A general-purpose protein design framework based on mining sequence–structure relationships in known protein structures

Jianfu Zhou[a], Alexandra E. Panaitiu[a], and Gevorg Grigoryan[a,b,1]

[a]Department of Computer Science, Dartmouth College, Hanover, NH 03755; and [b]Department of Biological Sciences, Dartmouth College, Hanover, NH 03755

Edited by David Baker, University of Washington, Seattle, WA, and approved December 6, 2019 (received for review May 31, 2019)

Current state-of-the-art approaches to computational protein design (CPD) aim to capture the determinants of structure from physical principles. While this has led to many successful designs, it does have strong limitations associated with inaccuracies in physical modeling, such that a reliable general solution to CPD has yet to be found. Here, we propose a design framework—one based on identifying and applying patterns of sequence–structure compatibility found in known proteins, rather than approximating them from models of interatomic interactions. We carry out extensive computational analyses and an experimental validation for our method. Our results strongly argue that the Protein Data Bank is now sufficiently large to enable proteins to be designed by using only examples of structural motifs from unrelated proteins. Because our method is likely to have orthogonal strengths relative to existing techniques, it could represent an important step toward removing remaining barriers to robust CPD.

protein design | data-driven protein design | structure-based analysis | protein structure | structure search

The robust engineering of protein molecules is a highly sought-after capability, with implications for a range of areas, from therapeutics to materials. Computational protein design (CPD) could be a particularly attractive means of fulfilling the need for such robust engineering, but CPD techniques have thus far lacked the reliability needed to incorporate them as "black-box" tools in downstream research and technology development. The basic idea behind the most ubiquitous approaches to CPD is to model structural phenomena (e.g., folding and binding), to the extent possible, based on physical principles. Since the initial demonstration of this concept by the Mayo group in the late 1990s (1), many groups have implemented significant advancements on the idea (2–8). Notably, the Baker laboratory developed and has continually refined the widely used Rosetta modeling suite, forming an entire community of researchers and programmers actively contributing to the project (2, 3). Advancements introduced over the years have aimed to improve the treatment of a range of physical effects toward a more realistic representation of proteins (3, 9–21). Nevertheless, despite many examples of successful designs in the literature (18, 22–31), it is still the case that CPD methods are not robust. State-of-the-art techniques, even in the hands of experts, fail too frequently, showing that significant inaccuracies are still present in the underlying models and motivating the development of alternative solutions.

In this work, we consider the possibility of performing protein design by directly observing and learning from sequence–structure relationships present in available protein structures, rather than aiming to synthesize them from atomistic interactions. This type of methodology is likely to have entirely orthogonal strengths and weaknesses, relative to the standard CPD approach. If sufficient structure and sequence data are available, a data-driven approach may be difficult to outperform in terms of robustness. However, it is unclear what "sufficient" means in this context and how close we may be to this threshold today. Thus, the 2 main objectives of this work are 1) to develop a general-purpose CPD framework that relies solely on previously available protein structures, and 2) thoroughly benchmark this framework as a means of understanding to what extent the present Protein Data Bank (PDB) is sufficiently large to support practical protein design.

As of March 2019, over 150,000 entries have been deposited into the PDB, with a yearly increase of ∼10,000. Experimental structures have always been a key source of fundamental insights on protein structure–sequence relationships, with degeneracies in structure space—i.e., repeated structural patterns or motifs and associated sequence preferences—proving especially insightful (11, 32–36). In a recent study, we showed that structural degeneracy extends beyond local-in-sequence motifs (e.g., backbone fragments) and into tertiary and quaternary geometries (37). Specifically, we found that local-in-space motifs, which we dubbed TERMs (tertiary motifs), are highly recurrent in the structural universe (37).

Here, we present a CPD framework dTERMen (design with TERM energies) that takes advantage of this degeneracy. As shown in Fig. 1, it systematically breaks the target structure into its constituent TERMs and accounts for the sequence preferences of each by analyzing sequences of closely matching backbone fragments in the PDB, identified using our structure search

Significance

Evolution has given us proteins that perform amazingly complex tasks in living systems, each molecule appearing "custom-built" for its particular purpose. Protein design seeks to enable the "custom building" of proteins at will, for specific tasks, without waiting for evolution. This is a grand challenge, because how a protein's 3-dimensional structure and function are encoded in its amino acid sequence is exceedingly difficult to model. In this paper, we argue that sequence–structure encodings can instead be learned directly from proteins of known structure, which enables an approach to design. We are at an exciting time in protein science, where emergent principles inferred from data may allow us to make headway in cases where application of first principles is challenging.

Author contributions: G.G. designed research; J.Z., A.E.P., and G.G. performed research; J.Z., A.E.P., and G.G. analyzed data; and J.Z. and G.G. wrote the paper.

Competing interest statement: Dartmouth College has filed a provisional patent application protecting the technology behind dTERMen. G.G. is an employee and shareholder of an early-stage biotechnology venture, which may have an interest in using the dTERMen technology.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: gevorg.grigoryan@dartmouth.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1908723117/-/DCSupplemental.

First published December 31, 2019.

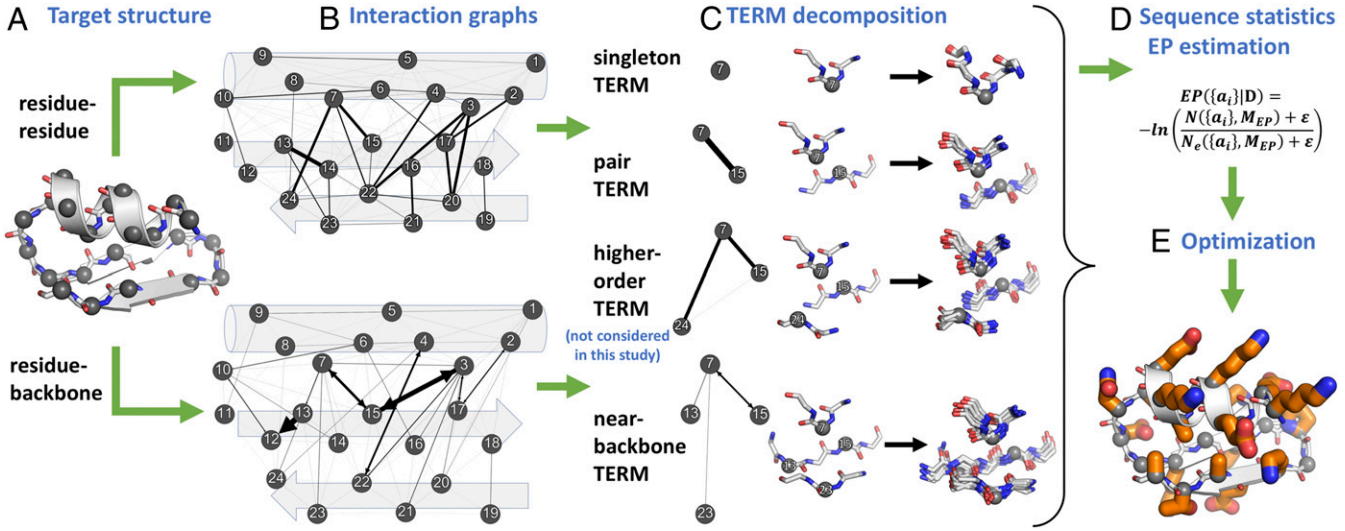

Fig. 1. Diagram of dTERMen procedure. Target structure (*A*) is decomposed into TERMs guided by the graph of its coupled residues (*B*, *Top*) and the graph of residue-backbone influences (*B*, *Bottom*). Close matches to each TERM from the structural database are identified (note, higher-order TERMs were not considered in this study) (*C*), and the sequence alignments implied by these matches are used to estimate EPs governing the sequence–structure relationship in the target structure (*D*). Combinatorial optimization is then used to produce the optimal sequence for the target (*E*) or can also be used to build a library of design variants or for other tasks.

engine MASTER (38). With this information, a sequence-level pseudoenergy table is generated, enabling the scoring of any sequence for compatibility with the target backbone, identification of the optimal sequence, and other optimization or search tasks.

The idea of data-driven CPD has been explored before. First, any statistical potential can be placed into this category of techniques, such that almost any existing CPD method can be thought of as partially data-driven (39–47). A fundamental difference between dTERMen and prior statistical approaches is that dTERMen goes beyond simple geometric descriptors and analyzes apparent sequence preferences in the context of larger well-defined backbone motifs, relying on their apparent quasi-digital nature (i.e., the "TERM hypothesis"). On the other hand, it is different from machine-learning (ML) approaches in that the TERM hypothesis effectively serves as a strong "prior" on the functional form of the model, which allows the method to bridge data sparsity issues.

In contrast to ML, it does not necessitate a model training step. The modularity of TERMs enables dTERMen to exploit entirely unrelated protein structures, broadening its applicability to a great extent. Furthermore, the potential for increased accuracy is effectively built into the method. More structures in the database produce more accurate and refined sequence preferences and, ultimately, more accurate sequence landscapes. Thus, we can expect better performance with time, as the PDB continues to grow.

Results

Results are organized as follows: the first 5 sections describe a series of computational benchmarks of dTERMen, and the sixth section presents the results of applying the method to the total surface redesign of mCherry. Details of experimental and computational procedures are provided in *Materials and Methods*.

dTERMen Procedure Summary. Given a target protein structure for which an appropriate amino acid sequence is needed, dTERMen works by building a table of effective pseudoenergies: self energies describe amino acid preferences at each position of the target, while pair energies capture effective interactions between amino acids at pairs of positions. The framework also supports the calculation of higher-order energies that describe collective contributions of amino acids at larger clusters of positions, but these were not considered in this study. We collectively refer to these pseudoenergy contributions as energy parameters (EPs) and their values are deduced from the statistics of structural matches in the PDB to appropriately defined TERMs comprising the target (Fig. 1 and *Materials and Methods*). The resulting pseudoenergy table is effectively a description of the sequence landscape associated with the target conformation, and can be used to obtain the optimal sequence for the target or perform other optimization or sampling tasks.

dTERMen Predicts Native-Like Sequence from NMR or X-ray Backbones. We first subject dTERMen to the classical "native sequence recovery" benchmark for CPD methods (21). The idea behind this test is that when presented with a native protein structure, a "good" method should propose sequences similar to the corresponding native sequence.

To this end, we curated a set of 90 X-ray and 31 NMR structures of globular proteins, ranging in length from 50 to 150 residues (*Materials and Methods*). dTERMen was applied to each backbone, and the globally optimal sequence was obtained by integer linear programing (ILP) optimization. For comparison, the same backbones were also used in designs by Rosetta, using the *talaris2013* energy function (48) (*Materials and Methods*). Table 1 summarizes the resulting native sequence recovery rates. The 2 methods perform similarly, with dTERMen giving slightly less native-like sequences for X-ray backbones (~29% relative to Rosetta's ~33%, on average) and slightly more native-like ones for NMR backbones (~24% relative to Rosetta's ~22%, on average). Thus, dTERMen performs on par with the state of the art, of which Rosetta Design is a great representative. Interestingly, however, the specific sequences proposed by dTERMen and Rosetta are quite different (see the fifth row of Table 1). This is in line with the fact that the 2 methods choose sequences based on entirely different principles, but it makes the comparable performance on native sequence recovery more interesting.

That dTERMen exhibits somewhat higher native sequence recovery rates on NMR backbones, compared to Rosetta, is consistent with its "fuzzier" interpretation of backbone coordinates (as sequence statistics are discovered in the context of similar, but not identical backbone structural matches). To investigate this

Table 1. dTERMen and Rosetta propose distinct, similarly native-like sequences given native backbone

	X-ray (90)	NMR (31)
Sequence identity		
dTERMen vs. native	28.6 ± 5.8%	23.9 ± 6.1%
Rosetta vs. native	32.6 ± 6.9%	22.2 ± 6.3%
dTERMen vs. Rosetta	26.8 ± 5.7%	22 ± 4.1%
Disulfide identity		
dTERMen vs. native	24/80 (30%)	1/7 (14.3%)
Rosetta vs. native	0/80 (0.0%)	0/7 (0.0%)

Shown are means and SDs of sequence identities between designed and native sequences, within respective datasets. The last 2 rows show the rate of recovering disulfide bonds (i.e., 2 cystine residues designed at locations occupied with disulfide-bonded cystines in the native structure).

apparent insensitivity to backbone noise, we compared sequences designed on alternative NMR backbones as well as those designed on X-ray and NMR structures of the same protein (*Materials and Methods*). Alternative NMR models or X-ray vs. NMR structures of the same protein can be seen as different experimental models of the same exact native state. An ideal CPD method should thus predict very similar sequence landscapes given these different structures. As shown in *SI Appendix*, Table S1, dTERMen is quite consistent across such experimentally equivalent backbones, producing sequences with 40 to 50% sequence identity, on average. Rosetta, on the other hand, shows much greater variability, with sequence identities from equivalent backbones in the range of 20 to 30%.

A closer look at native sequence recovery based on the degree of burial reveals that Rosetta's high performance for X-ray backbones is dominated by core positions, where the method achieves the very high rate of ∼52%, on average, whereas the performance of dTERMen is more uniform across position types (*SI Appendix*, Table S2). The performance of the 2 methods is comparable for interfacial positions and dTERMen produces slightly higher rates for surface positions (see *Materials and Methods* for position type definitions). Relative trends are similar for NMR structures, with the overall performance shifted toward dTERMen (*SI Appendix*, Table S2).

As shown in Table 1, dTERMen has a high rate of disulfide-bond recovery—e.g., 24 out of 80 disulfides (30%) were recovered from X-ray structures (the rate is lower for NMR structures, but it is out of only 7 disulfides occurring in this set). The rate seems especially high when considering that it refers to the simultaneous recovery of 2 residues (in fact, based on the 0.8% frequency of cystines in dTERMen designs, the random expected disulfide recovery rate would be 6.7×10^{-5}). Modeling the energetics of disulfide-bond formation, and balancing it with conformational energetics of the protein, is a challenge and generally an unsolved problem. By contrast, dTERMen effectively sidesteps this challenge, enabling the design of disulfides as a special case in the general strategy of inferring sequence–structure patterns observed in the database.

In addition to proposing native-like sequences for native backbones, the model underlying dTERMen also predicts a pattern of amino acid utilization that is quite close to the native amino acid distribution (see Fig. 2 and *SI Appendix*, Figs. S9 and S10; detailed analysis in *SI Appendix*, Supplementary Results). While this is not entirely unexpected, given that dTERMen is based on native instances to TERM matches, the result nevertheless validates the specific statistical framework used to extract effective pseudoenergy contributions from structural data.

Because dTERMen is entirely based on structural statistics, we reasoned that it may be less capable of making good amino acid choices in regions with few structural representatives in the PDB. However, we do not find a discernable correlation between how structurally well represented a given template is and the rate of native sequence recovery when designing on the template (*SI Appendix*, Fig. S12). Analyzing this on a per-position level, we see that residues with few local structural matches are, on average, slightly more likely to be assigned the native amino acid by dTERMen, than residues with a large number of local matches (*SI Appendix*, Fig. S13; detailed analysis in *SI Appendix*, Supplementary Results). This may suggest that common motifs are inherently more designable, in that they are compatible with a broader sequence space, which makes identifying the native residue more difficult. Loop residues also appear to have a higher sequence recovery rate compared to all positions (*SI Appendix*, Table S7), which is consistent with many loop conformations known to have strong positional amino acid preferences.

dTERMen-Designed Sequences Predicted to Fold to Desired Structures. Folding into the correct structure involves not only forming favorable interactions in the context of the target backbone, but also requires the sequence to disfavor the multitude of available alternative conformations. The latter property, which has been referred to as "fold specificity" (49), is particularly difficult to achieve in CPD, and this is a likely reason behind many design failures. The best way to assess this and other qualities of a designed sequence is to characterize it experimentally. Short of spending the time and resources toward this, however, one can assess whether the designed sequence is at least predicted to fold into the desired structure in silico, using cutting-edge structure prediction methods. Of course, such a prediction cannot serve as ground truth on its own. However, if such a test by structure prediction is performed on a large set of designed sequences, emerging from diverse templates, and used to compare sequences produced by different CPD methods, then statistically significant differences in performance may be interpreted as meaningful.

We performed de novo structure prediction for each sequence from the previous section using a standalone copy of I-TASSER, making sure that data from homologs of the protein whose backbone was used as the design target did not contribute to the calculation (*Materials and Methods*). Each I-TASSER run, which took ∼20 CPU hours on average, was asked to produce 10 models and each was subsequently compared with the desired target structure to extract its template modeling (TM) score (50). Each dTERMen and Rosetta designed sequence was subjected to the same treatment, with Fig. 3*A* comparing the results. As expected, TM scores were not usually close to 1.0, which represents both the difficulty of structure prediction and the fact that some designs may not fold into the desired structure. However, dTERMen design performed better, on average, with their TM scores exceeding the TM score of the corresponding Rosetta design in 58% of cases. The mean TM scores over dTERMen and Rosetta designs were 0.48 and 0.45, respectively ($P = 0.003$), with medians showing a similar trend (Table 2). Furthermore, 43.2% of dTERMen designs exhibited a TM score over 0.5 (a value typically chosen for delineating a roughly correct fold), and only 38% of Rosetta designs reached this value. Models derived from dTERMen sequences also exhibited higher fractions of correct secondary-structure types (Fig. 3*B*).

To address how significant the above differences may be (beyond mere statistical significance) and how good the performance is in an absolute sense, we ran a control calculation, repeating the above analysis for native sequences. Because native sequences do, in fact, fold to the desired structure, their performance in the test can be thought of as that of a "perfect" design method, allowing us to quantify both how far from ideal the methods are and how significant their performance differences are. Fig. 3 *C* and *E* compare the performance of native sequences with that of dTERMen designs and Rosetta designs, respectively, with summary metrics shown in Table 2. Native sequences perform better than both dTERMen and Rosetta, validating our test, dTERMen

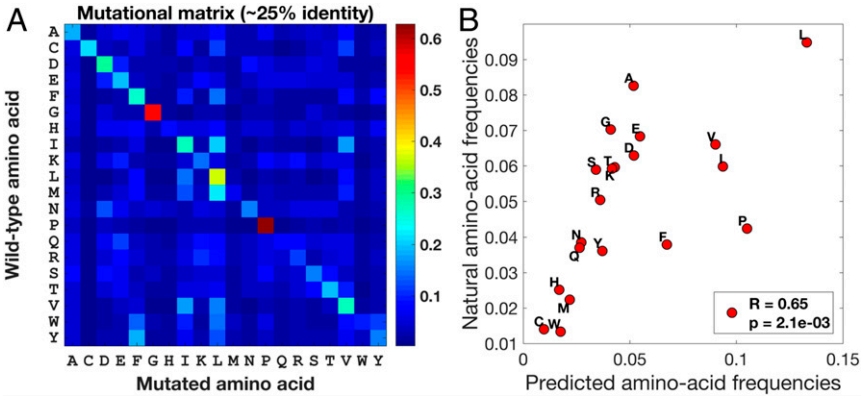

Fig. 2. Pattern of amino acid substitutions predicted by dTERMen is consistent with native amino acid utilization. Shown in *A* is the mutational matrix predicted by dTERMen. Each entry in the matrix is the conditional probability $\boldsymbol{p}(\boldsymbol{X}|\boldsymbol{Y})$, as described in the main text, where $\boldsymbol{X}$ and $\boldsymbol{Y}$ are the amino acids indicated on the *x* and *y* axes, respectively. Color indicates value in accordance with the show color bar. In *B*, the stationary amino acid distribution implied by the matrix in *A* is plotted against the native amino acid distribution found in the PDB. Analogous results obtained with Rosetta Design are shown in *SI Appendix*, Fig. S1.

is second best, and Rosetta is third. Furthermore, the performance of dTERMen, by all metrics, is about halfway between native sequences and Rosetta. For example, 51% of models from native sequences have a TM score above 0.5, while this number is 43% and 38% for dTERMen and Rosetta sequences, respectively. This suggests that the difference between dTERMen and Rosetta sequences is indeed significant. Finally, the difference between dTERMen and native sequences is at the edge of statistical significance. For example, mean TM score is 0.51 for native sequences and 0.48 for dTERMen sequences (*P* value of 0.05; Table 2). In fact, in terms of recovery of the correct secondary structures, dTERMen sequences perform slightly better than native ones, while Rosetta sequences perform worse than native ones (compare *D* and *F* in Fig. 3).

dTERMen Statistical Energy Indicates Design Quality. In a recent tour-de-force study, Baker and coworkers (26) designed de novo and experimentally characterized ∼16,000 sequences for 4 distinct topologies (*SI Appendix*, Fig. S2 *A–D*, *Top*). Each design, along with an approximately equal number of negative-control sequences, was tested, in high throughput, for the ability to form folded, stable, protease-resistant structures. These data represent an unprecedented opportunity for testing design methods, and we apply them to test dTERMen here. De novo design is a challenging task. So, while each of the ∼16,000 designs represented a sequence predicted to be well compatible with the desired target backbone by Rosetta, most designs failed to fold (26). We sought to test whether dTERMen would distinguish between successful and failed designs. To this end, we ran dTERMen on each of the ∼16,000 backbone structures deposited by Baker and coworkers (one for each of their designs) (26). Next, the dTERMen energy score was computed for each designed sequence on its respective backbone, divided by sequence length to facilitate comparison across different topologies. *SI Appendix*, Fig. S2 *A–D* shows, for each of the 4 topologies, the correlation between the resulting score and the experimental "stability score"—a protease resistance-based metric the authors developed to estimate design stability in high throughput, having shown it to correlate closely with thermodynamic stability (26). In each case, the correlation is highly statistically significant (*P* values in legends; *SI Appendix*, Fig. S2 *A–D*). In contrast, Rosetta scores for these sequences, computed using the scoring function used to design them (*talaris2013*), exhibit notably weaker correlations that are statistically insignificant or of the wrong sign in 3 out of 4 cases (*SI Appendix*, Fig. S2 *E–H*). Rocklin et al. also deposited scores from a different Rosetta scoring function, *beta_nov15*, which they had found to perform much better in postevaluating designs in this study. Accordingly, we found that this scoring function exhibits higher and statistically significant correlations in all cases (*SI Appendix*, Fig. S2 *I–L*), beating dTERMen in 3 out of 4 cases. Perhaps more interestingly, dTERMen and *beta_nov15* scores, across all designed sequences, are highly correlated (unlike dTERMen and *talaris2013*; *SI Appendix*, Fig. S8). This is especially remarkable given how fundamentally different the 2 scoring approaches are. The apparent confluence of molecular mechanics-based and structural statistics-based evaluations is encouraging for both types of approaches.

Despite the above correlation, the dTERMen best-scoring sequences for each of the ∼16,000 designed backbones differed considerably from the corresponding Rosetta-based designs (i.e., on average, only ∼16% of positions were identical between Rosetta- and dTERMen-chosen sequences). The fact that dTERMen scores quantify design quality even for sequences that are far from the optimality region of its own predicted sequence landscape validates the generality of the method and the sequence–structure relationships it quantifies. *SI Appendix*, Fig. S3 further shows that the dTERMen score correlates closely with thermodynamic stability, using the same 120 sequence variants of 4 native domains that Rocklin et al. (26) used to establish the quantitative nature of their experimental stability score.

Thus, dTERMen scores appear competitive with state-of-the-art atomistic scoring functions on the highly challenging task of evaluating design quality (especially when the best scoring function to use is not known a priori). Importantly, atomistic scoring functions are applied in conjunction with structural relaxation (i.e., enabling the starting template to minimize, in the context of the specific sequence being evaluated, before computing the final score). This is absolutely required to achieve any reasonable predictability (and the scores for both *talaris2013* and *beta_nov15* were calculated after such relaxation). In contrast, dTERMen scores were derived from the design template, as it was deposited by the authors, without the need for relaxation with respect to the dTERMen scoring function. Thus, to some extent, dTERMen accounts for structural relaxation implicitly, by deriving statistical energies from similar but not identical matches to template substructures. We have previously demonstrated the advantages of such structural "fussiness" in the context of predicting and designing protein–peptide interactions (51).

A Case Study in De Novo Design by dTERMen. Since dTERMen designs sequences based on information from available native protein structures, would the method still apply if the design target is a de novo generated backbone and not a native one? To

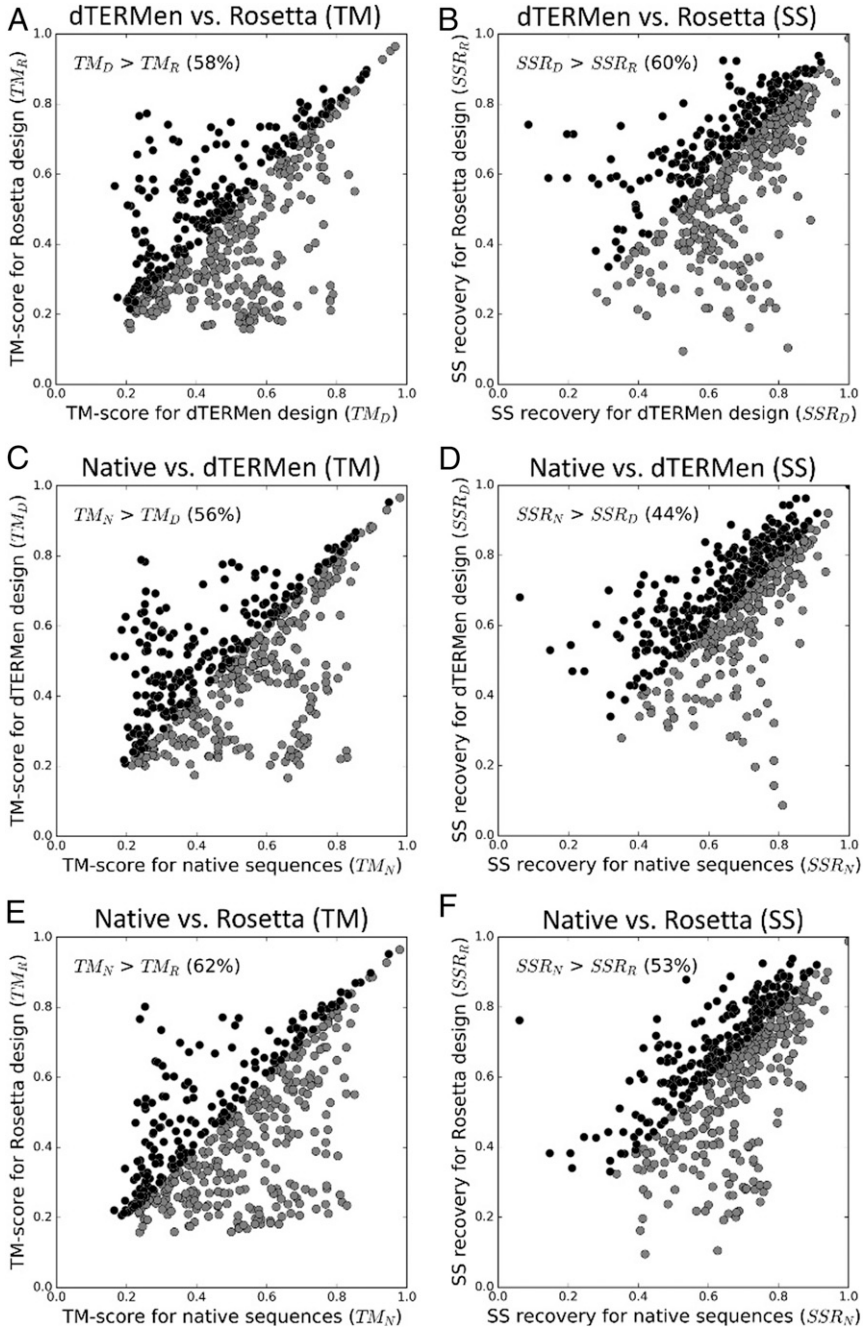

Fig. 3. Testing of dTERMen-designed sequences in structure prediction using I-TASSER. Structures were predicted for 3 sequences corresponding to each target structure (dTERMen-designed, Rosetta-designed, and native), with I-TASSER being asked to predict top 10 models. Models for each sequence were numbered (in the order returned by I-TASSER), allowing us to compare the *i*th model between any 2 sequences (e.g., the top model by dTERMen vs. Rosetta). Each point in each plot represents a comparison between some model i ($i \in [1; 10]$) for 2 sequences from the same template (gray and black points map below and above the diagonal, respectively). (*A* and *B*) Compare dTERMen and Rosetta sequences, (*C* and *D*) compare native and dTERMen sequences, and (*E* and *F*) compare native and Rosetta sequences. In *A*, *C*, and *E*, the comparison is by TM score of the model relative to the native structure; in *B*, *D*, and *F*, the comparison is by fraction of residues with the correct secondary-structure classification. The legend of each plot indicates the fraction of times one set of compared sequences outperforms the other.

interrogate this issue, we considered one of the de novo generated backbones for which Rocklin et al. (26) reported a successfully designed sequence in their recent large-scale design study (*SI Appendix*, Fig. S4*A*). Running dTERMen on this specific backbone, letting it choose any natural amino acid at any of the positions (for a total sequence space of $\sim 10^{52}$), identifies the solution shown in *SI Appendix*, Fig. S4*B* as optimal. The modeled structure of the designed sequence looks biophysically reasonable upon close inspection (*SI Appendix*, Fig. S4*B*). Furthermore, submitting the designed sequence to HHpred, a powerful structure prediction method that relies on the ability to identify remote "homologies" between the modeled sequence and a protein of known structure (52, 53), reveals PDB entry 5UP5 as the closest match (with a probability of over 97% and alignment coverage of 90%)—the very experimental structure of the corresponding sequence designed by Rocklin et al. (26) (*SI Appendix*,

Table 2. Summary of structure prediction performance of dTERMen-designed, Rosetta-designed, and native sequences

	% with TM > 0.5*	Mean TM†	Median TM‡
Native	50.7%	0.508	0.503
dTERMen	43.2%	0.484	0.474
Rosetta	38.0%	0.449	0.427

*Fraction of models built from either sequence set that achieved a TM score above 0.5 (relative to the native structure).
†Mean TM score across all predicted models within each sequence set. The *P* values for the null hypothesis that the true means of underlying distributions are identical are 0.05 for comparing dTERMen and native sequences, 0.003 for comparing dTERMen and Rosetta sequences, and 0.000002 for comparing Rosetta and native sequences.
‡Median TM score across all predicted models within each sequence set.

Fig. S4*C*). Importantly, 5UP5 was not itself used in the database of proteins from which dTERMen sought TERM-based sequence statistics (and, because it itself is a de novo design, no homologs of it were in the database either). Incidentally, the second match revealed by HHpred, PDB entry 1UTA, is a native structure with a fold highly reminiscent of the target (*SI Appendix*, Fig. S4*D*). This strongly suggests that the dTERMen-designed sequence has the necessary features to be especially favoring of the target structure.

Redesign of mCherry Surface. Protein surfaces—i.e., the set of residues exposed to solvent—are important in determining a multitude of biophysical properties, including solubility, immunogenicity, self-association, propensity for aggregation, stability, and fold specificity. It is, therefore, sometimes useful to redesign just the surface of a given protein, so as to modulate one or more of these properties, while preserving its overall structure and function. As an example, let us consider the task of redesigning the surface (resurfacing) of a red fluorescent protein (RFP). RFPs are proteins that naturally fluoresce, with the emission spectrum centered around ∼600 nm. Like other fluorescent proteins (FPs), RPFs are of high utility as biological imaging tags and in optical experiments (54). It may therefore be useful to modulate the surface residues of an RFP depending on the environment (or cell type) in which it has to function.

The crystal structure of RFP mCherry [PDB code 2H5Q (55)] was used as the design template. A total of 64 positions were chosen as being on the surface (corresponding approximately to positions with values of our freedom metric above 0.42; *SI Appendix*, Supplementary Methods); these are shown as spheres in *SI Appendix*, Fig. S5*A*. dTERMen was used to compute a statistical energy table, allowing all of the 64 surface positions to vary among the 20 natural amino acids, with the remaining positions fixed to their identities in the PDB entry 2H5Q. ILP was used to optimize over the resulting space of $20^{64} \approx 2{\cdot}10^{83}$ sequences. The globally optimal-scoring sequence, with 48 out of the 64 variable positions modified relative to mCherry, is shown in *SI Appendix*, Table S3. Comparing surface shapes and *in vacuo* electrostatic potentials between the original mCherry and the design model (*SI Appendix*, Fig. S5 *B* and *C*) reveals the latter to be a significant perturbation.

The designed sequence was cloned into *Escherichia coli*, followed by expression and purification using standard techniques. Size exclusion chromatography (SEC) showed the protein to be monomeric in solution, just as the native mCherry (Fig. 4 *A* and *B*), and the far-UV circular dichroism (CD) spectrum was consistent with a native-like secondary-structure distribution (*SI Appendix*, Fig. S6). Despite harboring 48 mutations and despite the fact that preservation of optical properties was not an explicit design constraint, the design still exhibited the chromophore features characteristic of the original protein (Fig. 4*C*). Furthermore, the designed protein was still fluorescent, with an emission spectrum of the same shape (but lower intensity) as that of mCherry (Fig. 4*D*). Finally, chemical denaturation by guanidinium hydrochloride (GuHCl) revealed that the protein's structure protects its chromophore approximately as well as the original mCherry—a hyperstable, highly engineered protein in its own right (*SI Appendix*, Fig. S7). Thus, by all measures, the designed protein preserved the original structure and even function. The ability to generate such diversity can be easily exploited to quickly engineer variants of RFP or other proteins that possess a range of desired properties.

Discussion

That protein structure can be designed computationally was first established some time ago (1) and demonstrated many times since (22, 56). It is also true that reliance on prior structural data has been broadly explored, both in terms of various statistics-based methods (39, 42, 45, 46) as well as in the creation of chimeras by the fusion of domains or larger fragments of structure (57–60). What is exciting about our results here is the marriage between the generality of our approach (i.e., its ability to design sequences for arbitrarily-defined structures) and its reliance on motif-based structural data. This combination is made possible by the fact that protein structure is not "analog" but "digital" in its nature (37, 61, 62)—local-in-space structural motifs, TERMs, tend to be broadly reused across unrelated proteins. These motifs are small enough to be well sampled in the PDB, but large enough to contain nontrivial sequence determinants of structure. One can thus consider an entirely novel structural template as a design target, while still relying purely on existing structures to select sequences well suitable for folding into it.

Our sequence recovery results, which compare performance on NMR vs. X-ray backbones (Table 1), alternative structures of the same protein (*SI Appendix*, Table S1), and dissect performance by position burial (*SI Appendix*, Table S2), suggest that, relative to dTERMen, Rosetta derives much insight from geometric fit. This is a consideration that arose as important early in the history of protein design, as researchers observed that X-ray structures generally exhibited jigsaw puzzle-like packed cores (63). However, such ideal packing is only feasible in the context of a ground state-like structure. When it comes to room temperature ensembles, the requirement for a crystalline packed core may not be appropriate. Backbone flexibility techniques have been proposed to address the issue that while rotamer-based methods are effectively modeling the ground state, a prespecified template may not represent such a state for any designed sequence (9, 64). In dTERMen, this issue is addressed implicitly, to an extent, by the fact that sequence statistics are gathered from ensembles of close TERM matches. Our extensive tests here (from native sequence recovery, Table 1 and *SI Appendix*, Table S1, to the prediction of design success and thermodynamic impacts of mutations, *SI Appendix*, Figs. S2 and S3, to a de novo design example, *SI Appendix*, Fig. S4, and the redesign of mCherry, Fig. 4 and *SI Appendix*, Figs. S5–S7) support this approach. As further support, when dTERMen-designed sequences are relaxed in Rosetta, they usually produce all-atom scores that are as good or better than corresponding native sequences relaxed in the same way (*SI Appendix*, Fig. S11 and *Materials and Methods*). However, more work is needed to identify the best means of representing the ensemble nature of structure while data mining in the context of TERMs.

In a traditional atomistic approach to design, specific important aspects of the physics underlying protein structure are recognized, parameterized, and included as part of the scoring function. Then, sequences are chosen based on this quantitative (albeit highly approximate) model. In dTERMen, the fundamental "reasons" behind sequence choice are not described beyond observed biases in sequence distributions among database substructures. This can be seen as a disadvantage. The corresponding advantage, however,

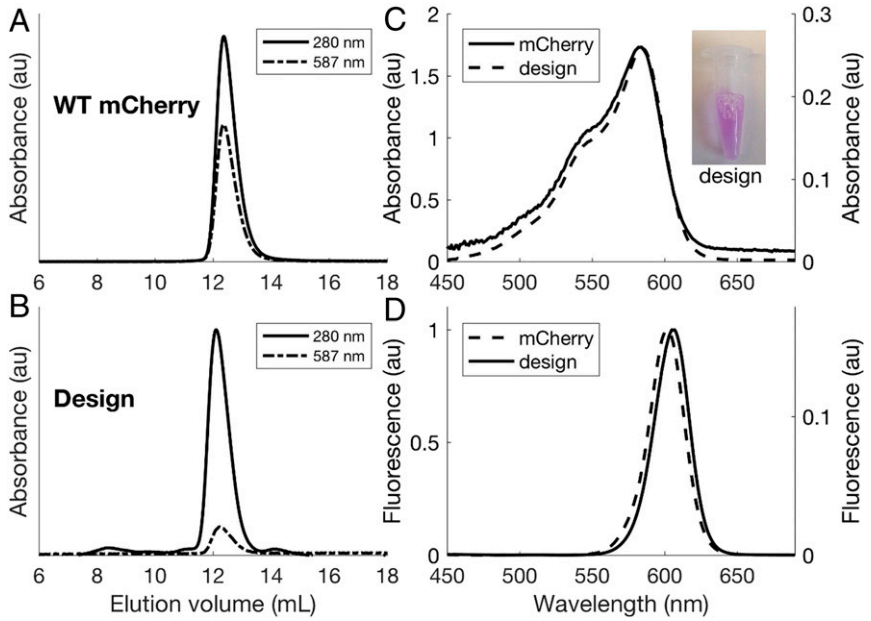

Fig. 4. Solution properties of designed mCherry. Shown in *A* and *B* are the size exclusion chromatograms of wild-type mCherry and the redesigned variant, respectively, run under identical conditions. The design elutes at a nearly identical volume as the wild type (difference in the ratios between absorbances at 280 and 587 nm reflect the lower brightness of the design). *C* and *D* further demonstrate that redesigned mCherry preserves photo properties of the wild-type fluorophore. In *C*, absorbance spectra of wild-type and redesigned mCherry are compared (absorbance values shown on the left and right *y* axes, respectively), while *D* compares fluorescence spectra of the 2 (left and right *y* axes, respectively). Spectra in both *C* and *D* were taken at equivalent concentrations for the 2 proteins, with *y*-axes units reflective relative intensities.

is that complex effects can be included without the need for understanding of their origin (or even being aware of them). A good example of this is disulfide bonds. The physics of these covalent bonds between sidechains of cystine residues is not trivial to model, and to strike the right balance between when to include such bonds and when not to in designing proteins is also not easy. However, as shown in Table 1, dTERMen frequently places disulfide bonds at locations where they appear natively. Importantly, this is not because dTERMen chooses cystines too often—in fact, Cys occurred at a frequency of ∼1% in dTERMen-designed sequences in this study, compared to the rate of ∼2.5% within corresponding native sequences. In addition, in general, amino acid utilization implied by the dTERMen model is in good agreement with the native distribution of amino acids (Fig. 2).

When inspecting design models manually, we frequently see other examples of dTERMen automatically recognizing and utilizing well-known sequence–structure patterns, such as helix-capping motifs (65), salt-bridge patterns within different secondary-structure combinations (66), β-turn preferences (34, 67), and π–cation interactions (68). However, these are just some of the patterns that we recognize, based on our experience with protein structure. It is interesting to consider what other important sequence–structure patterns—those not already well known (by us)—may be automatically included in dTERMen designs.

In summary, the evidence presented here strongly points to the fact that design of protein structure based entirely on sequence patterns mined from the PDB is feasible and practical. A recent study further validates dTERMen on the challenging task of designing protein–protein interactions (69). Based on extensive benchmarking, our general-purpose design framework dTERMen performs on par with or better than the state of the art in CPD. What is most exciting about this finding is that the "top-down" TERM-based insights that dTERMen relies upon are quite distinct from the "bottom-up" molecular mechanics (MM)-based models that are typically used in CPD. We can thus reasonably expect that the 2 methodological classes will have orthogonal strengths and weaknesses. There should be ample opportunity to improve the overall robustness of CPD as a whole by combining TERM- and MM-based insights and by further optimizing the specifics of TERM-based structure mining.

Code Availability

dTERMen is implemented as a Python-based program that makes extensive use of our structure search engine MASTER (38) to identify TERM matches and extract their sequence statistics. The code is freely available for noncommercial purposes from https://grigoryanlab.org/dtermen.

Materials and Methods

dTERMen Procedure. The procedure recognizes several types of effective energetic contributions at play in defining protein sequence–structure relationships: the propensity of an amino acid residue for the general environment of a position, such as the burial state (environmental energy); interactions between an amino acid at a position and its surrounding backbone, which are further broken into contribution from its local-in-sequence backbone fragment (the own-backbone component) and contributions from spatially proximal backbone fragments (the near-backbone component); and interactions between pairs and higher-order clusters of amino acids (note, higher-order interactions were not considered in this study). Environmental own-, and near-backbone energies are self contributions, whereas the remaining ones constitute pair and higher-order contributions.

Once the target structure, *D*, is appropriately decomposed into a set of overlapping TERMs (see below and Fig. 1), and structural matches are identified for each TERM from the database, EP values are deduced following 2 general principles. Principle 1 states that sequence statistics within TERM matches are driven only by the EPs involving positions contained in the TERM (e.g., a pair EP influences the statistics of a TERM if and only if the corresponding pair of positions are contained within the TERM). This assumption is reasonable in cases where the matches arise from a large diversity of structural backgrounds, such that context effects average out. Certain redundancy-removal steps are key to making sure that this assumption holds well in practice (see below). It follows from principle 1 that EP values should be sought to maximally describe the sequence data observed in TERM matches. Principle 2 stipulates that higher-order parameters be involved only when needed—i.e., models involving only lower-order

parameters are preferred, all else being equal. This means that higher-order EPs act as correctors to lower-order contributions. For example, pair energies are needed only to describe those aspects of sequence statistics that are not satisfactorily described with self contributions.

TERM decomposition: Main idea. Within the sequence space compatible with folding into *D*, some residue pairs are coupled—i.e., the optimal amino acid identity of one residue depends on the identity of the other. Such coupled positions can be identified through the structure of *D*, by finding position pairs capable of hosting amino acids that have an influence on each other via direct or indirect physical interactions (see below). In addition, in some systems with sufficiently large multiple sequence alignments, evolutionary covariation can suggest coupled positions. Finally, experimental evidence identifying specific coupled position pairs may also be available.

Whatever the source of the inference, the coupling relationships in *D* can be thought of as an undirected graph, where nodes represent residues and edges signify coupling, with edge weights optionally indicating the strength of coupling (inferred from structure or known); let us call this graph *G*. The final pseudoenergy model should involve self contributions for all nodes, pair contributions for all edges, and (optionally) higher-order contributions for a subset of connected subgraphs of *G*. Furthermore, to describe near-backbone interactions, we define a directed graph, *B*, in which nodes represent residues and a directed edge between *a* and *b*, $a \rightarrow b$, signifies that the backbone of *b* can influence the amino acid choice at *a*. As with coupling, such pairs of positions can be identified through a structural analysis of *D* (see below). A TERM decomposition of *D* should respect the structures of *G* and *B* and enable the extraction of EPs for the above contribution types. Specifically, a complete set of TERMs describing *D* must be such that every residue and every pair of coupled residues be covered by at least one TERM. In addition, if higher-order coupling contributions are desired, TERMs covering corresponding connected subgraphs of *G* should be included as well. Similarly, if higher-order near-backbone contributions for position *i* are desired, TERMs covering *i* and all (or a subset of) nodes to which it has directed edges in *B* should be included as well.

TERM decomposition: Specifics. Here, we describe the specific TERM decomposition procedure used in this study (Fig. 1), noting that many other procedures that follow the above principles can be appropriate. We define TERMs via connected subgraphs of *G* or *B*. If a subgraph includes the node corresponding to residue *i*, then the resulting TERM includes residues $(i-n)$ through $(i+n)$, where *n* is a parameter (we generally use $n=1$ or 2, and exclusively $n=1$ in this study) (Fig. 1). We first define a TERM for each node in isolation (i.e., treating it as a one-node subgraph); we refer to these as singleton TERMs. Singletons are used to deduce own-backbone contributions (see below). Next, to capture near-backbone contributions at residue *i*, we create a TERM that involves node *i* and all nodes to which it has directed edges in *B*; let us call this set $\beta(i)$—the "influencing" residues. If such a TERM does not have a sufficient number of close structural matches in the database (see below for details of match definition), the effect of the neighboring backbone on *i* needs to be captured with multiple TERMs. In particular, we start by defining TERMs containing *i* and each residue in $\beta(i)$, independently. Of these TERMs, the one with the most structural matches [suppose it is the one containing nodes *i* and $j \in \beta(i)$] is chosen for expansion, with each remaining node $k \in \beta(i) \cap j$ considered for inclusion into the subgraph, one at a time. Once again, of these, the one with the most matches is selected, and this procedure is repeated until no more nodes can be included into the growing subgraph. Once this occurs, the expanded TERM is accepted into the overall TERM decomposition, with all of the influencing residues involved in it marked as covered. The procedure is then repeated, using only uncovered influencing residues, until all residues in $\beta(i)$ are covered. This technique is a generalization of considering a single TERM that covers *i* and all $\beta(i)$, splitting the near-backbone effect into as few TERMs as needed to retain sufficiently good sequence statistics, while capturing as much of the near-backbone environment simultaneously as possible. TERMs generated for capturing near-backbone effects are referred to as near-backbone TERMs (Fig. 1).

We next define one TERM for each pair of nodes in *G* connected by an edge. These are referred to as pair TERMs and used to deduce pair interaction EPs. Finally, higher-order TERMs are defined for select connected subgraphs of *G* and used in deducing higher-order interactions. Individual higher-order subgraphs can either be chosen manually, based on prior knowledge of the system or inspection of structure *D*, or automatically using an appropriate structure-based rule (e.g., only fully connected 3-residue subgraphs, potentially filtered by edge weights). These TERMs are only included if they possess a sufficient number of structural matches (see below for details). While our method can extract higher-order couplings (provided enough data are available), we have generally found it unnecessary to do so in practice, and all of the examples presented in this study included only up to pair contributions.

Computing EPs. Following the 2 general principles outlined in *dTERMen Procedure*, many specific computational procedures can be formulated to extract EP values from the data provided by a TERM decomposition (i.e., TERM matches and their sequence statistics). Here, we employ a procedure that considers pseudoenergetic contributions in a hierarchy, with each next type of contribution introduced only to describe what is not already captured by previous ones. By including higher-order contributions later in the hierarchy, we make sure that these are only used as correctors (to the extent necessary) over what is already described by lower-order contributions. Furthermore, the earliest contributions in the hierarchy are those associated with the strongest sequence statistics, such that highest-confidence effects are captured first, relatively unaffected by statistical noise. The specific order of contributions in the hierarchy used here is: 1) amino acid backbone φ/ψ dihedral angle propensities, 2) amino acid backbone ω dihedral angle propensities, 3) pseudoenergy associated with the general environment (burial state) of a residue, 4) own-backbone contributions, 5) near-backbone contributions, 6) residue pair contributions, and 7) any considered higher-order contributions (not computed in this study). The details of pseudoenergy calculation are presented in *SI Appendix*, Supplementary Methods.

Native Sequence Recovery. To arrive at the list of templates used in native sequence recovery tests, the full list of domains in the CATH database (version 4.2.0) was downloaded on February 11, 2018 (70). The list was filtered using the following criteria: 1) each domain had to be an entire chain of the corresponding PDB entry that was nondeprecated and monomeric (i.e., both biological and asymmetric units containing a single chain), 2) domains in the "few secondary structures" CATH class were excluded, 3) domains corresponding to membrane-protein PDB entries [i.e., those listed in the OPM database (71)] were excluded, 4) only domains ranging from 50 to 150 residues in lengths, consisting entirely of natural amino acids (including MSE, HSC, HSD, HSE, and HSP), and with no missing nonhydrogen backbone atoms were allowed, and 5) for X-ray PDB entries, only those with resolution of 2.6 Å or better were allowed. The resulting list was split into 10 bins by domain length (i.e., [50, 60), [60, 70), [70, 80), ..., [130, 140), and [140, 150]), with 8 X-ray and 2 NMR structures selected from each bin manually, making sure that structures chosen from the same bin belonged to different CATH topologies. This gave a set of 100 monomeric, single-domain, water-soluble structures with 80 X-ray and 20 NMR entries (set I; *SI Appendix*, Table S4).

We also considered the 11 pairs of structures, each pair representing one NMR and one X-ray structure of the same protein, curated in our earlier work [i.e., sets X-ray-2 and NMR-2 from Mackenzie et al. (37)], here referred to as set II (*SI Appendix*, Table S5). Tests comparing design performance on NMR vs. X-ray structures or alternative NMR models used set II structures, while all other sequence-recovery tests used the union of set I and set II, containing 90 X-ray and 31 NMR structures (X-ray entry 1TTZ occurred in both set I and set II). Matching entries 3IBW (X-ray) and 2KO1 (NMR) from set II are homodimers (all others being monomeric), so one of the monomers was kept at its wild-type sequence during design with both dTERMen and Rosetta. Each NMR entry in set II contained 20 models, so each gave rise to 190 model-to-model comparisons, giving a total of 2,090 such comparisons across set II. For NMR-to-X-ray comparisons, each X-ray entry was compared to each of the 20 models of the corresponding NMR entry.

Positions were classified into surface, interface, and core using solvent-accessible surface area (SASA) values computed in the context of the native protein used as the template. Specifically, Stride (downloaded on June 24, 2018) was used to calculate absolute SASA values for each residue (72), and these were divided by "standard" reference SASA values for each amino acid type to obtain relative SASAs. Standard values were taken from GetArea (i.e., for each residue type X, its solvent-accessible surface area in the tripeptide Gly-X-Gly, averaged over a set of 30 random conformations) (73). Residues were labeled as surface if the relative SASA exceeded 40%, as core if the value was below 20%, and interface for cases between 20% and 40%.

Disulfide bonds in native structures were identified as instances of 2 cystine residues with SG atoms within 3.0 Å of each other, resulting in a total of 80 and 7 Cys–Cys bonds in all X-ray and NMR structures considered, respectively. Disulfide bond recovery was computed as the fraction of times the designed sequences retained 2 cystines at position pairs that were natively disulfide bonded.

Design and Relaxation with Rosetta. We used pyRosetta (Linux release r56316.64Bit) in all Rosetta Design tests, as well as to repack dTERMen-designed sequences onto target backbones (e.g., for visualization in *SI Appendix*, Fig. S5). Specifically, we performed fixed-backbone design using the

talaris2013 force-field and default parameters in pyRosetta via "standard_packer_task" and "PackRotamersMover" objects (for building structural models of dTERMen designs, only the single amino acid from the designed sequence was allowed at each position). Specifically, the relevant portion of Python code we used is shown in *SI Appendix*, Table S6. Rosetta *Relax* protocol (74–77) (Rosetta 3.8 Linux release 2017.08.59291) was used to minimize both native and repacked dTERMen-designed structures, with *beta_nov15* as scoring function.

Structure Prediction Test. Sequences designed for the 100 structures in set I (*SI Appendix*, Table S4), as well as their native counterparts, were subjected to structure prediction using standalone I-TASSER (version 5.1, downloaded on June 4, 2018) (78). Specifically, I-TASSER was run in fast mode, with at most 5 h for each round of simulation, producing at most 10 final models. Information from homologs of the protein used as the design template (i.e., the native sequence) was excluded from I-TASSER prediction. To this end, we used *blastpgp* from the standalone BLAST packages (version 2.2.26) to search the PDB (i.e., the preformatted BLAST database file *pdbaa* downloaded from National Center for Biotechnology Information on June 6, 2018) for homologs of the native sequence using the E-value cutoff of 1 (79). Chains corresponding to any matches, as well as the design template itself, were then removed from the I-TASSER template library during runs using the -temp_excl flag. With these settings, an I-TASSER run took around ∼20-h wall-clock time, on average.

All of the predicted models were further compared to their respective design templates via TM score and secondary structure recovery. The former was calculated using TM-align (downloaded on June 24, 2018) (80). Stride (downloaded on June 24, 2018) was used to identify the secondary structure for each residue in models and native structures (72).

While I-TASSER was asked to return up to 10 best models for each sequence, fewer models were produced in some cases (due to the inability of the method to identify a sufficient number of structural templates). When comparing models across different sequences categories (e.g., in Fig. 3 and Table 2), the same index model was always compared. For example, model 3 for the dTERMen sequence designed on the backbone of protein *X* was compared with model 3 of the Rosetta sequence designed on this backbone. Thus, if (for example) the dTERMen sequence resulted in 10 models and the Rosetta sequence produced 9 models, only the first 9 were compared. In total, there were 481 models that were successfully produced for all 3 sequence types (dTERMen, Rosetta, and native), and all comparisons were made using only these. This included models for 83 targets (I-TASSER produced no models in 13/100, 13/100, and 15/100 cases for dTERMen, Rosetta, and native sequences, respectively).

Experimental Characterization of mCherry Design. Both wild-type and design mCherry construct genes were synthesized, sequence-verified, and cloned into plasmids by Gen Script (pUC57 for wild-type mCherry and a modified pET28b for the design variant; in this modified plasmid, the factor Xa cleavage site was replaced with a tobacco etch virus or TEV protease site). Wild-type mCherry was subcloned into a standard pET28b using Agilent's QuikChange Lightning site-directed mutagenesis kit through a PCR insertion method relying on distal end homology between insert and template. The cloned construct sequence was confirmed by DNA sequencing (Dartmouth College Molecular Biology Core Facility).

Protein expression and purification. Both proteins were expressed in *E. coli* Rosetta 2 (DE3) cells made competent in-house. Expression was carried out through induction for 17 h at 20 °C by addition of 0.2 mM IPTG at an OD_{600} around 0.7 to 0.9. Cells were subsequently harvested by centrifugation at 3,000 rpm for 25 min, and the pellets were resuspended in 30 mL of FPLC binding buffer (50 mM Tris·HCl, 250 mM sodium chloride, 20 mM imidazole, pH 8.0). Cells were lysed using a Microfluidizer. The soluble protein fraction was cleared by centrifugation at 20,000 rpm for 40 min. The proteins from the lysed cultures were purified by means of affinity chromatography on a GE Healthcare Akta PureM FPLC system on Ni-NTA–conjugated resin (GE Healthcare HisTrap HP 5-mL column) followed by SEC on a GE Healthcare Superdex-75 16/600 prep-grade column or a Superdex Increase 10/300 GL column.

CD. Folding and stability of wild-type and design mCherry constructs were assessed by CD on a Jasco J-815 instrument. All samples contained 10 to 20 μM protein in 25 mM sodium phosphate, 150 mM sodium chloride, pH 7.5. CD scans were acquired at 20 °C with 4 accumulations each in the 250- to 200-nm UV range, at 100 nm/min, and with a 1-nm bandwidth, and a pitch of 0.1 nm.

Fluorescence. Fluorescence spectra were recorded on a synchronous scanning Jasco FP-8000 fluorometer. All samples contained either 38 μM protein (wild type) or 150 μM protein (design variant) in 25 mM sodium phosphate, 150 mM sodium chloride, pH 7.5. Scans were acquired over a wavelength range of 400 to 700 nm, with excitation and emission bandwidths of 5 nm, a 50-ms response time, and a 200-nm/min scan speed.

ACKNOWLEDGMENTS. This work was funded by NSF Award DMR1534246 (to G.G.) and NIH Award P20-GM113132 (to G.G.).

1. B. I. Dahiyat, S. L. Mayo, De novo protein design: Fully automated sequence selection. *Science* **278**, 82–87 (1997).
2. A. Leaver-Fay *et al*., ROSETTA3: An object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol.* **487**, 545–574 (2011).
3. R. F. Alford *et al*., The Rosetta all-atom energy function for macromolecular modeling and design. *J. Chem. Theory Comput.* **13**, 3031–3048 (2017).
4. M. A. Hallen *et al*., OSPREY 3.0: Open-source protein redesign for you, with powerful new features. *J. Comput. Chem.* **39**, 2494–2507 (2018).
5. O. Sharabi, C. Yanover, A. Dekel, J. M. Shifman, Optimizing energy functions for protein-protein interface design. *J. Comput. Chem.* **32**, 23–32 (2011).
6. B. I. Dahiyat, D. B. Gordon, S. L. Mayo, Automated design of the surface positions of protein helices. *Protein Sci.* **6**, 1333–1337 (1997).
7. T. Simonson *et al*., Computational protein design: The Proteus software and selected applications. *J. Comput. Chem.* **34**, 2472–2484 (2013).
8. J. Van Durme *et al*., A graphical interface for the FoldX forcefield. *Bioinformatics* **27**, 1711–1712 (2011).
9. D. J. Mandell, T. Kortemme, Backbone flexibility in computational protein design. *Curr. Opin. Biotechnol.* **20**, 420–428 (2009).
10. M. V. Shapovalov, R. L. Dunbrack, Jr, A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure* **19**, 844–858 (2011).
11. R. L. Dunbrack, Jr, Rotamer libraries in the 21st century. *Curr. Opin. Struct. Biol.* **12**, 431–440 (2002).
12. T. Lazaridis, M. Karplus, Effective energy function for proteins in solution. *Proteins* **35**, 133–152 (1999).
13. G. Archontis, T. Simonson, A residue-pairwise generalized born scheme suitable for protein design calculations. *J. Phys. Chem. B* **109**, 22667–22673 (2005).
14. L. Jiang, B. Kuhlman, T. Kortemme, D. Baker, A "solvated rotamer" approach to modeling water-mediated hydrogen bonds at protein-protein interfaces. *Proteins* **58**, 893–904 (2005).
15. G. Grigoryan, Absolute free energies of biomolecules from unperturbed ensembles. *J. Comput. Chem.* **34**, 2726–2741 (2013).
16. I. Georgiev, R. H. Lilien, B. R. Donald, The minimized dead-end elimination criterion and its application to protein redesign in a hybrid scoring and search algorithm for computing partition functions over molecular ensembles. *J. Comput. Chem.* **29**, 1527–1542 (2008).
17. P. Gainza *et al*., OSPREY: Protein design with ensembles, flexibility, and provable algorithms. *Methods Enzymol.* **523**, 87–107 (2013).
18. N. H. Joh *et al*., De novo design of a transmembrane Zn^{2+}-transporting four-helix bundle. *Science* **346**, 1520–1524 (2014).
19. T. Kortemme, A. V. Morozov, D. Baker, An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. *J. Mol. Biol.* **326**, 1239–1259 (2003).
20. E. I. Shakhnovich, A. M. Gutin, A new approach to the design of stable proteins. *Protein Eng.* **6**, 793–800 (1993).
21. B. Kuhlman, D. Baker, Native protein sequences are close to optimal for their structures. *Proc. Natl. Acad. Sci. U.S.A.* **97**, 10383–10388 (2000).
22. P. S. Huang, S. E. Boyken, D. Baker, The coming of age of de novo protein design. *Nature* **537**, 320–327 (2016).
23. B. Kuhlman *et al*., Design of a novel globular protein fold with atomic-level accuracy. *Science* **302**, 1364–1368 (2003).
24. P. S. Huang *et al*., De novo design of a four-fold symmetric TIM-barrel protein with atomic-level accuracy. *Nat. Chem. Biol.* **12**, 29–34 (2016).
25. N. Koga *et al*., Principles for designing ideal protein structures. *Nature* **491**, 222–227 (2012).
26. G. J. Rocklin *et al*., Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science* **357**, 168–175 (2017).
27. K. E. Roberts, P. R. Cushing, P. Boisguerin, D. R. Madden, B. R. Donald, Computational design of a PDZ domain peptide inhibitor that rescues CFTR activity. *PLoS Comput. Biol.* **8**, e1002477 (2012).
28. S. Q. Zhang *et al*., De novo design of tetranuclear transition metal clusters stabilized by hydrogen-bonded networks in helical bundles. *J. Am. Chem. Soc.* **140**, 1294–1304 (2018).
29. B. Dang *et al*., De novo design of covalently constrained mesosize protein scaffolds with unique tertiary structures. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 10852–10857 (2017).
30. C. Y. Chen, I. Georgiev, A. C. Anderson, B. R. Donald, Computational structure-based redesign of enzyme activity. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 3764–3769 (2009).
31. P. Lu *et al*., Accurate computational design of multipass transmembrane proteins. *Science* **359**, 1042–1046 (2018).
32. K. F. Han, D. Baker, Global properties of the mapping between local amino acid sequence and local structure in proteins. *Proc. Natl. Acad. Sci. U.S.A.* **93**, 5814–5818 (1996).

33. K. T. O'Neil, W. F. DeGrado, A thermodynamic scale for the helix-forming tendencies of the commonly occurring amino acids. *Science* **250**, 646–651 (1990).
34. E. G. Hutchinson, J. M. Thornton, A revised set of potentials for beta-turn formation in proteins. *Protein Sci.* **3**, 2207–2216 (1994).
35. D. E. Engel, W. F. DeGrado, Alpha-alpha linking motifs and interhelical orientations. *Proteins* **61**, 325–337 (2005).
36. C. O. Mackenzie, G. Grigoryan, Protein structural motifs in prediction and design. *Curr. Opin. Struct. Biol.* **44**, 161–167 (2017).
37. C. O. Mackenzie, J. Zhou, G. Grigoryan, Tertiary alphabet for the observable protein structural universe. *Proc. Natl. Acad. Sci. U.S.A.* **113**, E7438–E7447 (2016).
38. J. Zhou, G. Grigoryan, Rapid search for tertiary fragments reveals protein sequence-structure relationships. *Protein Sci.* **24**, 508–524 (2015).
39. P. Xiong *et al*., Protein design with a comprehensive statistical energy function and boosted by experimental selection for foldability. *Nat. Commun.* **5**, 5330 (2014).
40. X. Zhou *et al*., Proteins of well-defined structures can be designed without backbone readjustment by a statistical model. *J. Struct. Biol.* **196**, 350–357 (2016).
41. P. Xiong, Q. Chen, H. Liu, Computational protein design under a given backbone structure with the ABACUS statistical energy function. *Methods Mol. Biol.* **1529**, 217–226 (2017).
42. C. M. Topham, S. Barbe, I. André, An atomistic statistically effective energy function for computational protein design. *J. Chem. Theory Comput.* **12**, 4146–4168 (2016).
43. P. Mitra *et al*., An evolution-based approach to De Novo protein design and case study on *Mycobacterium tuberculosis*. *PLoS Comput. Biol.* **9**, e1003298 (2013).
44. P. Mitra, D. Shultis, Y. Zhang, EvoDesign: De novo protein design based on structural and evolutionary profiles. *Nucleic Acids Res.* **41**, W273–W280 (2013).
45. J. R. Brender, D. Shultis, N. A. Khattak, Y. Zhang, An evolution-based approach to de novo protein design. *Methods Mol. Biol.* **1529**, 243–264 (2017).
46. V. Potapov, J. B. Kaplan, A. E. Keating, Data-driven prediction and design of bZIP coiled-coil interactions. *PLoS Comput. Biol.* **11**, e1004046 (2015).
47. J. Wang, H. Cao, J. Z. H. Zhang, Y. Qi, Computational protein design with deep learning neural networks. *Sci. Rep.* **8**, 6349 (2018).
48. A. Leaver-Fay *et al*., Scientific benchmarks for guiding macromolecular energy function improvement. *Methods Enzymol.* **523**, 109–143 (2013).
49. J. O. Wrabl, S. A. Larson, V. J. Hilser, Thermodynamic environments in proteins: Fundamental determinants of fold specificity. *Protein Sci.* **11**, 1945–1957 (2002).
50. Y. Zhang, J. Skolnick, Scoring function for automated assessment of protein structure template quality. *Proteins* **57**, 702–710 (2004).
51. F. Zheng *et al*., Computational design of selective peptides to discriminate between similar PDZ domains in an oncogenic pathway. *J. Mol. Biol.* **427**, 491–510 (2015).
52. L. Zimmermann *et al*., A completely reimplemented MPI bioinformatics toolkit with a new HHpred server at its core. *J. Mol. Biol.* **430**, 2237–2243 (2018).
53. A. Meier, J. Söding, Automatic prediction of protein 3D structures by probabilistic multi-template homology modeling. *PLoS Comput. Biol.* **11**, e1004343 (2015).
54. R. N. Day, M. W. Davidson, The fluorescent protein palette: Tools for cellular imaging. *Chem. Soc. Rev.* **38**, 2887–2921 (2009).
55. X. Shu, N. C. Shaner, C. A. Yarbrough, R. Y. Tsien, S. J. Remington, Novel chromophores and buried charges control color in mFruits. *Biochemistry* **45**, 9639–9647 (2006).
56. I. Samish, "Achievements and challenges in computational protein design" in *Computational Protein Design*, I. Samish, Ed. (Springer, New York, 2017), pp. 21–94.
57. O. Khersonsky, S. J. Fleishman, Why reinvent the wheel? Building new proteins based on ready-made parts. *Protein Sci.* **25**, 1179–1187 (2016).
58. E. Verschueren *et al*., Protein design with fragment databases. *Curr. Opin. Struct. Biol.* **21**, 452–459 (2011).
59. P. Heinzelman, P. A. Romero, F. H. Arnold, Efficient sampling of SCHEMA chimera families to identify useful sequence elements. *Methods Enzymol.* **523**, 351–368 (2013).
60. S. Shanmugaratnam, S. Eisenbeis, B. Höcker, A highly stable protein chimera built from fragments of different folds. *Protein Eng. Des. Sel.* **25**, 699–703 (2012).
61. F. Zheng, J. Zhang, G. Grigoryan, Tertiary structural propensities reveal fundamental sequence/structure relationships. *Structure* **23**, 961–971 (2015).
62. F. Zheng, G. Grigoryan, Sequence statistics of tertiary structural motifs reflect protein stability. *PLoS One* **12**, e0178272 (2017).
63. S. F. Betz, D. P. Raleigh, W. F. DeGrado, De novo protein design: From molten globules to native-like states: Current opinion in structural biology. *Curr. Opin. Struct. Biol.* **3**, 601–610 (1993).
64. J. T. MacDonald, P. S. Freemont, Computational protein design with backbone plasticity. *Biochem. Soc. Trans.* **44**, 1523–1529 (2016).
65. R. Aurora, G. D. Rose, Helix capping. *Protein Sci.* **7**, 21–38 (1998).
66. J. E. Donald, D. W. Kulp, W. F. DeGrado, Salt bridges: Geometrically specific, designable interactions. *Proteins* **79**, 898–915 (2011).
67. S. T. Phillips, G. Piersanti, P. A. Bartlett, Quantifying amino acid conformational preferences and side-chain-side-chain interactions in beta-hairpins. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 13737–13742 (2005).
68. J. P. Gallivan, D. A. Dougherty, Cation-pi interactions in structural biology. *Proc. Natl. Acad. Sci. U.S.A.* **96**, 9459–9464 (1999).
69. V. Frappier, J. M. Jenson, J. Zhou, G. Grigoryan, A. E. Keating, Tertiary structural motif sequence statistics enable facile prediction and design of peptides that bind anti-apoptotic bfl-1 and mcl-1. *Structure* **27**, 606–617.e5 (2019).
70. N. L. Dawson *et al*., CATH: An expanded resource to predict protein function through structure and sequence. *Nucleic Acids Res.* **45**, D289–D295 (2017).
71. M. A. Lomize, I. D. Pogozheva, H. Joo, H. I. Mosberg, A. L. Lomize, OPM database and PPM web server: Resources for positioning of proteins in membranes. *Nucleic Acids Res.* **40**, D370–D376 (2012).
72. D. Frishman, P. Argos, Knowledge-based protein secondary structure assignment. *Proteins* **23**, 566–579 (1995).
73. R. Fraczkiewicz, W. Braun, Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules. *J. Comput. Chem.* **19**, 319–333 (1998).
74. L. G. Nivón, R. Moretti, D. Baker, A Pareto-optimal refinement method for protein design scaffolds. *PLoS One* **8**, e59004 (2013).
75. P. Conway, M. D. Tyka, F. DiMaio, D. E. Konerding, D. Baker, Relaxation of backbone bond geometry improves protein energy landscape modeling. *Protein Sci.* **23**, 47–55 (2014).
76. F. Khatib *et al*., Algorithm discovery by protein folding game players. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 18949–18953 (2011).
77. M. D. Tyka *et al*., Alternate states of proteins revealed by detailed energy landscape mapping. *J. Mol. Biol.* **405**, 607–618 (2011).
78. J. Yang *et al*., The I-TASSER suite: Protein structure and function prediction. *Nat. Methods* **12**, 7–8 (2015).
79. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
80. Y. Zhang, J. Skolnick, TM-align: A protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302–2309 (2005).