An enumerative algorithm for de novo design of proteins with diverse pocket structures

Benjamin Basanta[a,b,1], Matthew J. Bick[a,b,2], Asim K. Bera[a,b], Christoffer Norn[a,b], Cameron M. Chow[a,b], Lauren P. Carter[a,b], Inna Goreshnik[a,b], Frank Dimaio[a,b], and David Baker[a,b,c,3]

[a]Institute for Protein Design, University of Washington, Seattle, WA 98195; [b]Biochemistry Department, School of Medicine, University of Washington, Seattle, WA 98195; and [c]Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195

Edited by William F. DeGrado, University of California, San Francisco, CA, and approved July 28, 2020 (received for review March 26, 2020)

To create new enzymes and biosensors from scratch, precise control over the structure of small-molecule binding sites is of paramount importance, but systematically designing arbitrary protein pocket shapes and sizes remains an outstanding challenge. Using the NTF2-like structural superfamily as a model system, we developed an enumerative algorithm for creating a virtually unlimited number of de novo proteins supporting diverse pocket structures. The enumerative algorithm was tested and refined through feedback from two rounds of large-scale experimental testing, involving in total the assembly of synthetic genes encoding 7,896 designs and assessment of their stability on yeast cell surface, detailed biophysical characterization of 64 designs, and crystal structures of 5 designs. The refined algorithm generates proteins that remain folded at high temperatures and exhibit more pocket diversity than naturally occurring NTF2-like proteins. We expect this approach to transform the design of small-molecule sensors and enzymes by enabling the creation of binding and active site geometries much more optimal for specific design challenges than is accessible by repurposing the limited number of naturally occurring NTF2-like proteins.

protein design | high-throughput screening | NTF2-like proteins | protein pockets

Proteins from the NTF2-like structural superfamily consist of an elongated β-sheet that, along with three helices, forms a cone-shaped structure with a pocket (Fig. 1*A*). This simple architecture is highly adaptable, as evidenced by the low-sequence homology among its members, and the many different functions they carry out (1). Natural NTF2-like proteins have been repurposed for new functions through design (2–4), further showing the adaptability of this fold. General principles for designing proteins with curved β-sheets have been elucidated, and used to design several de novo NTF2-like proteins (5).

De novo design of protein function starts with an abstract description of an ideal functional site geometry (for example, a catalytic active site), and seeks to identify a protein backbone conformation with geometry capable of harboring this site. The extent to which the ideal site can be realized depends on the number and diversity of backbone conformations that can be searched (6, 7). A promise of de novo protein design is to generate a far larger and more diverse set of designable backbones for function than is available in the largest public protein structure database, the Protein Data Bank (PDB) (8, 9). This has been achieved for protein–protein binding due to the simplicity of small globular proteins (10). However, protein structures with pockets are considerably more complex, and since only a small number of de novo designed pocket-containing proteins have been characterized, this vision has not yet been realized for small-molecule binder or enzyme design. Here we develop a rule-based algorithm, akin to those used in generative design (11) that generates NTF2-like protein structures, exploring structure space by enumerating all possible combinations of high-level structural parameters that describe this fold. This algorithm samples the structural space available to the NTF2 fold systematically and widely, and the generated protein models surpass native NTF2-like proteins in pocket diversity.

Results

De novo protein design is a two-step process: First, a protein backbone conformation is generated, and second, low-energy amino acid sequences for this backbone are found by combinatorial side-chain packing calculations. In Rosetta (12, 13), new backbones can be constructed by Monte Carlo assembly of short peptide fragments based on a structure "blueprint," which describes the length of the secondary structure elements, strand pairings, and backbone torsion ranges for each residue (14, 15). Because this process is stochastic, each structure generated is distinct. We previously showed that NTF2-like proteins can be designed from scratch using this approach (5), but the diversity and number of designs to date (on the order of tens) is too limited to provide pockets for arbitrary function design. For a given blueprint, the resulting set of structures is generally more homogeneous than that observed in naturally occurring proteins within a protein family, where differences in secondary structure lengths and tertiary structure give rise to considerable diversity. Hence while large numbers of backbones can be generated for a particular blueprint, for example those previously used to design NTF2-like proteins, the overall structural diversity will be limited.

The NTF2 Enumerative Algorithm. To access a much broader range of protein backbones, we sought to develop an algorithm that samples a wider diversity of structures than natural NTF2-like

Significance

Reengineering naturally occurring proteins to have new functions has had considerable impact on industrial and biomedical applications, but is limited by the finite number of known proteins. A promise of de novo protein design is to generate a larger and more diverse set of protein structures than is currently available. This vision has not yet been realized for small-molecule binder or enzyme design due to the complexity of pocket-containing structures. Here we present an algorithm that systematically generates NTF2-like protein structures with diverse pocket geometries. The scaffold sets, the insights gained from detailed structural characterization, and the computational method for generating unlimited numbers of structures should contribute to a new generation of de novo small-molecule binding proteins and catalysts.

Author contributions: B.B. and D.B. designed research; B.B., M.J.B., A.K.B., C.N., C.M.C., L.P.C., and I.G. performed research; B.B., M.J.B., A.K.B., C.N., and F.D. analyzed data; and B.B. and D.B. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

This open access article is distributed under Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

[1]Present address: Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, CA 92037.

[2]Present address: Lyell Immunopharma, Seattle, WA 98109.

[3]To whom correspondence may be addressed. Email: dabaker@uw.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2005412117/-/DCSupplemental.

First published August 24, 2020.

Downloaded at UNIVERSITY OF WASHINGTON on December 7, 2020

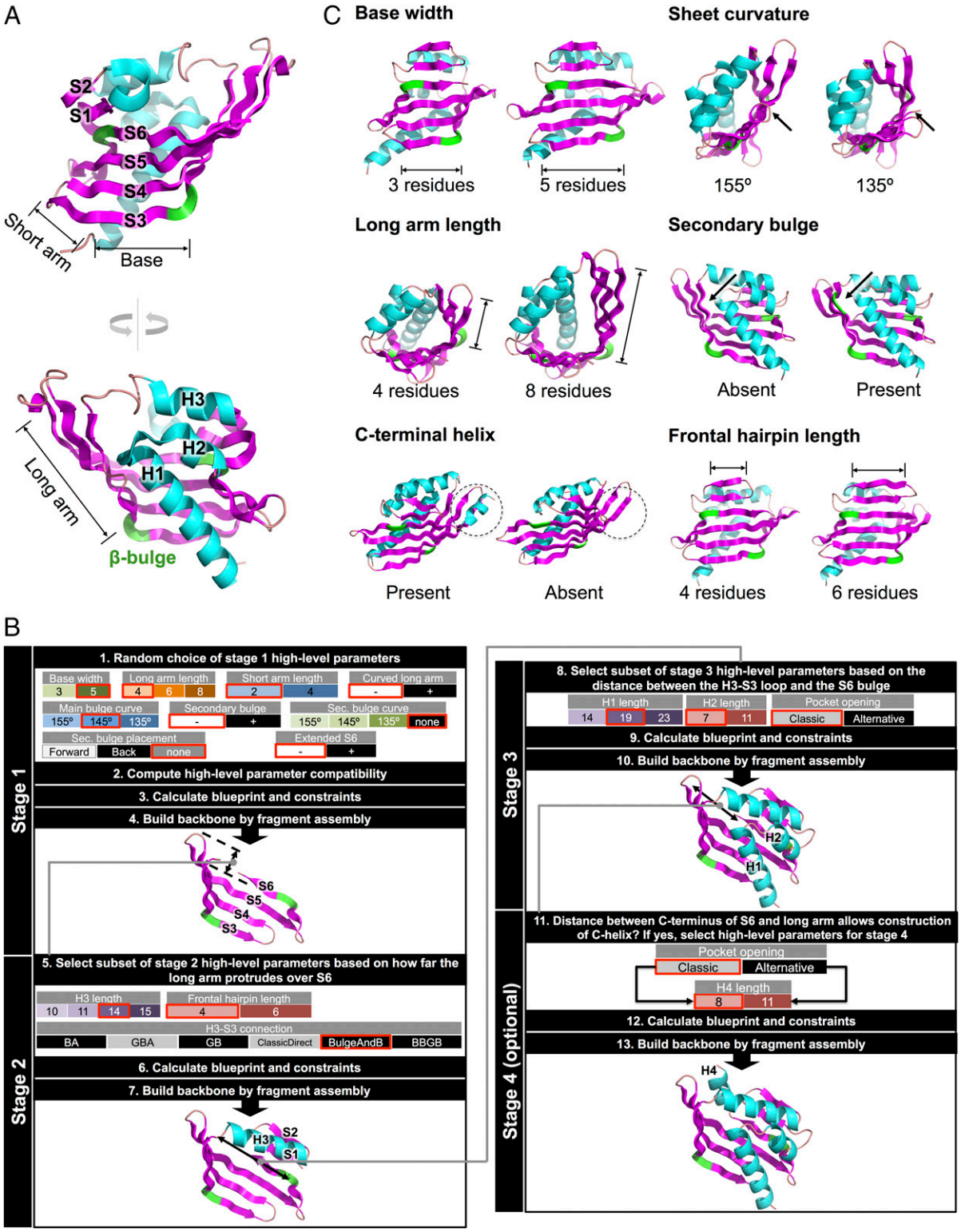

Fig. 1. High-level description of the NTF2 enumerative algorithm. (*A*) Canonical NTF2-like structural elements, labeled on the structure of scytalone dehydratase from *Magnaporthe grisea* (PDB ID 1IDP). (*B*) Overview of enumerative algorithm. At each stage of hierarchical backbone assembly, high-level parameters and local structure variation are sampled. (*C*) Examples of fold parameters sampled at the higher levels, and structures representing two extreme values for each.

proteins by carrying out backbone sampling at two levels (Fig. 1*B*). At the top level, sampling is carried out in the space of high-level parameters that define the overall properties of the NTF2 fold: For example, the overall sheet length and curvature, the lengths of the helices that pack on the sheet, the placement of the pocket opening, and the presence or absence of C-terminal elements (Fig. 1*C*). We then convert each choice of high-level parameters into structure blueprint/constraints pairs (hereon referred to

Downloaded at UNIVERSITY OF WASHINGTON on December 7, 2020

simply as blueprints), which guide backbone structure sampling at successive stages of fold assembly (see next paragraphs) (Fig. 1*B*). In total, there are 18 high-level fold parameters (*SI Appendix*, Table S1), and each unique combination gives rise to a specific blueprint. At the lower level, backbone structures are generated according to these blueprints through Monte Carlo fragment assembly; the blueprints dictate the secondary structure and torsion angle bins of the fragments, as well as a number of key residue–residue distances (*SI Appendix*, Figs. S1–S4). In a final sequence design step, for each generated backbone, low-energy sequences are identified through combinatorial sequence optimization using RosettaDesign.

We generate structure blueprints from the high-level parameters using a hierarchical approach (Fig. 1*B*). First, the four main strands of the sheet are constructed, then helix 3 and the frontal hairpin, finally, the two N-terminal helices. If the backbone to be assembled has a C-terminal helix, it is added in a fourth step.

In the first step, the length and curvature of the sheet are the primary high-level parameters sampled (Fig. 1 *C*, *Top* and *Middle*). For each choice of high-level sheet length and curvature parameters, compatible sets of low-level parameters (secondary structure strings and angle and distance constraints) are generated to guide Rosetta fragment assembly. The translation from sheet length to secondary structure length is straightforward as longer strands generate longer sheets. To realize a specified sheet curvature, bulges are placed at specific positions on the edge strands, where they promote sheet bending (5, 16, 17). Bulges are specified by a residue with α-helical ϕ/Ψ torsion values in the blueprint, leading to a backbone protuberance with two adjacent residues pointing in the same direction. As shown in Fig. 1*A*, there are always at least two bulges on the NTF2 sheet, delimiting the base and arms, and marking the axes at which the sheet bends. An additional bulge can exist on the long arm, further bending the sheet. To control the degree of bending centered at these points, angle constraints are placed on C_α carbons on center strands, at positions adjacent to bulges (*SI Appendix*, Fig. S1). Not all combinations of sheet length and curvature values are compatible with a closed pocket-containing structure: For example, long sheets with low curvature cannot generate a cone-shaped structure. These incompatibilities are identified by attempting to construct sheet structures (as described above) across the full parameter space, and then assessing the success in generating a pocket-containing structure. (Note: The region with no solutions at the bottom left of Fig. 3*A* reflects the incompatibility of long sheets and low curvature with the formation of a pocket; See *SI Appendix*, Table S2 for the complete set of rules dictating high-level parameter combinations).

The range of possibilities for helix 3 and the frontal hairpin, which are generated next, is limited by the geometric properties of the sheet constructed in the first step. In order to determine which parameter combinations lead to folded proteins, we generated and evaluated backbone structures based on a wide variety of parameter combinations, and extracted the following rules. Structures where the sheet does not protrude outwards beyond the pocket opening require longer loops between helix 3 and strand 3 (*SI Appendix*, Fig. S2*A*). Conversely, sheets that protrude outwards over the opening of the pocket require shorter loops between helix and strand 3, to avoid placing helix 3 too far from the rest of the structure (*SI Appendix*, Fig. S2*A*). The length of helix 3 is coupled to the torsional angles of the loop that connects it to strand 3, such that hydrogen bonds form between the backbone of the loop and the C terminus of helix 3 (*SI Appendix*, Fig. S2*B* and Table S3). Independent from helix 3 length and its connection to strand 3, the length of the frontal hairpin strands (two possible values: four or six residues) depends on the length of the sheet base: Narrow sheet bases allow only short hairpins, as all positions on strand 1 must be paired to strand 6 (*SI Appendix*, Fig. S2*C*).

Stage 3, the construction of the N-terminal helices, is likewise constrained by the geometric properties of the structure built so far. If the distance between the bulge on strand 6 and the loop connecting helix 3 with strand 3 is more than 25 Å, then helix 1 and 2 are elongated by a full turn (4 amino acids) to close the cone described by the sheet (*SI Appendix*, Fig. S3). The constraints that control the placement of H1 and H2 are adapted based on the shape of the current structure in order to position H1 and H2 such that good side-chain packing is favored during sequence design, and occluding backbone polar atoms on the outward-facing edge of S3 is avoided (*SI Appendix*, Fig. S3).

In cases where the backbone to be assembled has a C-terminal helix (has_cHelix = True), if the pocket opening is, like in most native NTF2-like proteins, between the frontal hairpin and H3 (Opening = Classic), the C-terminal helix is set to eight residues long and rests against the long arm. If the opening is set to be between the termini of H1 and H2, and H3 (Opening = Alternative), then the C-terminal helix length is set to 11 residues long, and closes the space between H3 and the frontal hairpin (Fig. 1*B* and *SI Appendix*, Fig. S4).

This four-step blueprint building procedure is implemented in a Python script that samples over the high-level degrees of freedom incorporating the logic described in the preceding paragraphs, and the improvements described throughout the remainder of the paper.

High-Throughput Characterization of the Known De Novo NTF2 Structure Space. The design of large pockets in de novo NTF2-like proteins is challenging and requires strategies to compensate for the loss of stabilizing core residues that would otherwise fill the space occupied by the pocket. Before setting out to experimentally sample the full range of structure space accessible to the enumerative algorithm, we chose to characterize the sequence and structure determinants of stability in the region of NTF2 space explored in our previous work (5), and its immediate vicinity. We generated 2,709 new NTF2-like proteins belonging to the blueprints previously described, plus a few variations (9 different blueprints) (*SI Appendix*, Fig. S5 and Table S4). We adapted a high-throughput stability screen based on folding-induced protease resistance on yeast cell surface, originally developed for small (<43 amino acid) domains (18) to the much larger (105 to 120 residues) NTF2-like protein family. This required optimizing current methods (19) for efficiently splicing long oligonucleotides (230 bases) from oligonucleotide arrays to form longer genes by limiting pairing promiscuity and, therefore, the number of chimeric design combinations (*Materials and Methods*).

A fifth (578, 21%) of the tested designs were stable (stability scores above 1), while only 2% of scrambled controls (randomly selected design sequences scrambled such that the hydrophobicity pattern is maintained) passed this stability threshold (Fig. 2*A*). All tested blueprints had representatives among the stable sequences (*SI Appendix*, Fig. S6). Analysis of the sequences and structures of the stable designs revealed several broad trends. There was a marked depletion of hydrophilic residues in positions oriented toward the protein core (*SI Appendix*, Fig. S7), suggesting that the stable proteins identified in this first round experiment are likely folded as modeled, but may not be able to accommodate a pocket with polar amino acids, limiting their potential to be designed for general function. A logistic regression model trained to distinguish between designs with stability scores above or below 1.0 identified total sequence hydrophobicity (see "hydrophobicity" feature definition in *SI Appendix*, *Supplementary Methods*), Rosetta energy ("score_res_betacart"), and local sequence-structure agreement (fragment quality, see "avAll") as key determinants of stability (*SI Appendix*, Fig. S8).

The importance of overall hydrophobicity is in agreement with the observed per-position amino acid enrichments, and suggests the composition or size of the designed protein cores is suboptimal. While Rosetta optimizes local sequence-structure agreement at single positions [p_aa_pp and rama_prepro energy function terms (20)], overall secondary structure propensity depends on stretches

Downloaded at UNIVERSITY OF WASHINGTON on December 7, 2020

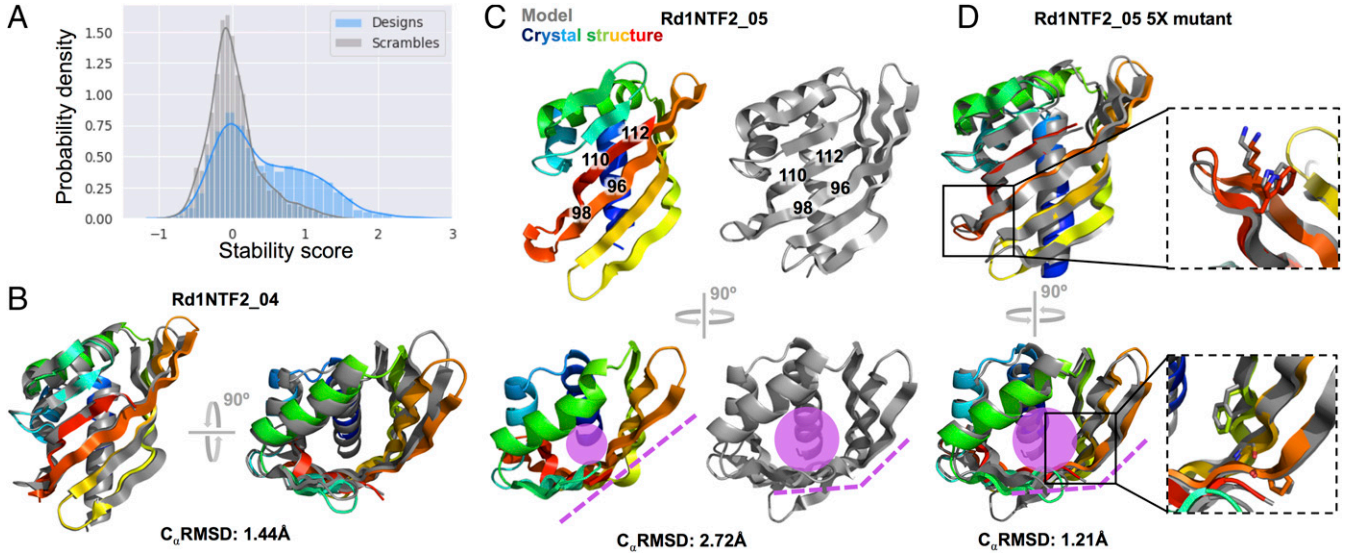

Fig. 2. High-throughput screening and structural characterization of de novo NTF2-like proteins. (*A*) Round 1 stability score distributions. Designs are more likely to have stability scores above 1.0 than scrambled sequences. (*B*) Crystal structure and computational model of design Rd1NTF2_04 (PDB ID code 6W3G); the protein backbone is in very close agreement. (*C*) Crystal structure and model of design Rd1NTF2_05 (PDB ID code 6W3D), showing significant differences between model and structure. Strands 5 and 6 are shifted two residues relative to each other (bold numbers, *Left*), resulting in a smaller space in the concave side of the flattened sheet (magenta sphere and dashed line, *Right*). (*D*) Crystal structure and model of design Rd1NTF2_05 fivefold mutant (PDB ID code 6W3F), showing agreement between model and structure for backbone and mutated side chains. As in *C*, a magenta circle and lines show how the concave side and sheet curvature fold as designed.

of several residues and cannot be decomposed in pairwise or single body energies. The detection of local sequence-structure agreement as a feature of stable designs suggests the design protocol produces sequences with suboptimal local sequence-structure relationship.

We selected 17 designs with a stability score above 1 for more thorough biophysical characterization (*SI Appendix*, *Supplementary Methods*). Seven of these expressed in soluble form in *Escherichia coli*, and were found to be folded by CD spectroscopy. Six of seven remained folded up to 95 °C, and had two-state unfolding transitions in guanidine hydrochloride denaturation experiments (*SI Appendix*, Figs. S9 and S10 and Table S5). The remaining 10 designs did not express or formed higher-order oligomers (*SI Appendix*, Table S5), indicating stability score values above those of most scrambles are no guarantee of soluble expression and folding in *E. coli* cytoplasm; aggregation is likely suppressed on yeast cell surface.

We obtained crystal structures for two of the above-mentioned hyperstable proteins with de novo NTF2 blueprints not characterized before (Fig. 2 *B* and *C* and *SI Appendix*, Fig. S11*A*). The crystal structure and model of design Rd1NTF2_04 are in close agreement both in terms of C_α atom positions and most core side-chain rotamers (Fig. 2*B* and *SI Appendix*, Fig. S11*B*). In contrast, the structure of design Rd1NTF2_05 shows a two-residue register shift between strands 5 and 6 relative to the design (Fig. 2*C*), which results in a flatter sheet and a smaller core, a shorter strand 5, and longer strand 6. While the overall shape of the structure and the relative orientations of the hydrophobic residues in strand 5 and 6 are preserved (Fig. 2*C*), the structure deviations would be significant for a designed functional pocket. The identification of a design that is stable but has a structure different from its model provides an opportunity to discover determinants of structural specificity not captured by the design method.

We hypothesized that the disagreement between model and structure for design Rd1NTF2_05 originates from a lack of core interactions favoring the modeled high sheet curvature around residue 94, as well as from lack of consideration of negative design in the sequence choice for the 5/6 strand hairpin, which allows the shortening of strand 5. We identified several mutations that could favor the modeled sheet curvature and strand register. Mutations D101K and L106W near the strand 5/6 connection make favorable interactions in the context of the designed conformation, and the replacement of leucine 106 with a large tryptophan side chain is incompatible with the observed crystal structure (*SI Appendix*, Fig. S12). Mutation A80G, at the most curved position of strand 4, favors bending by removing steric hindrance between the alanine 80 side chain and the backbone at position 66, but leaves a void in the core, which modeling suggests should be rescued by I64F (*SI Appendix*, Fig. S12) (see description in refs. 6 and 21). A phenylalanine side chain at position 64 makes favorable interactions in the designed conformation, and is likely to not fit in the core and be exposed in the observed conformation. Finally, the rigidity imparted by proline in position 94 limits the Ramachandran angles to those compatible with the designed conformation, as well as preventing strand 5 and 6 pairing beyond residue 92.

Experimental characterization of the Rd1NTF2_05 fivefold mutant showed a higher ΔG of unfolding than the original design (*SI Appendix*, Fig. S13), and its crystal structure is in close agreement with the model (Fig. 2*D*). The side chains at the five mutated positions were in the exact designed conformation, supporting our structural hypothesis and the incorporation of negative design to increase structural specificity (Fig. 2 *D*, *Right*). The fivefold mutant also displays a large cavity, present in the design, a unique example of a de novo-designed monomeric NTF2 with a large pocket that does not require additional stabilizing features, such as a disulfide bond or a dimer interface (*SI Appendix*, Fig. S14). We incorporated the principles used to improve the design Rd1NTF2_05 in the enumerative algorithm to increase the probability that the generated designs fold as modeled.

High-Throughput Characterization of New Regions of NTF2 Structure Space Explored by the Enumerative Algorithm. Armed with the insights from high-throughput characterization of known de novo NTF2 structural space, we set out to design proteins from hundreds of backbone blueprints created using our enumerative algorithm that explore a much larger structure space. We incorporated the

Downloaded at UNIVERSITY OF WASHINGTON on December 7, 2020

lessons learned in the sequence design stage, with the goal of generating more stable and diverse designs that fold as modeled. To address the low sequence hydrophobicity, we added an amino acid composition term to the Rosetta energy function to favor sequences with 30% nonalanine hydrophobic amino acids on average, with different hydrophobicity targets for core, interface, and surface positions. We increased the consistency of the predicted and design target secondary structure by increasing sampling (*SI Appendix*, Fig. S15). Finally, guided by the experience with design Rd1NTF2_05, we incorporated steps in the design process that detect strand curvature ranges that require glycine placement to reduce strain. We used this improved method to generate a second round of designs exploring a much larger set of 1,503 blueprints. These designs span a wide range of pocket volumes that are modulated by sheet length and curvature (Fig. 3*A*, *x* and *y* axes). There are two main modes by which the specification of sheet structure by the high-level parameters modulates pocket volume. First, as sheets of similar length become more curved, the helices come closer to the sheet, resulting in smaller pocket volumes (*SI Appendix*, Fig. S16). Second, as sheets with similar curvature elongate, they wrap around the concave face and extend the pocket outwards (*SI Appendix*, Fig. S16). The "sheet periodic table" in *SI Appendix*, Fig. S16*A* shows how the high-level parameter values can be arranged to track pocket volume. The rest of the high-level parameters have less impact on pocket volume (*SI Appendix*, Fig. S17).

Due to gene length limitations, we were able to test designs for 323 unique parameter combinations of the possible 1,503; these yield proteins of 120 amino acids or less in length. We synthesized genes for 5,188 proteins generated from these 323 blueprints, and subjected the designed proteins and scrambled versions to the protease stability screen. The increased protease resistance of the scrambled sequences likely reflects their increased hydrophobicity (*SI Appendix*, Fig. S18). Roughly one-third (29%) of the designs had stability values above those of most scrambled sequences (Fig. 3*B*) (98% of all scrambles have stability score <1.55), a larger fraction than the 21% of stable designs in the initial screen, increasing our dataset of stable NTF2-like designs from a total of 578 to 2,077. These stable designs belong to 236 different parameter combinations, a very large increase over the 9 combinations in the previous round, with most of the missing combinations having fewer than 10 initial samples (*SI Appendix*, Fig. S19). The new parameter combinations result in structural features not sampled before, such as a secondary bulge on the long arm, new H3–S3 connections, and elongated frontal hairpins. The pocket volume distribution of stable designs is very similar to the distribution for all tested designs (Fig. 3*C*), suggesting that pocket volume is not a limiting factor, and spans most of the native NTF2 range (*SI Appendix*, Fig. S20). Stable designs not only sample the native pocket volume range, but also the native range for several other pocket properties (*SI Appendix*, Fig. S21). The amino acid identities in stable designs show much lower levels of enrichment and depletion at individual positions than in the first round of high-throughput screening (*SI Appendix*, Fig. S22); in particular, polar amino acids are not depleted in core positions (*SI Appendix*, Fig. S22), suggesting that polar residues are likely better tolerated in pocket positions perhaps due to the improved core packing resulting from the optimized sequence design protocol.

With the large increase in diversity in the second round, the stable designs created by the enumerative algorithm span a very wide range of structures. To visualize the space spanned by our generated structures compared to native NTF2 structures, we used the uniform manifold approximation and projection (UMAP) algorithm (22) to project similarity in backbone structure [TM-score (23)] into two dimensions (see Fig. 4 and *SI Appendix*, Fig. S23 for plots generated using different UMAP hyperparameters). The grouping of structures with similar features in different map regions provides an indication of which model parameters lead to novel NTF2 structures (*SI Appendix*, Fig. S24). Inspection of the map shows that our algorithm samples most of the native space, as well as completely uncharted regions. Most native proteins form clusters that overlap with de novo ones, likely reflecting overall structural similarity between these, with differences that can be attributed to loop structure: Native NTF2-like proteins often have long, heterogeneous loops, while our designs tend to have short, homogeneous loops. The subset of designs tested by high-throughput screening sample a wide range of structures within the accessible protein length, and stable representatives from the 236 unique NTF2 parameter combinations are found across the sampled space (Figs. 3*A* and 4). Overall, the number and diversity of de novo-designed NTF2-like structures is considerably larger than that of the NTF2 structures in the PDB.

A logistic regression model trained on stability of second-round designs suggests the lessons from the first round of high-throughput screening proved effective, and provides new suggestions for improvement (*SI Appendix*, *Supplementary Information Text* and Fig. S25). Features based on the high-level parameters of the enumerative algorithm (e.g., H3 length, sheet curvatures, sheet length, and hairpin length) did not contribute significantly to stability prediction, suggesting stable proteins can be designed across all of the considered structural space (*SI Appendix*, *Supplementary Information Text*).

We biochemically characterized 37 stable designs from the second round of high-throughput screening; 43%, similar to the 41% in round 1, expressed solubly in *E. coli* and had CD spectra consistent with the folded state (the remaining 20 second-round designs did not express or formed higher-order oligomers) (*SI Appendix*, Table S6). Most of the folded designs retained their folded state CD spectrum above 95 °C (*SI Appendix*, Figs. S26 and S27 and Table S6). The length of helix 3 in two of the second-round stable designs, Rd2NTF2_06 and Rd2NTF2_19, is the longest of the values we allowed, supporting the designability of this feature despite it being slightly disfavored by the stability model (*SI Appendix*, *Supplementary Information Text* and Fig. S25). Overall, the folded designs sample a wide range of pocket shapes and sizes (*SI Appendix*, Fig. S21).

More than half of the designs we attempted to express in *E. coli* did not express or formed soluble aggregates, indicating that a high stability score does not necessarily translate to folding in *E. coli* cytoplasm. While stability score has no significant correlation with $\Delta G_{unfolding}$ for these larger proteins, it has some capacity to discriminate designs that fold from those that do not (*SI Appendix*, Fig. S28*C*). Nine out of nine designs with low stability scores that we attempted to express in *E. coli* did not fold, supporting the use of stability score as a metric to improve the design of these pocket-containing proteins. In an attempt to improve the power of the stability score to predict folding and stability of proteins expressed in *E. coli*, we trained an alternative unfolded state protease-resistance model based on the protease resistance of scrambled sequences (*SI Appendix*, *Supplementary Methods* and Fig. S28 *D* and *E*). As expected, this model predicts NTF2 scrambled sequence stability better than the published unfolded-state model, but using it to recalculate stability scores did not lead to better prediction of $\Delta G_{unfolding}$ or folding in *E. coli* (*SI Appendix*, Fig. S28 *B* and *C*).

For two of the folded hyperstable designs (Rd2NTF2_20 and Rd2NTF2_16), we obtained high-resolution crystal structures. Overall, the crystal structures are very close to the computational design models (Fig. 3 *D*–*G*). Both designs feature structural elements that were not present in the known de novo NTF2 structural space. Rd2NTF2_20 has an extended connection between H3 and S3, recapitulated in the crystal structure (Fig. 3*D*), which enables the use of a short helix 3. Rd2NTF2_16 features two new structural elements, a bulge on the long arm (in addition to the ones flanking the base), and an extended frontal hairpin, both recapitulated in the crystal structure (Fig. 3*F*). The additional bulge enables higher curvature on the long arm, contributing significant diversity to long-arm structure, which is further increased by allowing different bulge

Downloaded at UNIVERSITY OF WASHINGTON on December 7, 2020

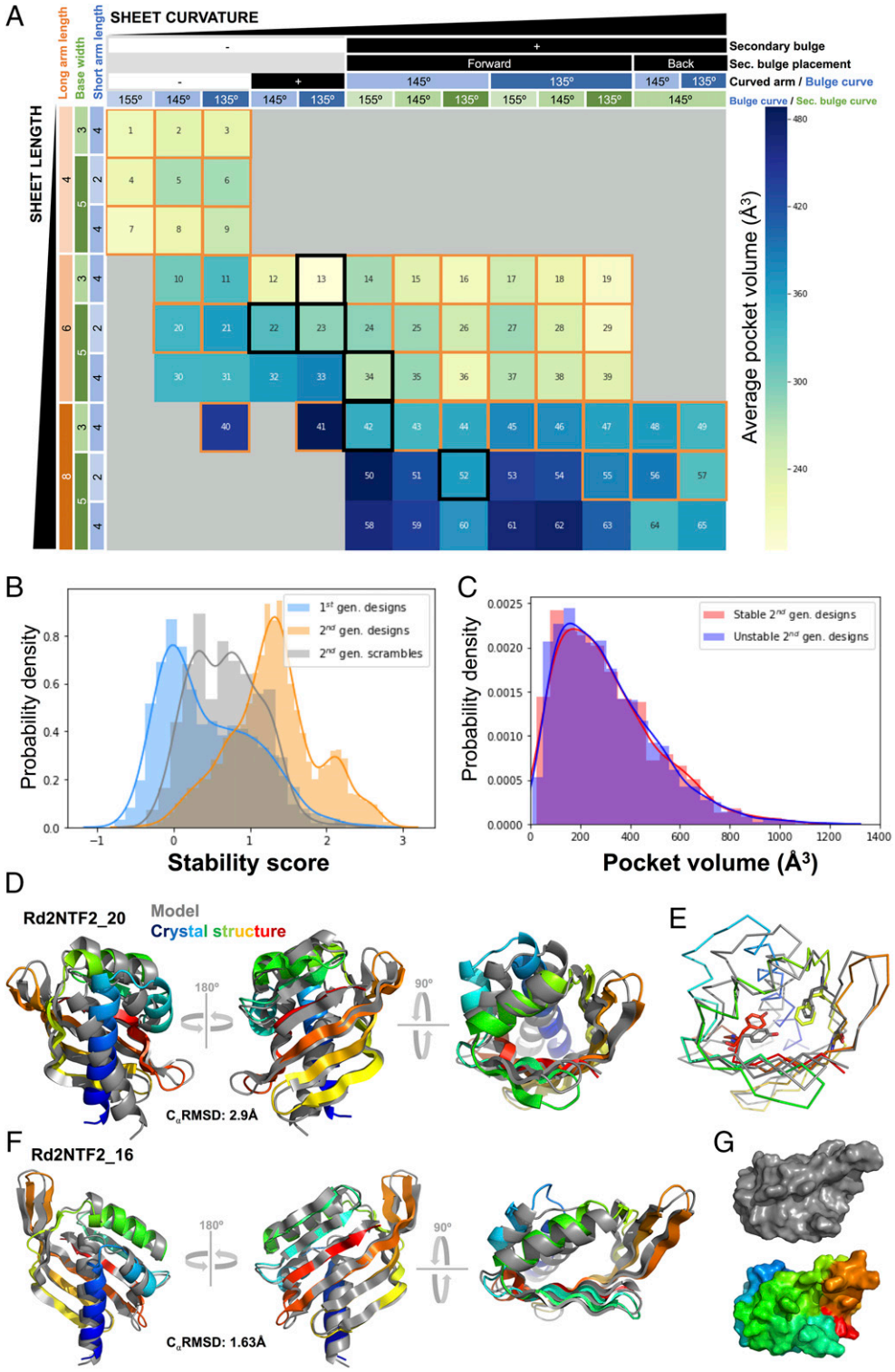

Fig. 3. Characterization of second-round designs. (*A*) De novo NTF2 designs sorted by sheet structure and ordered by sheet curvature and length. Each quadrant is colored by the average pocket volume of designs belonging to it. Orange frames denote quadrants for which stable designs were identified. Black frames denote designs were tested, but no stable design was identified. (*B*) Stability score of algorithm designs (orange), compared to controls (gray) and designs from the initial screening (blue). (*C*) Volume distribution of stable and unstable designs. (*D*) Crystal structure of stable design Rd2NTF2_20 (PDB ID code 6W3W), which features a new, elongated helix three-strand connection. Despite significant differences between the model and structure in the N-terminal helices, the new loop and the sheet are well recapitulated. (*E*) Core rotamers of Rd2NTF2_20. TYR101 (red, sticks) shows a significant deviation from the model, and enables the change in location of helix 1. In contrast, PHE61 and GLY77 interact as modeled, showing the glycine rescue feature can be designed from scratch. (*F*) Crystal structure of stable design Rd2NTF2_16 (PDB ID 6W40), which has a secondary bulge and an elongated frontal hairpin, features not designed before. Both of these features are recapitulated in the crystal structure. As in Rd2NTF2_20, but not as dramatic, the Rd2NTF2_16 crystal structure presents significant deviations from the model in the N-terminal helices. (*G*) Surface rendering of the model and crystal structure of Rd2NTF2_16, showing the shallow pocket formed by the long arm and the frontal hairpin is recapitulated by the crystal structure.

Downloaded at UNIVERSITY OF WASHINGTON on December 7, 2020

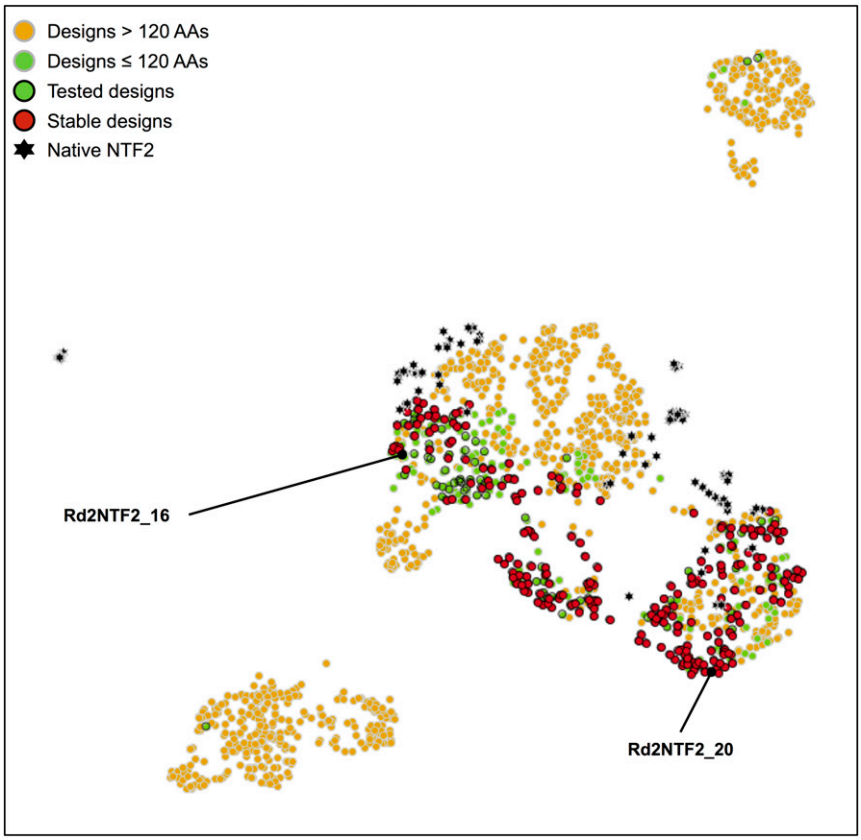

Fig. 4. Global comparison of de novo designed and native NTF2 structures. TMscores were computed for all pairs of structures, and the resulting distance map was projected into two dimensions using UMAP. For the de novo designs, each parameter combination is represented by a single structure randomly selected from the ensemble generated for that combination. Structurally similar proteins are closer together; the designs span a larger range of structural variation than the native structures.

placements. The extended hairpin, which is only designable when the base is sufficiently long, extends the pocket outwards, thereby increasing its volume. In the case of Rd2NTF2_16, the combination of these features yields a protein with a shallow groove instead of a pocket (Fig. 3*G*). The ability to generate proteins with shallow grooves with two open ends should enable design of binding sites for polymers, such as peptides or polysaccharides. The properties of the pockets in the crystal structures obtained in this work (*SI Appendix*, Table S7) span a broad range (*SI Appendix*, Fig. S22), confirming the ability of the enumerative algorithm to generate a diversity of pocket geometries.

The accuracy of the Rd2NTF2_20 and Rd2NTF2_16 computational models follows directly from the insights gained in the first large-scale design round. Both proteins feature a glycine on strand 4, enabling high curvature between the base and the long arm, as described for the design Rd1NTF2_05 fivefold mutant, and consequently incorporated in the enumerative algorithm. In order to implement the glycine placement on strand 4 as generally as possible, the design protocol searches for large hydrophobic side chains to fill the void left by the glycine. In Rd2NTF2_20, this is achieved by a phenylalanine in the same conformation as the one observed in the design 0589 fivefold mutant, while in Rd2NTF2_16 a void is left in the core. Unlike design Rd1NTF2_05, in the Rd2NTF2_20 and Rd2NTF2_16 crystal structures the highly curved sheet conformation is in close agreement with the model. In addition to generally supporting the models created by the enumerative algorithm, the two crystal structures provide information to improve the design method (*SI Appendix*, *Supplementary Information Text* and Fig. S29). The ability to design and properly model the sheet in de novo NTF2-like proteins is of great importance, as this structural element is the most involved in pocket structure.

Most of the 1,503 possible high-level parameter combinations yield proteins that are too long to be encoded by assembling two 240-base pair oligonucleotides (the limit of what can be synthesized at very large scale). To explore the parameter space that generates these longer proteins, we characterized 10 designs that are predicted to be stable by a logistic regression model trained on the second high-throughput screening experiment data, and have large pockets (500 to 1,200 Å^3). Two of the 10 were monomeric and remained folded above 95 °C, a success rate similar to that of the biochemical characterization of designs identified in the second high-throughput experiment, suggesting that de novo NTF2-like proteins longer than 120 amino acids with large pockets are also designable using the enumerative algorithm (*SI Appendix*, Figs. S30 and S31 and Table S8).

The goal of widely sampling NTF2 structural space is to produce structurally diverse pockets that can, in turn, harbor diverse binding and active sites. Most effective methods to design such sites do not rely on finding a preformed pocket with side chains of the correct identity, in perfect arrangement and configuration. Instead, they evaluate the ability of the protein backbone to harbor a binding site—represented as a precalculated constellation of side chains—for the small-molecule ligand or substrate of interest (6, 7). Because of this focus on backbone structure rather than complete atomic structure, we choose to analyze the utility of the NTF2-like proteins we generate in terms of the diversity of the backbone positions lining the inside of the cone

Downloaded at UNIVERSITY OF WASHINGTON on December 7, 2020

formed by the backbone. The enumerative algorithm outputs, alongside each model generated, a list of positions near the opening of the cone, which exclude loops, and whose C_α-C_β vectors point toward the concave side of the sheet (*Materials and Methods*). We can compare the geometric diversity of these positions to the positions lining the pockets of native proteins, as detected by CLIPPERS (24) (*SI Appendix*, *Supplementary Methods*). *SI Appendix*, Fig. S32 shows the distributions of three parameters (angles α, β, and distance *D*; defined in *SI Appendix*, *Supplementary Methods* and Fig. S32*A*), which describe the geometry of pocket positions around the pocket center of mass. Aside from differences that can be attributed to the heuristics used for choosing pocket positions (see further analysis in *SI Appendix*), the set of de novo proteins sample geometry space more thoroughly than native NTF2-like structures (*SI Appendix*, Fig. S32 *B*–*F*).

To further compare pocket geometries of our de novo-generated structures to those of native NTF2 proteins, we binned the C_α-C_β vectors lining each pocket based on their coordinates, and used UMAP (22) to project these features into two dimensions for visualization. Maps without and with loop positions are shown in *SI Appendix*, Fig. S33 *A* and *B*, respectively; in the former, the de novo designed and native distributions overlap (we do not match the loop variation in native structures in our designs); the structure and sequence of loops can be crucial for de novo-designed protein stability and folding (this work and refs. 15, 18, 25, and 26) and controlling complex loop structures, either by designing them from scratch or by mutation, can be challenging (27, 28). See *SI Appendix*, Fig. S33 *C* and *D* for UMAP hyperparameter exploration. The distribution of C_α-C_β geometries in the native and de novo proteins sets, and grouping of structures by these features, show that our de novo models sample geometry space more thoroughly than native NTF2-like proteins, suggesting scaffolds better suited for de novo design of binding and active sites may be found among our de novo models.

Suitability of Designed Scaffolds for Harboring Small-Molecule Binding Sites. To probe the capability of the designed proteins to host binding sites, we chose a set of 50 ligands from protein–small-molecule complexes in the PDB, docked them into designed and native NTF2 scaffolds using RIFDOCK (6), and designed the surrounding amino acids to make favorable interactions with the docked ligand (*Materials and Methods* and *SI Appendix*, Figs. S34 and S35). In these calculations, we used all second-round, stable de novo models, and native NTF2 structures with pockets larger than 30 $Å^3$ (790 and 64, respectively). By limiting the docking and design calculations to these 790 de novo NTF2 models, we are conservatively assessing only the diversity associated with the de novo proteins experimentally demonstrated to be stable. Because RIFDOCK and other binding site design algorithms only take into account the scaffold backbone, the existence of a pocket in the model is not a requirement, but for this in silico experiment we limit the designable positions to those lining the pocket designed by Rosetta (selected by CLIPPERS using the full-atom model), to provide a conservative estimate of the quality of binding sites that can be designed in these scaffolds (less conservative approaches to binding site design are described in *Discussion*).

After docking and design, we investigated whether the pocket with the most favorable interactions was based on a de novo or native NTF2-like protein for each ligand. The predicted binding energy scales with the size of the ligand; for a ligand size-independent measure, we calculated the mean and standard deviation (SD) of protein interaction energy for all designs for each ligand, and computed *z*-scores for each individual design [z = (interaction energy − mean)/SD)]. As a summary statistic for comparing the docks in native and designed proteins, we used the difference in *z*-scores of the lowest energy (most favorable) de novo designed protein dock and the lowest energy native protein dock. Larger positive Δ*z*-scores indicate a larger advantage of the best de novo scaffold over the best native scaffold. Despite the conservative choice of de novo designs, the de novo proteins provide a better (lower ligand interaction energy) pocket for 76% of all tested ligands (38 of 50), without obvious biases in ligand molecular weight, charge, chemical groups, or hydrophobicity (Fig. 5 and *SI Appendix*, Figs. S36 and S37). As controls for this docking test, we included two small molecules in the ligand set that are bound by the native NTF2 scaffolds (PDB ligand codes EQU and AKV, bound by 1OH0 and 2F99, respectively) and found that native-like poses are recovered when the bound ligand conformer found in the crystal structure is used (*SI Appendix*, Fig. S38). The de novo scaffold with the largest number of top ranking docks is Rd2NTF2_03, one of the designs found to be folded and highly stable (*SI Appendix*, Fig. S39). The observed advantage of de novo structures in binding site scaffolding should increase with the number of de novo designed structures generated, while the rate of growth of the native set is limited to what has been sampled by evolution.

As the overarching goal of this work is to expand the set of available protein structures with pockets, we generated a final set of scaffolds that incorporates all of the lessons from previous experiments. Improvements in the enumerative algorithm, both in sequence design and backbone generation resulted in increased diversity (1,619 unique parameter combinations) and improved stability-related metrics (see *SI Appendix*, *Supplementary Methods* and Figs. S33, S34, and S40 for pocket diversity). We have made this set of 32,380 scaffolds (20 models with different sequences per parameter combination) available for general use as starting points for ligand binding and enzyme design.

Discussion

Our enumerative algorithm may be viewed as encoding the "platonic ideal" of the NTF2-like structural superfamily along with a method for essentially unlimited sampling structures belonging to it, in a fashion directly tied to pocket structure. In terms of Structural Classification of Proteins–Extended (SCOPe) categories (29), each combination of top-level parameters can be thought of as a protein family, and the set of all combinations, the de novo NTF2-like structural superfamily. Whereas in our previous work four NTF2 structure blueprints were manually constructed, the new enumerative algorithm samples through over 1,600 unique blueprints that result in well-formed backbones. This represents a qualitative jump in the structural diversity that can be achieved for complex folds by de novo protein design.

Our experience in developing the enumerative algorithm for NTF2-like structures suggests guidelines for developing similar enumerative algorithms for other folds. First, determine the common structural elements that are part of all proteins in the target family: each stage of our algorithm builds one of such elements for the NTF2 fold. Second, identify a subset of elements that together form a central hub: In NTF2-like proteins, the hub is the curved sheet, and all stages use it as a reference point. Third, analyze how the properties of the elements covary due to larger structural constraints: For example in de novo NTF2-like proteins, at stage 2, the length of strands 1 and 2 is limited by the width of the sheet base, and the length of helix 4 and its connection to the sheet are dictated by the shape of the sheet. Fourth, simplify and adapt structural elements and their connections to rules that are well understood: We base the sheet construction on previously described principles, limit the length and torsions of many structural elements to a few easy-to-pair options, and make use of the knob-socket packing description (30) to arrange structural elements relative to each other (*SI Appendix*, Fig. S3). Even with these simplifications, the combination of relatively simple elements leads to a high level of structural diversity.

The generative approaches to de novo protein structure design so far described in the literature, rule- or model-based, either focus exclusively on helical structures (31–33), are not geared

Downloaded at UNIVERSITY OF WASHINGTON on December 7, 2020

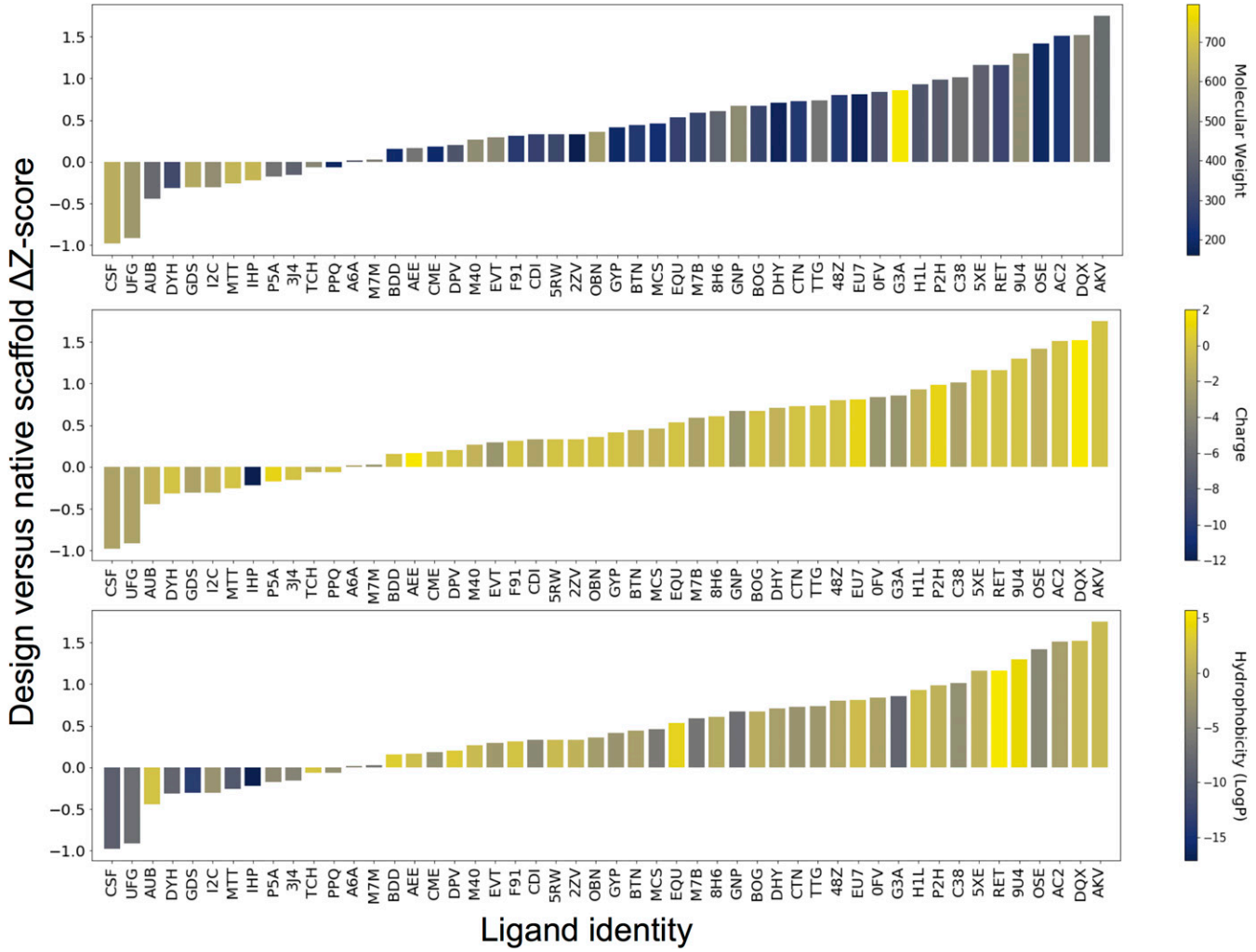

Fig. 5. Comparison of de novo designs to native structures for ligand docking and design. Following docking and design into our de novo designed and native protein scaffolds, ligand binding energies were computed and converted to *z*-scores. The *y* axis is the difference between the *z*-scores obtained for the best designed and best native scaffolds; higher values indicate that the best design had a more favorable binding energy and hence was a better scaffold for the ligand. Ligands are arranged along the *x* axis in order of Δ*z*-score. In each panel bars are colored by ligand properties, from top to bottom: molecular weight (Da), charge at pH 7.5, and hydrophobicity (LogP).

toward atomic-detail modeling and design (34), or sacrifice fine-grained structural control for structural diversity (35). Machine-learning–based generative models show considerable promise (35, 36), but have not yet been applied to the direct generation of full atomic structures with specific features of interest, as we do here for scaffolds containing a varied geometry of binding pockets. We hope the experimental data generated in this work will aid the development of models that more efficiently produce protein structures with finer control over atomic detail and greater diversity.

We provide several sets of de novo NTF2 models, and an algorithm to generate an unlimited number of them, to help the community address the challenge of finding an ideal scaffold in which to design a binding or active site. Because our algorithm can sample NTF2 space at different structural resolutions, we propose a hierarchical strategy to find the best-fitting scaffold for binding a specific ligand: First, use RIFDOCK to quickly dock and design binding sites on a set of scaffolds that sample a wide range of high-level parameters, and select a subset of parameter combinations that fit the ligand most favorably. Then, use the enumerative algorithm to create more models with these high-level parameter combinations, sampling the selected subspace more deeply, and dock and design with more exhaustive RIFDOCK settings. We should note that, as indicated by the differences between the experimental structures we obtained and their computational models, after a binding or active site is designed in a de novo NTF2 model, protein structure refinement and/or evaluation by independent measures, such as molecular dynamics simulations, is advisable to increase the likelihood that the desired active or binding site is recapitulated in the protein structure (37). Up to now, protein design for a specific function has relied either on searching through the scaffolds in the PDB, or generating small variations of a limited set of de novo scaffolds. Our approach now enables going far beyond both approaches by searching through an essentially unlimited set of generated scaffolds.

The experimental characterization of many of our designs shows that the enumerative algorithm samples a wide range of feasible structure space, and that designs usually fold as modeled. The insights we gained in learning to produce these diverse proteins can be harnessed to improve the success rate in future protein design efforts. Furthermore, our approach could be implemented for other protein folds to expand structural diversity even further. In combination with existing docking and design methods, the enumerative algorithm here presented should open the door to

Downloaded at UNIVERSITY OF WASHINGTON on December 7, 2020

design of novel functions by eliminating the limitations imposed by current protein structural databases, and enabling scaffold generation custom-tailored to function.

Materials and Methods

Enumerative Algorithm for Proteins from the NTF2-like Superfamily. All code can be downloaded from GitHub (https://github.com/basantab/NTF2Gen).

The NTF2Gen repository contains all of the tools for de novo design of NTF2-like proteins. The main script is CreateBeNTF2_backbone.py, which manages the construction of NTF2 backbones, followed by DesignBeNTF2.py, which designs sequence on a given backbone generated by the previous script. To generate backbones from a specific set of parameters, use CreateBeNTF2PDBFromDict.py. The fundamental building blocks of the backbone generation protocol are Rosetta XML protocols (included in the repository) that are specialized instances of the BlueprintBDRMover Rosetta fragment assembly mover. All checks and filters mentioned in *Results* previous to design are implemented either in the XML files or the Python scripts. Additional backbone quality controls are run after each step (*SI Appendix, Supplementary Methods*). The design script is also based on a set of XML protocols, one for each of three stages. The glycine placement in highly curved strand positions and the selection of pocket positions are managed by DesignBeNTF2.py (see the BeNTF2seq/Nonbinding directory). Pocket positions are selected by placing a virtual atom in the midpoint between the H3–S3 connection and the S6 bulge, and choosing all positions whose C_α-C_β vector is pointing toward the virtual atom (the V_{atom}-C_α-C_β angle is smaller than 90°), excluding positions in loops, and their C_α is closer than 8 Å, this information is stored in the each model PDB file under the PDB-Info labels, with the tag "Pckt".

De Novo NTF2 Backbone Generation and Sequence Design for the First Round of High-Throughput Screening. Backbones were constructed as described in Marcos et al. (5). For families not described in said paper (i.e., BBM2nHm* designs), the same backbone construction algorithms were used, but parameters were changed accordingly. Scripts for producing all these backbones can be found at https://github.com/basantab/NTF2Analysis, NewSubfamiliesGeneration. The sequence design protocol for the first round of designs can be found in the above-mentioned GitHub repository. Briefly, the design protocol begins by generating four different possible sequences using the Rosetta FastDesign mover in core, interface, and surface layers separately. Then, random mutations are tested, accepting only those that improve secondary structure prediction without worsening score, introducing Ramachandran outliers, or worsening the shape complementarity between helices and the rest of the protein.

Design of Gene Fragments for Multiplex Gene Assembly. In order to obtain full-length genes from fragments synthetized in DNA microarrays, they must be assembled from halves, as described in Klein et al. (19). To generate highly orthogonal overlaps, we generated DNA sequences using DNAWorks (38), then split the gene in half and altered the composition of around 20 overlapping nucleotides to have as low homology as possible with other halves in the pool, while maintaining an adequate melting temperature, GC content, and staying below the maximum oligonucleotide length (230 nucleotides). This optimized version of the algorithm described in Klein et al. (19) can be found at https://github.com/basantab/OligoOverlapOpt.

Protease-Based High-Throughput Stability Screening. The protease-based high-throughput stability screening was carried out as described in Rocklin et al. (18). Briefly, genes encoding for thousands of different de novo NTF2 sequences cloned in the pETCON2 vector, which has the protein of interest expressed as a chimera of the extracellular wall yeast protein AgaII, on its C terminus, connected by a "GS" linker of alternating glycine and serine. The protein of interest is followed by a myc-tag (EQKLISEEDL). This library is transformed in yeast for surface display in a one-pot fashion using electroporation. Different aliquots of the yeast culture are then subject to increasing concentrations of trypsin and chymotrypsin, and labeled with an anti-myc tag antibody conjugated to fluorescein. Cells still displaying full proteins (myc-tag–labeled) after this treatment are then isolated by FACS. Deep-sequencing of the sorted populations reveals which sequences are protease-resistant and to what degree, providing an estimate for folding free energy. The metric reported by this assay is the stability score, an estimate of how much protease is necessary to degrade a protein over that expected if the protein was completely unfolded. A stability score of 0 indicates that the protein is degraded by the same amount of protease as expected if it was unfolded, (i.e., it is likely completely unfolded). A stability score of 1 indicates that 10 times more protease is required to degrade the protein than expected if it was completely unfolded.

LASSO Logistic Regression Model Training on Stability Data. To identify features that predict stability, we trained LASSO (Least Absolute Shrinkage and Selection Operator) logistic regression models (39) using the features described in *SI Appendix*, Tables S9–S16, after normalization. A logistic regression model predicts the probability of a binary outcome using a logistic function that depends on a weighted summation of features. By sampling a series of L1 regularization values, we obtained models with varying degrees of parsimony, and for each of those L1 values we also generated different random partitions of our dataset. This way, for each L1 value we obtained models with a spread on accuracy, which we used for selecting an L1 regularization value that maximizes accuracy and minimizes complexity (i.e., the number of features with weight different from 0). The simplest measure of the importance of each feature is the magnitude of the assigned coefficient.

The data and code for this analysis derived from the first high-throughput experiment can be found at https://github.com/basantab/NTF2analysis, ProteaseAnalysisExp1/LassoLogisticRegression.ipynb. Analysis of data from the second high-throughput experiment can be found at: https://github.com/basantab/NTF2analysis, ProteaseAnalysisExp2/LassoLogisticRegression_new_version.ipynb.

Crystallography Data Collection and Analysis Metrics. To prepare protein samples for X-ray crystallography, the buffer of choice was 25 mM Tris, 50 mM NaCl, pH 8.0. Proteins were expressed from pET29b$^+$ constructs to cleave the 6xHis tag with tobacco etch virus (TeV) protease. Proteins were incubated with TeV protease (1:100 dilution) overnight at room temperature and cleaved samples were loaded to a Ni-NTA column preequilibrated in 25 mM Tris, 50 mM NaCl, pH 8.0+30 mM Imidazole. Flow-through was collected and washed with one to two column volumes. Proteins were further purified by FPLC size-exclusion chromatography using a Superdex 75 10/300 GL (GE Healthcare) column, and specific cleavage of the 6xHis tag was verified by SDS/PAGE.

Purified proteins were concentrated to ∼10 to 20 mg/mL for screening crystallization conditions. Commercially available crystallization screens were tested in 96-well sitting or hanging drops with different protein:precipitant ratios (1:1, 1:2, and 2:1) using a mosquito robot. When possible, initial crystal hits were grown in larger 24-well hanging drops. Obtained crystals were flash-frozen in liquid nitrogen. X-ray diffraction data sets were collected at the Advanced Light Source. Crystal structures were solved by molecular replacement with Phaser (40) using the design models as the initial search models. The structures were built and refined using Phenix (41, 42) and Coot (43). Crystallization conditions and data collection and refinement statistics can be found in the *SI Appendix, Supplementary Methods* and Table S17.

UMAP Embedding of NTF2 Designs. UMAP (22) is a dimension reduction technique widely used for visualization of high-dimensional data. We obtained the code for running UMAP by following instructions in https://umap-learn.readthedocs.io/en/latest/. For generating the embedding, UMAP requires a distance measure between points, for which we provided 1-TMscore between all analyzed structures. We ran UMAP in a Jupyter notebook with different metaparameter combinations and verified that the general cluster structure was conserved among all of them, and that structural features were reflected in the groupings. The code and files necessary for generating the UMAP-related figures can be found in the GitHub repository https://github.com/basantab/NTF2analysis, UMAP_embedding and Pocket_position_vector_analysis.

Ligand In Silico Docking Test. The goal of the ligand in silico docking test is to provide an estimate of how de novo NTF2-like proteins compare to native ones in terms of their ability to harbor arbitrary binding sites. We used RIFDOCK (6) for simultaneous docking and design based on a set de novo and native protein backbones. As RIFDOCK only uses backbone coordinates and a list of pocket positions to dock the ligand and design a binding site around it, it can be used in a sequence-agnostic way. We selected and prepared (see ligand preparation in *SI Appendix, Supplementary Methods*) a subset of 50 ligands from all nonpolymeric PDB ligands (Ligand Expo, ligand-expo.rcsb.org) using *k*-means clustering on physical and chemical features (see *SI Appendix*, Fig. S30, and the 50_ligand_table.html file at https://github.com/basantab/NTF2analysis/tree/master/ligandInSilicoDockingTest). The number of ligands tested was limited to 50 for computational tractability, as RIFDOCK uses a significant amount of resources per ligand and scaffold: >3 h in 32 cores and 64 GB of RAM on average per ligand, to generate the initial rotamer interaction field (RIF), and ∼2 h in 32 cores using >20 GB of RAM, per ligand for docking in a subset of 12 scaffolds. As

Downloaded at UNIVERSITY OF WASHINGTON on December 7, 2020

NTF2-like native representatives, we selected 64 structures with pockets (pockets detected and defined as described in the *SI Appendix, Supplementary Methods*) from the SCOPe2.05 database (described in *SI Appendix, Supplementary Methods*). In order to provide a conservative estimate of pocket diversity and aid computational tractability, we limited the set of de novo designs used for docking to those stable (stability score > 1.55) and with detectable pockets in the concave side of the sheet (>25% overlap between CLIPPERS-detected pocket and backbone-based pocket positions, and >30 Å^3 volume), resulting in 790 different de novo sequences (see https://github.com/basantab/NTF2analysis "ligandInSilicoDockingTest" for relevant files). Pocket residues were detected using CLIPPERS, as described in *SI Appendix*, *Supplementary Methods*, and only positions lining the pocket of the scaffolds this way, including loops, were considered for binding site design by RIFDOCK. We generated five binding site designs per scaffold per ligand, and sorted them by "packscore," a measure of favorable Van der Waals interactions and hydrogen bonds, with bonuses for bidentate (one side chain contacting two hydrogen-bonding ligand atoms) interactions. We measured the capacity of de novo scaffolds to accommodate binding sites better than natives by subtracting the best (lowest) de novo packscore *z*-score from the best native packscore *z*-score, as described in the main text.

Data and Code Availability. The atomic coordinates have been deposited in the Protein Data Bank, www.wwpdb.org (PDB codes: 6W3D, 6W3F, 6W3G, 6W3W, and 6W40). In order to facilitate reproducibility, improvement, further analysis and use of the models and information in this work, we have made all relevant data and code publicly available on basantab/NTF2Analysis and basantab/NTF2Gen GitHub repositories (GitHub repositories: https://github.com/basantab/NTF2Gen and https://github.com/basantab/NTF2analysis). All sequences, PDB models, analysis scripts, and data tables for the first high-throughput experiments can be found in the ProteaseAnalysisExp1 folder of NTF2Analysis, and ProteaseAnalysisExp2 for the second high-throughput experiment. The set of 32,380 scaffolds, available for general use as starting points for ligand binding and enzyme design, is available in the BeNTF2seq/design_with_PSSM/final_set folder in the basantab/NTF2Gen GitHub repository.

ACKNOWLEDGMENTS. We thank current and former D.B. laboratory members, in particular, Anastassia Vorobieva, Gabriel Rocklin, Mark Lajoie, Enrique Marcos, Ivan Anishchanka, Brian Coventry, William Sheffler, Hahnbeom Park, and Sergey Ovchinnikov for insightful conversations and suggestions. This work was funded by Defence Advanced Research Projects Agency Synergistic Discovery and Design (SD2) HR0011835403 contract FA8750-17-C-0219, Eric and Wendy Schmidt by recommendation of the Schmidt Futures program, and the Institute for Protein Design Directors Fund. C.N. is supported by Novo Nordisk Foundation Grant NNF17OC0030446. This work was facilitated through the use of advanced computational, storage, and networking infrastructure provided by the Hyak supercomputer system at the University of Washington.

1. R. Y. Eberhardt *et al*., Filling out the structural map of the NTF2-like superfamily. *BMC Bioinformatics* **14**, 327 (2013).
2. J. Dou *et al*., Sampling and energy evaluation challenges in ligand binding protein design. *Protein Sci.* **26**, 2426–2437 (2017).
3. R. Obexer *et al*., Active site plasticity of a computationally designed retro-aldolase enzyme. *ChemCatChem* **6**, 1043–1050 (2014).
4. C. E. Tinberg *et al*., Computational design of ligand-binding proteins with high affinity and selectivity. *Nature* **501**, 212–216 (2013).
5. E. Marcos *et al*., Principles for designing proteins with cavities formed by curved β sheets. *Science* **355**, 201–206 (2017).
6. J. Dou *et al*., De novo design of a fluorescence-activating β-barrel. *Nature* **561**, 485–491 (2018).
7. F. Richter, A. Leaver-Fay, S. D. Khare, S. Bjelic, D. Baker, De novo enzyme design using Rosetta3. *PLoS One* **6**, e19230 (2011).
8. P.-S. Huang, S. E. Boyken, D. Baker, The coming of age of de novo protein design. *Nature* **537**, 320–327 (2016).
9. D. N. Woolfson *et al*., De novo protein design: How do we expand into the universe of possible protein structures? *Curr. Opin. Struct. Biol.* **33**, 16–26 (2015).
10. A. Chevalier *et al*., Massively parallel de novo protein design for targeted therapeutics. *Nature* **550**, 74–79 (2017).
11. S. Krish, A practical generative design method. *Comput. Aided Des.* **43**, 88–100 (2011).
12. B. Kuhlman, Designing protein structures and complexes with the molecular modeling program Rosetta. *J. Biol. Chem.* **294**, 19436–19443 (2019).
13. J. K. Leman *et al*., Macromolecular modeling and design in Rosetta: Recent methods and frameworks. *Nat. Methods* **17**, 665–680 (2020).
14. P. S. Huang *et al*., RosettaRemodel: A generalized framework for flexible backbone protein design. *PLoS One* **6**, e24109 (2011).
15. N. Koga *et al*., Principles for designing ideal protein structures. *Nature* **491**, 222–227 (2012).
16. J. S. Richardsont, E. D. Getzofft, D. C. Richardsont, The beta bulge: A common small unit of nonrepetitive protein structure. *Proc. Natl. Acad. Sci. U.S.A.* **75**, 2574–2578 (1978).
17. K. Fujiwara, S. Ebisawa, Y. Watanabe, H. Fujiwara, M. Ikeguchi, The origin of β-strand bending in globular proteins. *BMC Struct. Biol.* **15**, 21 (2015).
18. G. J. Rocklin *et al*., Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science* **357**, 168–175 (2017).
19. J. C. Klein *et al*., Multiplex pairwise assembly of array-derived DNA oligonucleotides. *Nucleic Acids Res.* **44**, e43 (2016).
20. R. F. Alford *et al*., The Rosetta all-atom energy function for macromolecular modeling and design. *J. Chem. Theory Comput.* **13**, 3031–3048 (2017).
21. J. S. Merkel, L. Regan, Aromatic rescue of glycine in beta sheets. *Fold. Des.* **3**, 449–455 (1998).
22. L. McInnes, J. Healy, J. Melville, UMAP: Uniform manifold approximation and projection for dimension reduction. arXiv:1802.03426 (6 December 2018).
23. Y. Zhang, J. Skolnick, TM-align: A protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302–2309 (2005).
24. R. G. Coleman, K. A. Sharp, Protein pockets: Inventory, shape, and comparison. *J. Chem. Inf. Model.* **50**, 589–603 (2010).
25. Y. Lin *et al*., Control over overall shape and size in de novo designed proteins. *Proc. Natl. Acad. Sci. U.S.A.* **112**, E5478–E5485 (2015).
26. E. Marcos, D.-A. Silva, Essentials of de novo protein design: Methods and applications. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **8**, e1374 (2018).
27. F. Richter *et al*., Computational design of catalytic dyads and oxyanion holes for ester hydrolysis. *J. Am. Chem. Soc.* **134**, 16197–16206 (2012).
28. X. Hu, H. Wang, H. Ke, B. Kuhlman, High-resolution design of a protein loop. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 17668–17673 (2007).
29. N. K. Fox, S. E. Brenner, J. M. Chandonia, SCOPe: Structural Classification of Proteins–Extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res.* **42**, D304–D309 (2014).
30. K. J. Fraga, H. Joo, J. Tsai, An amino acid code to define a protein's tertiary packing surface. *Proteins* **84**, 201–216 (2016).
31. T. J. Brunette *et al*., Exploring the repeat protein universe through computational protein design. *Nature* **528**, 580–584 (2015).
32. G. Grigoryan, W. F. Degrado, Probing designability via a generalized model of helical bundle geometry. *J. Mol. Biol.* **405**, 1079–1100 (2011).
33. T. M. Jacobs *et al*., Design of structurally distinct proteins using strategies inspired by evolution. *Science* **352**, 687–690 (2016).
34. W. R. Taylor, A "periodic table" for protein structures. *Nature* **416**, 657–660 (2002).
35. J. G. Greener, L. Moffat, D. T. Jones, Design of metalloproteins and novel protein folds using variational autoencoders. *Sci. Rep.* **8**, 16189 (2018).
36. N. Anand, R. R. Eguchi, A. Derry, R. B. Altman, P.-S. Huang, Protein sequence design with a learned potential. bioRxiv:2020.01.06.895466 (7 January 2020).
37. Y. Zhang, Protein structure prediction: When is it useful? *Curr. Opin. Struct. Biol.* **19**, 145–155 (2009).
38. D. M. Hoover, J. Lubkowski, DNAWorks: An automated method for designing oligonucleotides for PCR-based gene synthesis. *Nucleic Acids Res.* **30**, e43 (2002).
39. F. Pedregosa *et al*., Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
40. A. J. McCoy *et al*., Phaser crystallographic software. *J. Appl. Cryst.* **40**, 658–674 (2007).
41. P. D. Adams *et al*., PHENIX: A comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 213–221 (2010).
42. P. H. Zwart *et al*., Automated structure solution with the PHENIX suite. *Methods Mol. Biol.* **426**, 419–435 (2008).
43. P. Emsley, K. Cowtan, Coot: Model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.* **60**, 2126–2132 (2004).

Downloaded at UNIVERSITY OF WASHINGTON on December 7, 2020