-difluoro-4-hydroxybenzylidene imidazolinone (DFHBI) binders based on a *de novo* β-barrel scaffold (Figure 1a)[7]. This set included examples of tight binders as well as unsuccessful designs, thus classified due to failures to fold properly or to bind to the ligand with high affinity (Table 1). Besides the modeled structures, we also performed simulations starting from the X-ray crystal structures of DIG10.2 (PDB code 4J8T, chains A and B) and DIG10.3 (PDB code 4J9A, chains C and F) to investigate possible errors introduced by the use of design models instead of experimentally-validated structures. Missing terminal residues in the crystal structures of DIG10.2 and DIG10.3 were modeled with Schrödinger's Maestro (version 10.4, Schrödinger, LLC, New York, NY) based on the known protein sequence, and missing side chains were added with Schrödinger's Prime[36,37].

The digoxigenin and DFHBI ligands were parametrized using Antechamber and the generalized Amber force field (GAFF)[38,39], with geometry optimization performed with Gaussian 09[40]. For DFHBI, the torsional parameters of the C2-C1-C7-C8 dihedral were increased to model the molecule's expected planarity due to its aromaticity in the bound fluorophore state. All starting structures were processed with Maestro-integrated PROPKA to assign protonation states at pH 7. Dowser[41] was used to hydrate the protein cavity following removal of the ligand's coordinates for apo simulations. The proteins were solvated in water boxes with a buffer distance of 13 Å (for the β-barrels) or 15 Å (DIG designs) to the box edge with counter ions for charge neutrality, and 150 mM NaCl to simulate the experimental ionic concentration. The Amber14SB force field[42,43] was used for the protein and NaCl, with TIP3P for the water molecules[44]. As a note, we re-checked the protein protonation in the 33 β-barrel simulations after 500 ns of sampling and observed that about 30% of the histidine residues were assigned a different protonation state due to structural rearrangements, evidencing the limitations of conventional MD when it comes to fixed protonation states[45].

Molecular dynamics simulation protocol

All systems were simulated in their apo and ligand-bound (holo) states. We used the MD Kepler Workflow developed by our lab to enable automated MD minimization and equilibration for such a large number of systems[46]. Minimization consisted of five stages: hydrogen only, solvent, solvent and ligand, side chains, and the full system resulting in 13,000 cycles using a combination of steepest decent and conjugate gradient methods. Since the majority of the starting structures were not experimentally-resolved conformations, we performed a long equilibration protocol and verified RMSD evolution to ensure system relaxation and convergence. Equilibration involved an initial heating to 100 K at constant

volume for 50 ps followed by heating to 298 K at constant pressure, 1 bar, for 200 ps. The systems were further equilibrated at 298 K and 1 bar for 2.25 ns.

Molecular dynamics simulations were run using GPU accelerated Amber14[42,47] as an NPT ensemble with periodic boundary conditions at 1 bar and 298 K to simulate experimental conditions. We used a non-bonded short-range interaction cutoff of 10 Å, and the long-range electrostatic interactions were approximated by particle mesh Ewald[48]. The simulations used a 2 fs time step with the SHAKE algorithm to constrain hydrogen atoms. The initial design data set was simulated for five replicas of 1,000 ns each in the apo and holo states, while additional validation systems were simulated in three 500 ns replicas of each state (Table 1), resulting in a total sampling of 184 μs. MD input files are available for download at https://github.com/emiliapb/Design_screening.

Analysis methods

Trajectory files were visualized in VMD[49] and analysis were conducted using Jupyter notebooks[50] and in-house scripts, using a variety of MD analysis functions from MDTraj[51], CPPTRAJ[52], PyEMMA[53] and MDAnalysis[54,55]. The Jupyter notebooks are available for download on GitHub (https://github.com/emiliapb/Design_screening). Specifics of the different analysis methods conducted are discussed below.

Protein structural flexibility—Root mean square fluctuation (RMSF) of Cα carbons was calculated using CPPTRAJ, following structural alignment of the protein to backbone atoms. To obtain a single value informative of structural flexibility for each design, we used PyEMMA's regular space clustering of the Cα coordinates with RMSD metric and a cutoff of 1.5 Å to obtain the number of clusters (NOC) sampled. Simulations were analyzed every 100 frames.

To inform on the solvent accessibility of hydrophobic residues, we calculated solvent accessible surface area (SASA) of Ala, Ile, Leu, Phe, Val, Pro, Gly, Met and Trp residues for every frame in the simulations, using MDTraj's SASA function.

Ligand dynamics—We investigated ligand displacement in the holo simulations through the calculation of ligand root mean square deviation (RMSD) from the starting conformation in the designed models. Each trajectory was aligned to the protein coordinates of the respective starting structure, and RMSD values calculated using CPPTRAJ.

Pocket organization—Cavity volume was investigated using POVME (Pocket Volume Measurer), version 2.0[56]. Volume calculations were performed for every 100 frames of the aligned trajectories. Inclusion spheres were defined to encompass the binding site, and seed spheres were selected to include the minimal definition of the pocket, which were placed roughly at the center of the ligand position in the binding pocket. To allow for comparison across designs, the same POVME spheres were used for each scaffold and POVME's convex hull option was turned off.

The side chain chi1 dihedral angles of residues designed to interact with the ligand were investigated using MDTraj. For protein-ligand hydrogen bond analysis, MDTraj was used to

calculate the distance between the hydrogen and acceptor atoms, and the angle formed between donor, H and acceptor atoms. H-bonds that fell within the definition of strong and moderately-strong bonds were counted (Strong = XH --- Y bond length of 1.2–1.5 Å and X-H---Y angle of 170–180°, moderate = bond length between 1.5–2.2 Å and angle 130–170°)[35,57].

Water analysis—To investigate the presence of water molecules inside the protein cavity, we counted the number of water molecules within a sphere delimiting the binding site using MDAnalysis. The delimiting region was selected as a sphere of radius 8 Å from the coordinates of the C7 atom in the digoxigenin ligand, and a sphere of radius 7 Å centered at DFHBI's C1 atom (Figure 1a). The survival probability function in MDAnalysis[58] was used to calculate water survival probability within the same defined spheres in the last 100 ns of the apo simulations.

Convergence analysis—To assess the influence of simulation length and number of replicas on the computed features, we used the protocol described by Knapp, Ospina and Deane[59], computing the average difference between the features for 100 rounds of random selection without replacement.

Machine learning—Python's scikit-learn library was used to perform unsupervised and supervised learning on the features extracted from 500 ns of simulations. For the designs that have been simulated for 1000 ns and 5 replicas, we used the first 500 ns of the simulation and the first three replicas (generated using random seeds) for the calculations. Feature scaling was performed among the designs of a particular scaffold to prevent dominance of larger-valued features. Logistic regression was performed with a tolerance of 10^{-4} and liblinear solver. K-nearest neighbors used k=2 or 5, uniform weights and Euclidian distance metric.

RESULTS AND DISCUSSION

Binding determinants for DIG designs

To investigate the dynamics of designed small-molecule binding proteins and understand the determinants affecting binding ability and affinity, we first investigated 4 designed proteins of the Nuclear Transport Factor 2 scaffold, which have been engineered to bind to the small molecule digoxigenin[5] (Figure 1a). We conducted extensive simulations in both the apo and holo states of two successful, tight binder designs (DIG10.2 and DIG10.3) and two designs that failed to bind to the ligand despite positive predictions by the computational methodology (DIG12 and DIG16). DIG10.2 and DIG10.3 are third and fourth generation designs, respectively, generated following design optimization, and exhibit binding constants in the nano-molar to pico-molar range (Table 1). The structures of DIG10.2 and DIG10.3 have been solved by X-ray crystallography and were the starting structures of the simulations. DIG12 and DIG16, being considered failed designs, did not have their structures solved experimentally and the starting structures for the simulations were modeled by Rosetta based on these designs sequences and the original scaffold from which they were engineered[5].

We first focused on the dynamics of the proteins in the simulations. The projection of the RMSF values on the protein structures evidences that the highly fluctuating regions are located in the structural motifs lining up the cavity entrance, with DIG12 in particular exhibiting a larger flexible region than the tight binders and thus suggesting a possible negative effect of the protein flexibility in the accessibility of the cavity for ligand binding (Supplementary Figure 1). We also looked at solvent accessibility of hydrophobic residues, since this is an important factor affecting protein stability. Figure 2a shows the average SASA/hydrophobic residue calculated for the designs. In line with what would be expected, tight binders show smaller SASA both in the apo and holo simulations, although DIG16 values are not as distinct from the tight binders as DIG12. This indicates that the successful designs not only tend to have a better organized hydrophobic core, but also confirms the importance of solvent shielding of nonpolar residues and promotion of hydrophobic interactions for adequate structural stability.

A key unanswered question in the design process of proteins functionalized for ligand binding is how stable and pre-organized the pocket remains in the absence of the ligand[35]. We set out to explore the survival of the organized pocket through the calculation of cavity volume throughout the simulations. All designs, regardless of binding affinity, showed large variations of cavity volume in the apo state, indicating that the pocket deviates from its designed conformation but that does not necessarily preclude ligand binding (Supplementary Figure 2a). The non-binders DIG12 and DIG16 have occasional complete closure of the cavity, but the same also happens for the successful binder DIG10.2. The volume density of DIG10.2 at 50% of the frames, for example, shows a partially collapsed pocket (Supplementary Figure 2b). Of the designs, DIG10.3 is the only one that shows apo pocket volume density that completely encompasses the volume occupied by the ligand in the bound state, demonstrating the pocket pre-organization achieved in the last round of design optimization. The C-alpha RMSF values (Supplementary Figure 1) from the MD simulations indicate that the lack of cavity pre-organization among the other designs is not only due to sidechain flexibility but also reflects backbone-level dynamics, which overcomes a major limitation of protein design protocols of not accounting for backbone flexibility.

On average, we observe much larger cavity volumes for the non-binders DIG12 and DIG16 than for the tight binders, both in the apo and holo simulations (Figure 2b). These designs' scaffolds have a large cavity opening, while the scaffold of DIG10, from which DIG10.2 and 10.3 were generated, presents a more enclosed cavity, such that the cavity volume results are a reflection of this. The simulations of DIG10.2 with ligand bound sampled a high number of cavity conformations with volumes smaller than what was originally designed, suggesting side chain rearrangements that result in a tighter interface around the ligand (Supplementary Figure 2a). The fact that the designs with the sterically most accessible cavities resulted in the lowest affinities with the ligand sheds light on an interesting question: cavity accessibility may not be as important a factor for ligand affinity given these designs innate flexibility, and a "close-fitting" pocket may play a bigger role as it allows for stronger interactions with the ligand when in the bound state.

We furthered our study of pocket pre-organization by looking at the dihedral angles sampled by the residues side chains specifically designed to hydrogen bond to the ligand. The DIG

designs present 3 interacting residues at the interior of the cavity in each of the monomer chains (Y34, Y101 and Y115 for DIG10.2 and DIG10.3, W57, H60 and H67 for DIG12 and Y39, H41 and N89 for DIG16), and we found that these remain in their designed conformer for a larger fraction of frames in the successful design simulations (Supplementary Figure 3).

Besides probing protein dynamics, the simulations provide interesting insights into the dynamics of the ligand as well. In the holo simulations of the tight binders DIG10.2 and DIG10.3, very small ligand fluctuations are seen, with the ligands remaining very tightly bound in their original conformation in the cavity (Figure 2c). In the simulations of the non-binder examples, on the other hand, the ligand showed a large degree of displacement from its starting position, probably influenced by the larger size of the cavity, including complete dissociation from one of the monomers in 2 out of the 5 holo DIG16 trajectories (Supplementary Figure 4). Somewhat surprisingly, the MD simulations were thus able to distinguish between tight binders and non-binder designs without requiring any ligand steering or information on the design's binding affinity.

We further investigated ligand-protein interaction by counting the number of hydrogen bonds stablished in the holo simulations. While there's a lot of fluctuation in the number of hydrogen bonds due to the dynamics of the ligands and side chains, DIG10.2 and DIG10.3 show a larger average number of hydrogen bonds than DIG12 and DIG16 (Figure 2d). For these successful designs, 3 stable hydrogen bonds are maintained with the designed interacting side chains located at the binding site interior, while additional transient hydrogen bonds are occasionally established with the ligand moiety located at the more solvent exposed opening of the cavity. Interestingly, DIG12 also establishes a large number of transient hydrogen bonds besides those modeled in the design structure due to its larger ligand dynamics.

Binding is not only highly influenced by the direct interactions between the protein and ligand, but also by the dynamics of the water molecules surrounding them[60]. Consequently, we also looked at the water molecules present in the apo and holo cavity interiors to investigate if there were any differences in water organization between the designs. The protein preparation steps preceding MD production of the apo state involved using Dowser[41] to incorporate water molecules into the void left by the removal of the ligand from the model structure. We accompanied the presence of waters in the binding site by counting the number of molecules within a sphere delimiting the protein cavity. The holo simulations showed a smaller degree of water insertion in the cavity than the apo simulations due to the presence of the ligand (Figure 2e). In both states, the non-successful designs allowed for a greater degree of water insertion, promoted by the larger pocket volumes sampled during these simulations (Figure 2b).

Finally, as the absolute water count inside the cavity likely does not provide the full picture of the energetics of interactions, we calculated the survival probability of the waters inside the binding pocket in the apo simulations to get information on the stability of these molecules. As seen in Figure 2f, survival probability for the tight DIG binders decay more slowly than that for the non-binders, indicating presence of longer-lived waters inside these

cavities and that successful design strategies involve promotion of favorable protein-water interactions.

Dynamics of modeled versus resolved crystal structures

A question might arise in the application of MD simulations for design screening regarding the accuracy and reliability of the results obtained from possibly inaccurately modeled starting structures. This is a particularly valid concern since the prospective application of the simulations involves using structures predicted by Rosetta or some other protein design software that are not experimentally validated, as it antecedes experimental assays. In the above section, we described the dynamics and results obtained from the simulations of the crystal structures of DIG10.2 and DIG10.3. To try to address this question, we also performed simulations starting from the corresponding Rosetta-modelled structures for these designs, here named DIG10.2a and DIG10.3a. Simulations were run for 500 ns both in apo and holo states, and here we compare results from equivalent simulation lengths of the crystallographic structures. Overall, the dynamics obtained from the modeled structures showed similar distribution profiles to those derived from the simulations of the crystal structures (Figure 3). Hydrophobic SASA average values of modeled and resolved structure simulations are very similar to each other, as well as pocket volume distributions (Figures 3a and 3b). Water count inside the cavity is also comparable between modeled and resolved structure simulations (Figure 3c). In terms of ligand RMSD, DIG10.3a showed a significant tail of higher ligand RMSD values (Figure 3d), due to a large transient ligand displacement in one of the runs, but nonetheless the distributions sampled are still very distinct from that observed for the non-binders DIG12 and DIG16 (Figure 2c). Our results suggest that simulations starting from modeled structures are therefore accurate enough in sampling protein dynamics to be used with confidence in the assessment of these designs.

Validation on a distinct scaffold

The analysis of the DIG designs suggests the existence of energetic factors influencing binding ability which manifest themselves in key dynamical properties exhibited by successful binders. To validate these findings in a larger dataset and test the universality of the properties, we performed simulations of 33 experimentally-validated designs of a β-barrel scaffold[7]. These *de-novo* designed proteins not only represent a completely different protein scaffold than the DIG binders, but have also been designed to bind to a distinct small-molecule ligand, DFHBI[7]. Our data set includes 24 first-generation designs, which were all predicted to be tight binders by the computational design methodology even though the majority was experimentally found to not be so: two were verified to be structurally unstable and thus not fold in the predicted β-barrel structure (HBI_38 and HBI_41), 20 fold properly but fail to bind to the ligand, and only two are successful tight binders, with ligand affinity values in the micro-molar range (Table 1). In the work of Dou *et al* these successful initial designs were further optimized, resulting in 9 second and third-generation designs with higher ligand affinity which have also been simulated and included in our analysis here (Table 1).

To first verify the necessary sampling time required for appropriate distinction between the designs, we performed convergence analysis of the initial DIG design results as well as a

small set of the β-barrel designs (containing the 2 first-generation tight binders and 3 non-binders) which were run for five 1 μs-long replicates in the apo and holo states, the same sampling length used for the previous designs. Analysis of the identified dynamical features indicated that the simulations do not need to be run so extensively, with results either reaching approximately constant values or maintaining constant relation among each other at around 500 ns (Supplementary Figure 5). Moreover, estimates of the reliability and reproducibility achieved using different number and combination of replicates[59] indicates that three or four replicas yield property mean values satisfactorily converged and independent, within small variations, of the identity of the replicate simulations (Supplementary Figure 6 shows results for HBI_10 and HBI_11). A key point in our exploration is that we do not intend to perform an exhaustive investigation of the design dynamics, as this would likely require extremely long simulations and defeat the purpose of using molecular dynamics to increase the efficiency of the design process. Instead, we aimed at obtaining sufficient sampling for insightful discrimination between the large number of design candidates. Therefore, in the interest of time efficiency, we performed subsequent simulations of the remaining β-barrel designs as 500 ns triplicate runs in the apo and holo states, and the following results will be discussed for equivalent sampling times for all simulations.

In the analysis of this larger dataset, it became evident the need for an additional feature that would describe the conformational flexibility of the different designs. While RMSF is useful to investigate structural fluctuation incurred during the simulations, we turned to RMSD clustering of the Cα coordinates to provide a single value to represent each design's flexibility and thus allow for a more direct comparison across the different proteins. As in the work of Demir *et al.*, the number of clusters (NOCs) thus obtained was used as a representation of structural flexibility since at least in principle more flexible proteins sample a larger conformational ensemble in the simulations, resulting in a higher number of clusters to represent the variation of the Cα positions[61]. Figure 4 shows the 33 designs in terms of the structural and dynamical properties discussed above: structural stability (represented by the number of clusters and SASA of hydrophobic residues), cavity pre-organization (probed by number of frames in the apo simulations with volumes smaller than a cutoff which would prevent ligand binding, and holo average volume), insertion of waters in the cavity in holo simulations, and ligand dynamics (in terms of ligand RMSD and average number of protein-ligand hydrogen bonds per frame).

The profiles in Figure 4 support and accentuate the trends observed from our initial reduced data set and evidence that these structural and dynamical descriptors can be useful in the classification of candidate designs. Importantly, the analysis of the same features computed from the original Rosetta-modelled design structures does not evidence any such distinction between the design categories (Supplementary Figure 7), such that the descriptors generated from the static structures are not sufficient to distinguish successes from failures. The incorporation of dynamics, however, indicates that the non-successful designs in general are much more flexible and explore a wider range of conformations (Figure 4a, some designs have equal values of apo and holo NOC and overlap in the graph), suggesting that for this scaffold, failure to bind to the small ligand may arise from the lack of accounting for structure dynamics in the structure prediction methods. We observed that several non-

binders were structurally destabilized by the introduction of the ligand in the holo simulations, leading to some dramatic structural deformations in some cases (Supplementary Figure 8b). Tight binders, on the other hand, tended to be stabilized by the ligand in the holo simulations as indicated by the dampening of fluctuations in the RMSF plots (Supplementary Figure 8). The number of clusters analysis is particularly promising in that it may allow for early identification of non-stable designs, since HBI_41, one of the two designs that did not fold in the β-barrel structure in our data set, displayed one of the largest number of cluster pair values. The solvent exposure of hydrophobic residues does not allow for such a clear distinction between the design classes, but it's possible to see in Figure 4b the suggestion of an empirical threshold at around 0.29 nm^2 apo and holo SASA beyond which only non-successful designs can be found.

The apo simulations of some of these designs showed such a large number of frames with completely collapsed pockets that it became evident that another useful discriminating metric would be something that could capture this phenomenon. Here, we chose that as the number of frames with cavity volumes below a cutoff of 30 $Å^3$, which represents a pocket volume too small to allow for ligand binding (the smallest pocket volume observed in the simulations with ligand bound was 33 $Å^3$). This descriptor, in conjunction with the average cavity volume in the holo simulations, permits successful distinction between most of the designs (Figure 4c). Comparison of the same average cavity volume with the number of water molecules that insert into the pocket in the holo simulations evidences that while there's a lot of variability for the non-binders, the successful designs cluster around smaller cavities and a reduced number of inserted water molecules (Figure 4d).

Finally, as for the DIG designs, probing ligand dynamics also provides valuable information for design identification (Figure 4e). While several outliers can be seen, all successful designs show a higher number of ligand-protein hydrogen bonds and reduced ligand dynamics as indicated by the low ligand RMSD values. The incorporation of protein and ligand dynamics into these designed scaffolds provide important additional information that can thus aid design selection, since all generated designs had been originally intended to form at least four hydrogen-bonding interactions with the ligand[7].

Discriminative models for design screening

For some of the features in Figure 4 it is possible to imagine cutoffs of acceptable or promising values exhibited by proteins with favorable ligand interaction that could be used for prospective design predictions. However, as would be imagined from the complexity of the process investigated, each of these descriptors is not perfect in its discernment of binding ability, and we can see the likelihood of both false positive and false negative assignments. We hypothesized that taking the features jointly into account would result in a more accurate design classification, given the multi-dimensionality of the problem. We performed Principal Component Analysis (PCA) on the scaled features and projected the β-barrel dataset into three principal components (PC) which describe 80% of the data variance (Figure 5a). Confirming our hypothesis, the successful binders cluster together in regions of smaller PC1 and PC2, while the non-binders are more spread along the principal components. This is in line with the general notion that protein-ligand binding can be negatively impacted by

several causes, and that only a specific (almost serendipitous) combination of the properties result in a tight interaction.

The contributions of each of the features to the principal components can be analyzed to try to rationalize the energetic causes most highly affecting ligand-binding (Figure 5b). Entropy seems to play the most pronounced role in determining binding, as properties such as water dynamics in the cavity, protein conformational flexibility, cavity volume and ligand dynamics show the highest contributions in the first principal component. Ligand-protein induced fit comes in as a second determining factor, with the number of frames with too-small cavity volumes to allow for ligand insertion showing negative correlation with design binding ability. Finally, enthalpic components appear in the third PC, encoded by the SASA and number of hydrogen bonds established with the ligand.

Even though the successful and non-binder designs concentrated in different areas of the PC map, the separation is not absolute and there are overlaps or outliers among the two classes. Interestingly, the two first generation successful binders are the ones located closer to the area occupied by the non-binders, while the second and third generation higher affinity binders cluster more closely together, evidencing the successful enhancement of the energetic properties by the experimental optimization. While by visual inspection it can be hard to define a separating line between the classes, we turned to unsupervised learning and clustering to see if such regions could be determined in an unbiased manner. Using the k-means algorithm, the designs were not accurately clustered when only two clusters were used, but assigning the data to three distinct clusters yielded interesting results with good clustering quality (average silhouette value of 0.45): One of the clusters was enriched in designs from the tight binding class, while the others contained only examples from the non-binding designs (Figure 5c).

One of the clusters, located in the area of higher PC1 values, included designs that showed clearly unstable dynamics from the simulations (such as the designs shown in Supplementary Figure 8a), and can be interpreted as the Failed (F) cluster. The second cluster of only non-binders contained members in the boundary region with the tight binders and exhibited dynamics that would be hard to be accurately classified by visual inspection of the simulations. For this reason, we termed this the Uncertain (U) cluster. Their main distinction from the successful binders is captured by the collapse of the cavity in apo simulations incorporated into Principal Component 2, as all of these designs show small pocket volumes or completely closed cavities for significant portions of the simulations. Finally, the cluster to which all of the tight binders were assigned, here termed Successful (S) cluster, contains only 4 incorrectly classified designs. One of the false positives in this classification is HBI_38, the non-stable design that did not display as different feature profiles as HBI_41 in Figure 4. HBI_38 and the tight binder HBI_11 differ by only two mutations in the N terminal that lead to the formation of a stabilizing intramolecular disulfide bridge in HBI_11, and thus the misclassification of HBI_38 is not surprising given the likely much longer timescale that would be required to properly sample the difference between these designs structural stability.

Remarkably for such a complex problem, the unsupervised learning approach here employed on the features measured from the simulations was thus able to identify the high affinity binders with only 4 inaccurate classifications and no false negative assignments. We estimate that the early identification of the 12 unsuccessful designs from the F cluster and the 6 designs from the U cluster could have saved about 6 weeks of work, including protein expression, purification, folding and binding assays. However, this is likely to be a low-bound estimate as the Baker lab is very well equipped for protein characterization and the entire process could probably take 2 to 3 times longer in a different lab. On the other hand, the MD system preparation, simulation and analysis workflow greatly automates the required steps such that the whole set of proteins can be simulated and analyzed in less than 2 weeks, using parallel GPUs and requiring minimal human intervention.

This unsupervised approach is useful to identify inherent differences between designs of the same structural scaffold, but lacks transferability with the DIG design results (Supplementary Figure 9). However, taking advantage of the availability of experimentally-validated labels, we explored the use of supervised learning for the classification of the joint design scaffold[62,63]. K-nearest neighbors and logistic regression classifiers were trained using 5-fold cross validation on the 10 dynamical fingerprints identified in the joint, 37 β-barrel and DIG design simulations, and showed good classification performance (Table 2). The precision values, the rate of true positive classifications over all positive assignments (including false positives), indicate the presence of misclassified non-binders. However, the recall metric at 1.0 for both algorithms, given as the ratio of true positive assignments over all assignments of the real positive class (including false negatives), indicates a complete absence of tight binders being classified as unsuccessful designs. In the same way, the high accuracy of the classifications and the Matthew correlation coefficient (MCC) and F_1 scores, all used as measures of a classifier performance and with a maximum value of 1.0 for a perfect classification, evidence the generality of the proposed approach. Moreover, the feature weights of the logistic regression model indicate that pocket dynamics plays the most determinant role for identifying non-binders, quantified by the insertion of water molecules in the cavity when in the ligand-bound state and the collapse of the cavity when in the absence of the ligand (Supplementary Figure 10). Logistic regression, in particular, resulted in good classifiers even when trained on small sets (50% or even 33% of the dataset, Supplementary Table 1), suggesting that not many designs need to be experimentally validated in order to yield accurate predictions in a prospective study.

Finally, to further test the universality of this approach, we constructed models solely on the β-barrel designs and checked the predictions on the DIG dataset. With a large set of β-barrel designs available, we further split the data into training and validation sets to verify absence of overfitting. Using logistic regression, training the model on 70% of the β-barrel designs yields perfect classification of the designs of the distinct DIG scaffold (Table 3). Conversely, models trained solely on the 4 DIG designs display lower accuracy and precision due to the much smaller training set in this case, but the recall still indicates a perfect absence of false negative classification (Table 4). Interestingly, the 12 non-binders correctly identified correspond exactly to the designs classified in the F cluster using unsupervised learning. Regardless of the classification approach employed, the computation of dynamical fingerprints[64] from molecular dynamics simulations of designed proteins, thus, emerges as a

potential general and scaffold-independent screening methodology to aid the challenging protein design process (Figure 1b).

CONCLUSIONS

In this work, we used MD simulations to investigate the dynamics of designed ligand-binding proteins as a source of insight into the failure of some of these designs to bind to the ligand with high affinity. It became evident that the design model generated by the protein design protocol may differ from the ensemble of structures accessed by the simulations, such that the modeled structural descriptions can be further enriched by the incorporation of dynamic fingerprints.

The results obtained here suggest that successful and non-successful designs differ in their dynamical properties. Entropic components play a significant role in determining ligand affinity, which are complex and often very challenging to incorporate in the empirical models of protein design. Easily measured MD-realized descriptors (including number of clusters, cavity volume, hydrophobic solvent-accessible surface area, water count in cavity and number of protein-ligand hydrogen bonds) allow for the investigation of multiple design candidates, and analysis of these enthalpic and entropic feature profiles in a data set of 33 β-barrel designs resulted in a 88% accuracy of binding ability classification using unsupervised learning. This data set included 24 first-generation designs, among which only two were found to bind with high affinity, and 9 optimized second and third-generation designs[7]. The application of the unsupervised learning method in the screening of the first generation designs would result in the identification of the two successful binders and 4 false positive non-binders, which constitutes a 4-fold enrichment ((2/6)/(2/24)) over the initial candidate design data set and a minimum net time and effort “savings” of one month of work. Moreover, the application of supervised learning in the form of k-nearest neighbors or logistic regression classifiers on the full dataset consisting of two different protein scaffolds resulted in accurate classification with no false negatives, suggesting the generality of this approach. The results here described emphasize how MD can act as a promising screening step in the protein design process, avoiding the experimental testing of non-stable and low affinity designs and increasing the efficiency of the pipeline.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

EPB thanks JiaJie Xiao and Jamel Meslamani for helpful discussions. This work was funded in part by the National Biomedical Computation Resource (NBCR) through NIH P41 GM103426. REA has an equity interest in, and is a cofounder and on the scientific advisory board of, Actavalon, Inc.

REFERENCES

(1). Huang P; Boyken SE; Baker D The Coming of Age of de Novo Protein Design. Nature 2016, 537, 320–327. [PubMed: 27629638]

(2). Woolfson DN; Bartlett GJ; Burton AJ; Heal JW; Niitsu A; Thomson AR; Wood CW De Novo Protein Design: How Do We Expand into the Universe of Possible Protein Structures? Curr. Opin. Struct. Biol 2015, 33, 16–26. [PubMed: 26093060]

(3). Huang P-S; Feldmeier K; Parmeggiani F; Fernandez Velasco DA; Höcker B; Baker D De Novo Design of a Four-Fold Symmetric TIM-Barrel Protein with Atomic-Level Accuracy. Nat. Chem. Biol 2016, 12, 29–34. [PubMed: 26595462]

(4). Lin Y-R; Koga N; Tatsumi-Koga R; Liu G; Clouser AF; Montelione GT; Baker D Control over Overall Shape and Size in de Novo Designed Proteins. Proc. Natl. Acad. Sci. USA 2015, 112, E5478–E5485. [PubMed: 26396255]

(5). Tinberg CE; Khare SD; Dou J; Doyle L; Nelson JW; Schena A; Jankowski W; Kalodimos CG; Johnsson K; Stoddard BL; Baker D Computational Design of Ligand-Binding Proteins with High Affinity and Selectivity. Nature 2013, 501, 212–216. [PubMed: 24005320]

(6). Thomas F; Dawson WM; Lang EJM; Burton AJ; Bartlett GJ; Rhys GG; Mulholland AJ; Woolfson DN De Novo-Designed A-helical Barrels as Receptors for Small Molecules. ACS Synth. Biol 2018, 7, 1808–1816. [PubMed: 29944338]

(7). Dou J; Vorobieva AA; Sheffler W; Doyle LA; Park H; Bick MJ; Mao B; Foight GW; Lee MY; Gagnon LA; Carter L; Sankaran B; Ovchinnikov S; Marcos E; Huang P-S; Vaughan JC; Stoddard BL; Baker D De Novo Design of a Fluorescence-Activating β -Barrel. Nature 2018, 561, 485–491. [PubMed: 30209393]

(8). Bjelic S; Nivon LG; Çelebi-Ölçum N; Kiss G; Rosewall CF; Lovick HM; Ingalls EL; Gallaher JL; Seetharaman J; Lew S; Montelione GT; Hunt JF; Michael FE; Houk KN, Baker D Computational Design of Enone-Binding Proteins with Catalytic Activity for the Morita-Baylis-Hillman Reaction. ACS Chem. Biol 2013, 8, 749–757. [PubMed: 23330600]

(9). Burton AJ; Thomson AR; Dawson WM; Brady RL; Woolfson DN Installing Hydrolytic Activity into a Completely de Novo Protein Framework. Nat. Chem 2016, 8, 837–844. [PubMed: 27554410]

(10). Strauch E-M; Bernard SM; La D; Bohn AJ; Lee PS; Anderson CE; Nieusma T; Holstein CA; Garcia NK; Hooper KA; Ravichandran R; Nelson JW; Sheffler W; Bloom JD; Lee KK; Ward AB; Yager P; Fuller DH; Wilson IA; Baker D Computational Design of Trimeric Influenza-Neutralizing Proteins Targeting the Hemagglutinin Receptor Binding Site. Nat. Biotechnol 2017, 35, 667–671. [PubMed: 28604661]

(11). Chevalier A; Silva DA; Rocklin GJ; Hicks DR; Vergara R; Murapa P; Bernard SM; Zhang L; Lam K; Yao G; Bahl CD, Miyashita S-I; Goreshnik I; Fuller JT; Koday MT; Jenkins CM; Colvin T; Carter L; Bohn A; Bryan CM; Fernández-Velasco DA; Steward L; Dong M; Huang X; Jin R; Wilson IA; Fuller DH; Baker D Massively Parallel de Novo Protein Design for Targeted Therapeutics. Nature 2017, 550, 74–79. [PubMed: 28953867]

(12). Koday MT; Nelson J; Chevalier A; Koday M; Kalinoski H; Stewart L; Carter L; Nieusma T; Lee PS; Ward AB; Wilson IA; Dagley A; Smee DF; Baker D; Fuller DH A Computationally Designed Hemagglutinin Stem-Binding Protein Provides in Vivo Protection from Influenza Independent of a Host Immune Response. PLOS Pathog 2016, 12, e1005409. [PubMed: 26845438]

(13). King NP; Sheffler W; Sawaya MR; Vollmar BS; Sumida JP; André I; Gonen T; Yeates TO; Baker D Computational Design of Self-Assembling Protein Nanomaterials with Atomic Level Accuracy. Science. 2012, 336, 1171–1174. [PubMed: 22654060]

(14). Fletcher JM; Harniman RL; Barnes FRH; Boyle AL; Collins A; Mantell J; Sharp TH; Antognozzi M; Booth PJ; Linden N; Miles MJ; Sessions RB; Verkade P; Woolfson DN Self-Assembling Cages from Coiled-Coil Peptide Modules. Science. 2013, 340, 595–599. [PubMed: 23579496]

(15). Hsia Y; Bale JB; Gonen S; Shi D; Sheffler W; Fong KK; Nattermann U; Xu C; Huang P-S; Ravichandran R; Yi S; Davis TN; Gonen T; King NP; Baker D Design of a Hyperstable 60-Subunit Protein Icosahedron. Nature 2016, 535, 136–139. [PubMed: 27309817]

(16). Bale JB; Gonen S; Liu Y; Sheffler W; Ellis D; Thomas C; Cascio D; Yeates TO; Gonen T; King NP; Baker D Accurate Design of a Megadalton-Scale Two-Component Icosahedral Protein Complexes. Science. 2016, 353, 389–394. [PubMed: 27463675]

(17). Regan L; Caballero D; Hinrichsen MR; Virrueta A; Williams DM; Hern CSO Protein Design: Past, Present, and Future. Pept. Sci 2015, 104, 334–350.

(18). Martin SF; Clements JH Correlating Structure and Energetics in Protein-Ligand Interactions: Paradigms and Paradoxes. Annu. Rev. Biochem 2013, 82, 267–293. [PubMed: 23746256]

(19). Zanghellini A De Novo Computational Enzyme Design. Curr. Opin. Biotechnol 2014, 29, 132–138. [PubMed: 24794534]

(20). Feng J; Jester BW; Tinberg CE; Mandell DJ; Antunes MS; Chari R; Morey KJ; Rios X; Medford JI; Church GM; Fields S; Baker D A General Strategy to Construct Small Molecule Biosensors in Eukaryotes. Elife 2015, 4, e10606. [PubMed: 26714111]

(21). Bick MJ; Greisen PJ; Morey KJ; Antunes MS; La D; Sankaran B; Reymond L; Johnsson K; Medford JI; Baker D Computational Design of Environmental Sensors for the Potent Opioid Fentanyl. Elife 2017, 6, e28909. [PubMed: 28925919]

(22). Roy A; Nair S; Sen N; Soni N; Madhusudhan MS In Silico Methods for Design of Biological Therapeutics. Methods 2017, 131, 33–65. [PubMed: 28958951]

(23). Entzminger KC; Hyun J; Pantazes RJ; Patterson-Orazem AC; Qerqez AN; Frye ZP; Hughes RA; Ellington AD; Liebermanl RL; Maranas CD; Maynard JA De Novo Design of Antibody Complementarity Determining Regions Binding a FLAG Tetra- Peptide. Sci. Rep 2017, 7, 10295. [PubMed: 28860479]

(24). Mondal J; Friesner RA; Berne BJ Role of Desolvation in Thermodynamics and Kinetics of Ligand Binding to a Kinase. J. Chem. Theory Comput 2014, 10, 5696–5705. [PubMed: 25516727]

(25). Dou J; Doyle L; Greisen PJ; Schena A; Park H; Johnsson K; Stoddard BL; Baker D Sampling and Energy Evaluation Challenges in Ligand Binding Protein Design. Protein Sci 2017, 26, 2426–2437. [PubMed: 28980354]

(26). Kiss G; Pande VS; Houk KN Molecular Dynamics Simulations for the Ranking, Evaluation, and Refinement of Computationally Designed Proteins. In Methods Enzymol; 2013; Vol. 523, pp 145–170. [PubMed: 23422429]

(27). Childers MC; Daggett V Insights from Molecular Dynamics Simulations for Computational Protein Design. Mol. Syst. Des. Eng 2017, 2, 9–33. [PubMed: 28239489]

(28). Privett HK; Kiss G; Lee TM; Blomberg R; Chica RA; Thomas LM; Hilvert D; Houk KN; Mayo SL Iterative Approach to Computational Enzyme Design. Proc. Natl. Acad. Sci. USA 2012, 109, 3790–3795. [PubMed: 22357762]

(29). Lindert S; Mccammon JA Improved cryoEM-Guided Iterative Molecular Dynamics-Rosetta Protein Structure Refinement Protocol for High Precision Protein Structure Prediction. J. Chem. Theory Comput 2015, 11, 1337–1346. [PubMed: 25883538]

(30). Leelananda SP; Lindert S Iterative Molecular Dynamics-Rosetta Membrane Protein Structure Refinement Guided by Cryo-EM Densities. J. Chem. Theory Comput 2017, 13, 5131–5145. [PubMed: 28949136]

(31). Yu H; Huang H Engineering Proteins for Thermostability through Rigidifying Flexible Sites. Biotechnol. Adv 2014, 32, 308–315. [PubMed: 24211474]

(32). Joo JC; Pack SP; Kim YH; Yoo YJ Thermostabilization of Bacillus Circulans Xylanase: Computational Optimization of Unstable Residues Based on Thermal Fluctuation Analysis. J. Biotechnol 2011, 151, 56–65. [PubMed: 20959126]

(33). Liu J; Yu H; Shen Z Insights into Thermal Stability of Thermophilic Nitrile Hydratases by Molecular Dynamics Simulation. J. Mol. Graph. Model 2008, 27, 529–535. [PubMed: 18948044]

(34). Chen J; Yu H; Liu C; Liu J; Shen Z Improving Stability of Nitrile Hydratase by Bridging the Salt-Bridges in Specific Thermal-Sensitive Regions. J. Biotechnol 2012, 164, 354–362. [PubMed: 23384947]

(35). Kiss G; Rothlisberger D; Baker D; Houk KN Evaluation and Ranking of Enzyme Designs. Protein Sci 2010, 19, 1760–1773. [PubMed: 20665693]

(36). Jacobson MP; Pincus DL; Rapp CS; Day TJF; Honig B; Shaw DE; Friesner RA; Friesner RA A Hierarchical Approach to All-Atom Protein Loop Prediction. Proteins 2004, 55, 351–367. [PubMed: 15048827]

(37). Jacobson MP; Friesner RA; Xiang Z; Honig B On the Role of the Crystal Environment in Determining Protein Side-Chain Conformations. J. Mol. Biol 2002, 320, 597–608. [PubMed: 12096912]

(38). Wang J; Wolf RM; Caldwell JW; Kollman PA; Case DA Development and Testing of a General Amber Force Field. J. Comput. Chem 2004, 25, 1157–1174. [PubMed: 15116359]

(39). Wang J; Wang W; Kollman PA; Case DA Automatic Atom Type and Bond Type Perception in Molecular Mechanical Calculations. J Mol Graph Model 2006, 25, 247–260. [PubMed: 16458552]

(40). Frisch MJ; Trucks GW; Schlegel HB; Scuseria GE; Robb MA; Cheeseman JR; Scalmani G; Barone V; Mennucci B; Petersson GA; et al. Gaussian 09, Revision C.01 Gaussian, Inc.: Wallingford CT 2010.

(41). Gumbart J; Trabuco LG; Schreiner E; Villa E; Schulten K Regulation of the Protein-Conducting Channel by a Bound Ribosome. Structure 2009, 17, 1453–1464. [PubMed: 19913480]

(42). Case DA; Babin V; Berryman JT; Betz RM; Cai Q; Cerutti DS; Cheatham TE, I.;Darden TA; Duke RE; Gohlke H; Goetz AW; Gusarov S; Homeyer N; Janowski P; Kaus J; Kolossváry I; Kovalenko A; Lee TS; LeGrand S; Luchko T; Luo R; Madej B; Merz KM; Paesani F; Roe DR; Roitberg A; Sagui C; Solomon-Ferrer R; Seabra G; Simmerling CL; Smith W; Swails J; Walker RC; Wang J; Wolf RM; Wu X; Kollman PA AMBER 14 University of California, San Francisco 2014.

(43). Maier JA; Martinez C; Kasavajhala K; Wickstrom L; Hauser KE; Simmerling C ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. J. Chem. Theory Comput 2015, 11, 3696–3713. [PubMed: 26574453]

(44). Jorgensen WL; Chandrasekhar J; Madura JD; Impey RW; Klein ML Comparison of Simple Potential Functions for Simulating Liquid Water. J. Chem. Phys 1983, 79, 926–932.

(45). Radak BK; Chipot C; Suh D; Jo S; Jiang W; Phillips JC; Schulten K; Roux B Constant-pH Molecular Dynamics Simulations for Large Biomolecular Systems. J. Chem. Theory Comput 2017, 13, 5933–5944. [PubMed: 29111720]

(46). Purawat S; Ieong PU; Malmstrom RD; Chan GJ; Yeung AK; Walker RC; Altintas I; Amaro RE A Kepler Workflow Tool for Reproducible AMBER GPU Molecular Dynamics. Biophys. J 2017, 112, 2469–2474. [PubMed: 28636905]

(47). Salomon-Ferrer R; Go AW; Poole D; Grand S. Le; Walker RC Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. J. Chem. Theory Comput 2013, 9, 3878–3888. [PubMed: 26592383]

(48). Darden TA; York D; Pedersen L Particle-Mesh Ewald: An N.log(N) Method for Ewald Sums in Large Systems. J. Chem. Phys 1993, 98, 10089–10092.

(49). Humphery W; Dalke A; Schulten K VMD-Visual Molecular Dynamics. J. Molec. Graph 1996, 14, 33–38. [PubMed: 8744570]

(50). Kluyver T; Ragan-Kelley B; Pérez F; Granger B; Bussonnier M; Frederic J; Kelley K; Hamrick J; Grout J; Corlay S; Ivanov P; Avila D; Abdalla S; Willing C Jupyter Notebooks - a Publishing Format for Reproducible Computational Workflows. Position. Power Acad. Publ. Play. Agents Agendas 2016, 87–90.

(51). Mcgibbon RT; Beauchamp KA; Harrigan MP; Klein C; Swails JM; Hernandez CX; Schwantes CR; Wang L-P; Lane TJ; Pande VS MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. Biophys. J 2015, 109, 1528–1532. [PubMed: 26488642]

(52). Roe DR; Cheatham TE PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. J. Chem. Theory Comput 2013, 9, 3084–3095. [PubMed: 26583988]

(53). Scherer MK; Trendelkamp-Schroer B; Paul F; Perez-Hernandez G; Hoffmann M; Plattner N; Wehmeyer C; Prinz J-H; Noe F PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. J. Chem. Theory Comput 2015, 11, 5525–5542. [PubMed: 26574340]

(54). Michaud-Agrawal N; Denning EJ; Woolf TB; Beckstein O MDAnalysis : A Toolkit for the Analysis of Molecular Dynamics Simulations. J. Comput. Chem. Chem 2011, 32, 2319–2327.

(55). Gowers RJ; Linke M; Barnoud J; Reddy TJE; Melo MN; Seyler SL; Domanski J; Dotson DL; Buchoux S; Kenney IM; et al. MDAnalysis : A Python Package for the Rapid Analysis of Molecular Dynamics Simulations. In Proc. of the 15th Python in Science Conf.; 2016; pp 98–105.

(56). Durrant JD; Votapka L; Amaro RE POVME 2.0: An Enhanced Tool for Determining Pocket Shape and Volume Characteristics. J. Chem. Theory Comput 2014, 10, 5047–5056. [PubMed: 25400521]

(57). Steiner T The Hydrogen Bond in the Solid State. Angew. Chem. Int. Ed 2002, 41, 48–76.

(58). Liu P; Harder E; Berne BJ On the Calculation of Diffusion Coefficients in Confined Fluids and Interfaces with an Application to the Liquid - Vapor Interface of Water. J. Phys. Chem. B 2004, 108, 6595–6602.

(59). Knapp B; Ospina L; Deane CM Avoiding False Positive Conclusions in Molecular Simulations: The Importance of Replicas. J. Chem. Theory Comput 2018, 14, 6127–6138. [PubMed: 30354113]

(60). Mobley DL; Dill KA Binding of Small-Molecule Ligands to Proteins: “What You See” Is Not Always “What You Get.” Structure 2009, 17, 489–498. [PubMed: 19368882]

(61). Demir Ö; Baronio R; Salehi F; Wassman CD; Hall L; Hatfield GW; Chamberlin R; Lathrop RH; Amaro RE Ensemble-Based Computational Approach Discriminates Functional Activity of p53 Cancer and Rescue Mutants. PLoS Comput. Biol 2011, 7, e1002238. [PubMed: 22028641]

(62). Domingos P A Few Useful Things to Know about Machine Learning. Commun. ACM 2012, 55, 78–87.

(63). Chicco D Ten Quick Tips for Machine Learning in Computational Biology. BioData Min 2017, 10, 1–17. [PubMed: 28127402]

(64). Riniker S Molecular Dynamics Fingerprints (MDFP): Machine Learning from MD Data to Predict Free-Energy Differences. J. Chem. Inf. Model 2017, 57, 726–741. [PubMed: 28368113]

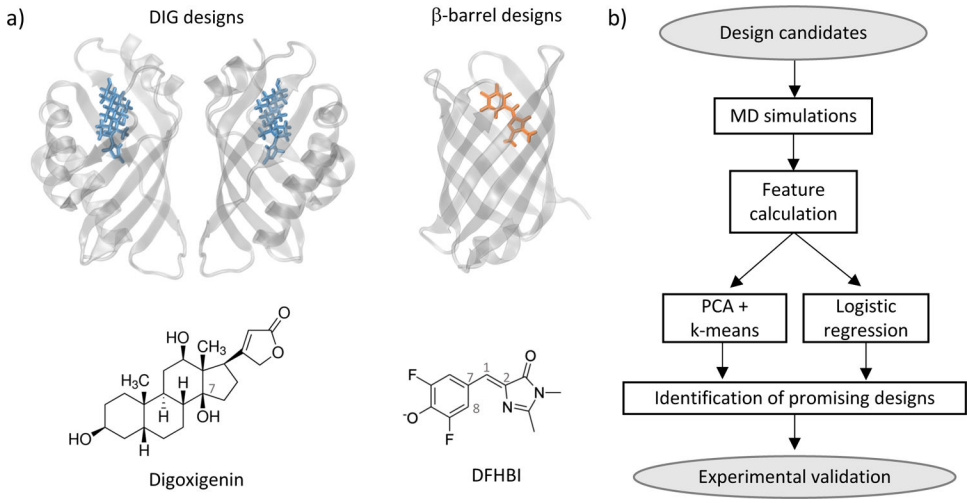


Figure 1.

(a) Summary of the design data set used in the simulations, constituting two protein scaffolds (β-barrels and DIG designs) designed to bind distinct ligands. Representative designs are shown on the top with ligands highlighted, and the structure of the ligands are shown on the bottom panels. DFHBI stands for fluorogenic 3,5-difluoro-4-hydroxybenzylidene imidazolinone. Key atoms mentioned later in the text are indicated by their respective numbering in the ligand molecule. **(b)** Schematics of the prosed use of MD as a screening tool in the protein design process.

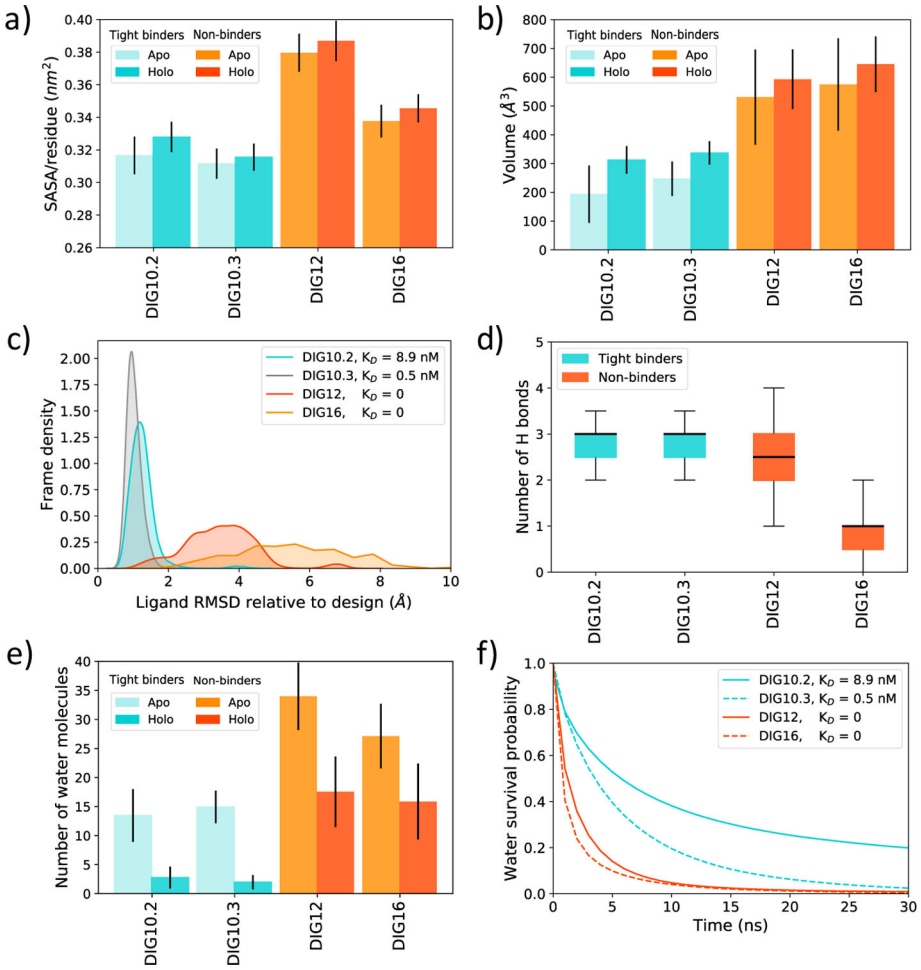


Figure 2. Evaluation of binding determinants for DIG designs

(**a)** Solvent accessible surface area (SASA) of hydrophobic residues. Tight binders are colored in turquoise and non-binders in orange. Results from apo simulations are shown in lighter colors, and holo simulations in darker. Error bars represent standard deviation across the five replicates. **(b)** Average cavity volumes for apo and holo simulations. Color scheme is the same as (a). **(c)** Ligand RMSD distribution for all replicates. **(d)** Box plot of the number of hydrogen bonds established between protein and ligand. Black line represents the mean value. The box extends to the lower and upper quartile and whiskers show the top 5 percentile and 95 bottom percentile of the data. **(e)** Average water count inside the cavity for apo and holo simulations. Coloring scheme is the same as (a). **(f)** Water survival probability in apo simulations for water molecules located within the cavity. Results are shown for one of the monomers only.

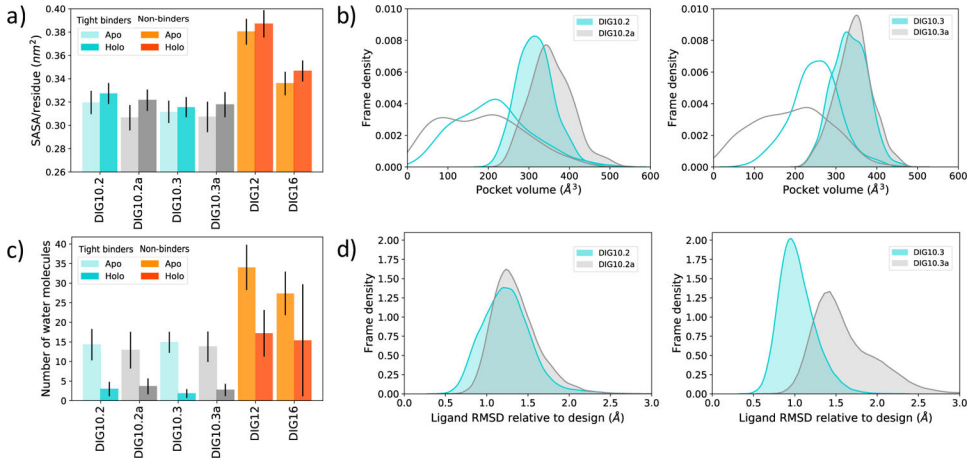


Figure 3. Comparison of results from the crystal (DIG10.2 and 10.3) and modeled structure (DIG10.2a and 10.3a) simulations

(a) Solvent accessible surface area (SASA) of hydrophobic residues. Tight binders are colored in turquoise (for simulations starting from the crystal structure) or gray (for simulations starting from the modeled structures) and non-binders in orange. Results from apo simulations are shown in lighter colors, and holo simulations in darker. Error bars represent standard deviation across the five replicates (**b)** Pocket volume distributions. DIG10.2 and DIG10.3 are represented in turquoise, DIG10.2a and DIG10.3a are shown in silver. Holo simulation results are shown with filled curves, and apo simulations with just the curve outline. **(c)** Cavity water count. Coloring scheme is the same as in (a) **(d)** Ligand RMSD distributions. Coloring scheme is the same as in (b).

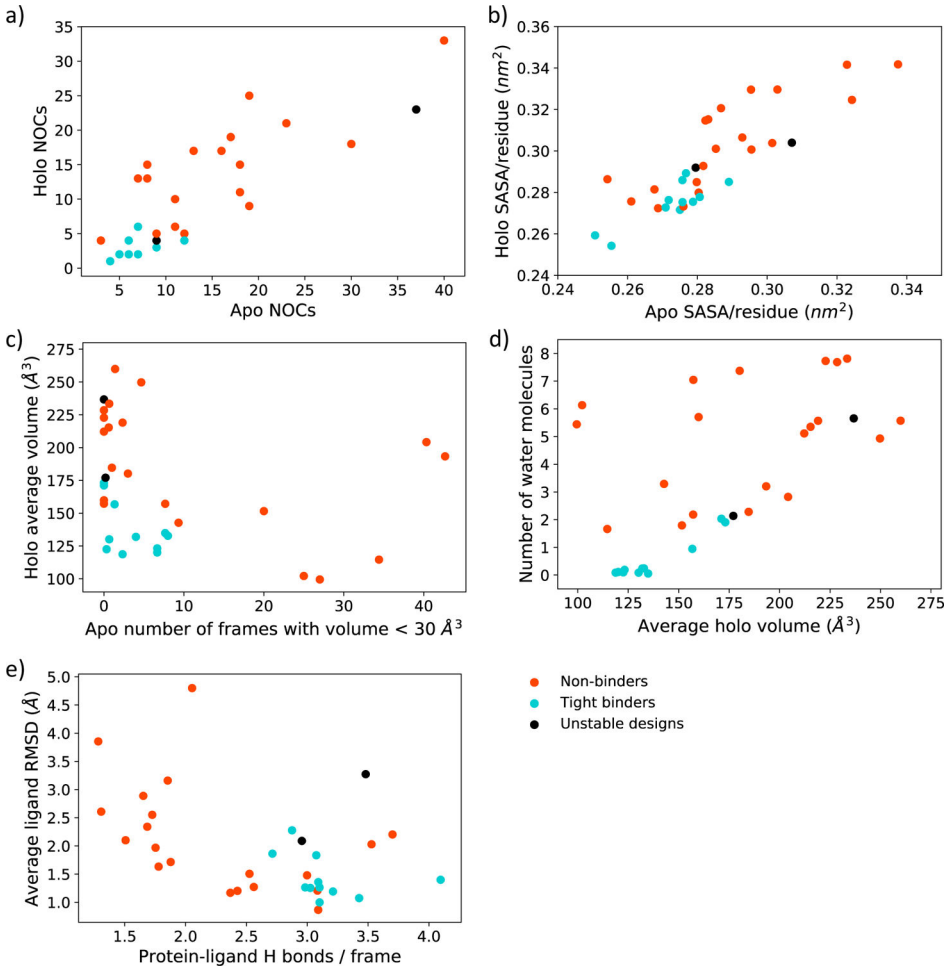


Figure 4. β-barrel designs distribution in terms of the identified discriminative features for design screening.

(a) Number of clusters (NOC) analysis, **(b)** SASA of hydrophobic residues, **(c)** Number of frames with volume below a cutoff of 30 $Å^3$ versus average holo cavity volume, **(d)** the same average holo cavity volume versus number of water molecules inserted in the pocket in the holo simulations and **(e)** average number of protein-ligand hydrogen bonds versus ligand RMSD. Non-binders are shown in orange, successful designs are shown in turquoise and structurally unstable designs in black.

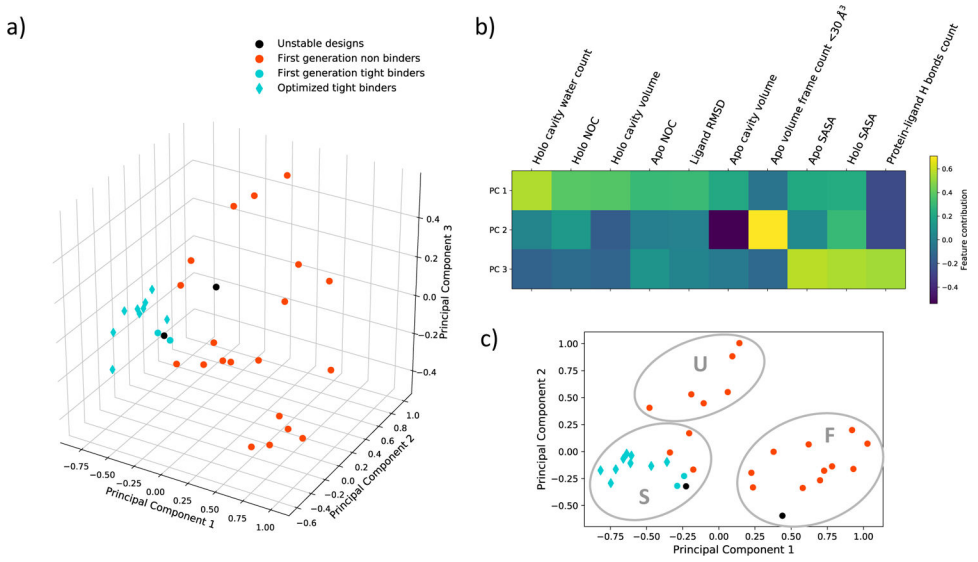


Figure 5. Unsupervised learning model for design classification.

(a) Three-dimensional plot of β-barrel simulations distribution in terms of the principal components of the discriminative features. Non-binders are shown in orange, structurally unstable designs in black, first-generation tight binders shown as turquoise circles and optimized tight binders as turquoise diamonds. **(b)** Color representation of feature contribution to the principal components. **(c)** Representation of cluster assignment on the 2-dimensional plot in terms of principal components 1 and 2. Clusters are named Successful (S), Uncertain (U) and Failed (F) clusters.

Table 1.

Summary of designed protein data set[5,7]

Design name	K_D	Classification*	Number of apo and holo replicates	Replicate simulation length (ns)
DIG10.2	8.9 nM	Binder	5	1,000
DIG10.2a	8.9 nM	Binder	5	500
DIG10.3	541 pM	Binder	5	1,000
DIG10.3a	541 pM	Binder	5	500
DIG12	-	Non-binder	5	1,000
DIG16	-	Non-binder	5	1,000
HBI_06	-	Non-binder	3	500
HBI_10	-	Non-binder	5	1,000
HBI_11	12.8 μM	Binder	5	1,000
HBI_15	-	Non-binder	3	500
HBI_19	-	Non-binder	3	500
HBI_21	-	Non-binder	3	500
HBI_22	-	Non-binder	3	500
HBI_24	-	Non-binder	3	500
HBI_26	-	Non-binder	5	1,000
HBI_27	-	Non-binder	3	500
HBI_32	49.8 μM	Binder	5	1,000
HBI_33	-	Non-binder	3	500
HBI_34	-	Non-binder	3	500
HBI_36	-	Non-binder	3	500
HBI_38	-	Unstable	3	500
HBI_41	-	Unstable	3	500
HBI_42	-	Non-binder	3	500
HBI_48	-	Non-binder	3	500
HBI_49	-	Non-binder	5	1,000
HBI_50	-	Non-binder	3	500
HBI_52	-	Non-binder	3	500
HBI_54	-	Non-binder	3	500
HBI_55	-	Non-binder	3	500
HBI_56	-	Non-binder	3	500
b11_loop	~0.5 μM**	Binder	3	500
b11L5F.1	~0.5 μM**	Binder	3	500
b11L5F_nC1	~0.5 μM**	Binder	3	500
b11L5F_nC2	~0.5 μM**	Binder	3	500
b11L5F_nC3	~0.5 μM**	Binder	3	500

Design name	K_D	Classification*	Number of apo and holo replicates	Replicate simulation length (ns)
b11L5F_nC4	~0.5 μM**	Binder	3	500
b11L5F.2	~0.5 μM**	Binder	3	500
mFAP0	~0.5 μM**	Binder	3	500
mFAP1	0.56 μM	Binder	3	500

*Unsuccessful designs are subdivided into two categories: “Unstable” for designs that showed improper folding or aggregation and “Non-binder” for folded designs that did not show ligand affinity within the sensitivity of the binding assays[5,7].

**estimated K_D values based on rough titration data from Dou, J. *et al*[7]

Table 2.

Evaluation of the supervised learning classifiers using 5-fold cross validation[a].

Validation metric	Classification algorithm	
	k-nn (k = 5)	Logistic regression
Accuracy	0.84 ± 0.16	0.93 ± 0.10
Precision	0.79 ± 0.21	0.87 ± 0.17
Recall	1.0 ± 0.0	1.0 ± 0.0
MCC	0.74 ± 0.25	0.87 ± 0.16
F1S	0.87 ± 0.13	0.92 ± 0.10

[a]Values correspond to average and standard deviation of the 5 cross validations.

Table 3.

Evaluation of the generality of the classifiers, with models trained exclusively on the β-barrel designs[a].

Validation metric	Training + validation set: 33 β-barrel designs (70:30 split) Test set: 4 NTF2 designs			
	k-nn (k = 5)		Logistic regression	
	Validation set	Test set	Validation set	Test set
Accuracy	0.83 ± 0.09	1.0 ± 0.0	0.91 ± 0.08	1.0 ± 0.0
Precision	0.70 ± 0.15	1.0 ± 0.0	0.80 ± 0.19	1.0 ± 0.0
Recall	0.98 ± 0.05	0.95 ± 0.15	1.0 ± 0.0	1.0 ± 0.0
MCC	0.71 ± 0.13	0.96 ± 0.13	0.83 ± 0.16	1.0 ± 0.0
F1S	0.81 ± 0.09	1.0 ± 0.0	0.87 ± 0.13	1.0 ± 0.0

[a]Values correspond to average and standard deviation of 10 rounds of random splits of the data set according to the 70%:30% training:validation ratio.

Table 4.

Evaluation of the generality of the classifiers, with models trained exclusively on the DIG designs.

Validation metric	Training set: 4 DIG designs Validation set: 33 β-barrel designs	
	k-nn (k = 2)	Logistic regression
Accuracy	0.70	0.70
Precision	0.52	0.52
Recall	1.0	1.0
MCC	0.53	0.53
F1S	0.69	0.69