DEEP SHARPENING OF TOPOLOGICAL FEATURES FOR DE NOVO PROTEIN DESIGN

Zander Harteveld
EPFL, CH
zander.harteveld@epfl.ch

Joshua Southern
Imperial College London, UK
jks17@ic.ac.uk

Michaël Defferrard
EPFL, CH
michael.defferrard@epfl.ch

Andreas Loukas
Prescient Design, Genentech
andreas.loukas@roche.ch

Pierre Vandergheynst
EPFL, CH
pierre.vandergheynst@epfl.ch

Michael M. Bronstein
University of Oxford, UK
michael.bronstein@cs.ox.ac.uk

Bruno E. Correia
EPFL, CH
bruno.correia@epfl.ch

ABSTRACT

Computational *de novo* protein design allows the exploration of uncharted areas of the protein structure and sequence spaces. Classical approaches to *de novo* protein design involve an iterative process where the desired protein shape is outlined, then sampled for structural backbones and designed with low energy amino acid sequences. Despite numerous successes, inaccuracies within energy functions and sampling methods often lead to physically unrealistic protein backbones yielding sequences that fail to fold experimentally. Recently, deep neural networks have successfully been used to design novel protein folds from scratch by iteratively predicting a structure and optimizing the sequence until a target protein structure is reached. These methods work well under circumstances where distributions of physically realistic target protein backbones can be readily defined, but lack the ability to *de novo* design loosely specified protein shapes. In fact, a major challenge for *de novo* protein design is to generate "designable" protein structures for defined folds, including native and artificial ("dark matter") folds that can then be used to find low energetic sequences in a generic manner. Here, we automate the task of creating designable backbones using a variational autoencoder framework, termed GENESIS, to denoise sketches of protein topological lattice models by sharpening their 2D representations in distance and angle feature maps. In conjunction with the trRosetta design framework, large pools of diverse sequences for different protein folds were generated for the maps. We found that the GENESIS-trDesign framework generates native-like feature maps for known and dark matter protein folds. Ultimately, the GENESIS framework addresses the protein backbone designability problem and could contribute to the *de novo* design of structurally defined artificial proteins that can be tailored for novel functionalities.

1 INTRODUCTION

Evolution is a slow and gradual process that has only sampled a very small fraction of the possible protein amino acid (AA) sequence space (Grant et al., 2004; Kolodny et al., 2013). In order to explore new sequences that fold into well-defined three-dimensional (3D) conformations outside the natural repertoire and are amenable to tailored functionalities, *de novo* protein design strategies have been developed (Huang et al., 2016a; Marcos & Silva, 2018). Established *de novo* protein design

methods employ an iterative process where (1) the protein shape is outlined and corresponding backbones are sampled, and (2) low energy AA sequences are fitted onto the generated backbones.

Despite numerous successes (Thomson et al., 2014; Huang et al., 2016b; Marcos et al., 2017; Dou et al., 2018), the stochastic nature of sampling methods and inaccuracies within current energy functions frequently drive the design simulations towards incorrect solutions, hence a significant number of design trajectories are needed (2,000–20,000 trajectories) in order to sufficiently probe the sequence and conformational landscape and to select potential low energy solutions (Rocklin et al., 2017; Chevalier et al., 2017; Bonet et al., 2018; Sesterhenn et al., 2020; Yang et al., 2021). In fact, many design failures arise from frustrated backbones that are *a priori* non-"designable". Designable backbones have optimal secondary structure configurations with favored tertiary structure symmetries such that they are physically realizable with the 20 natural AAs (Li et al., 1996; England & Shakhnovich, 2003; Wingreen et al., 2004; Grigoryan & Degrado, 2011).

Quantifying designability is challenging as it includes properties that are difficult to measure, such as fold specificity (Govindarajan & Goldstein, 1996; Wingreen et al., 2004), or native-like structural arrangements (Simons et al., 1997). It has been observed that highly designable backbones can accommodate a large variety of energetically favorable sequences (Govindarajan & Goldstein, 1996; Zhang et al., 2014; Helling et al., 2001). To craft designable backbones, empirical rules have been derived from analysis of protein databases and simulations (Koga et al., 2012), and, together with small structural protein backbone fragment (3-mers and 9-mers) assembly protocols (Rohl et al., 2004), have led to the design of "ideal" protein folds. These rules are based on loop lengths that embed the packing of local tertiary motifs such as β/β-, β/α-, and α/β-units to secondary structure elements (SSEs). Since, the design rules have been steadily updated, for instance, loops can be structurally defined to bridge non-local motifs (Lin et al., 2015; Marcos et al., 2018), cavities can be created by inducing strong curvatures into β-strands through controlling resisters shifts and β-bulges (Huang et al., 2016b; Marcos et al., 2017), and strategically placed residues relieving strain from the backbone allow the design of β-barrels (Dou et al., 2018; Vorobieva et al., 2021).

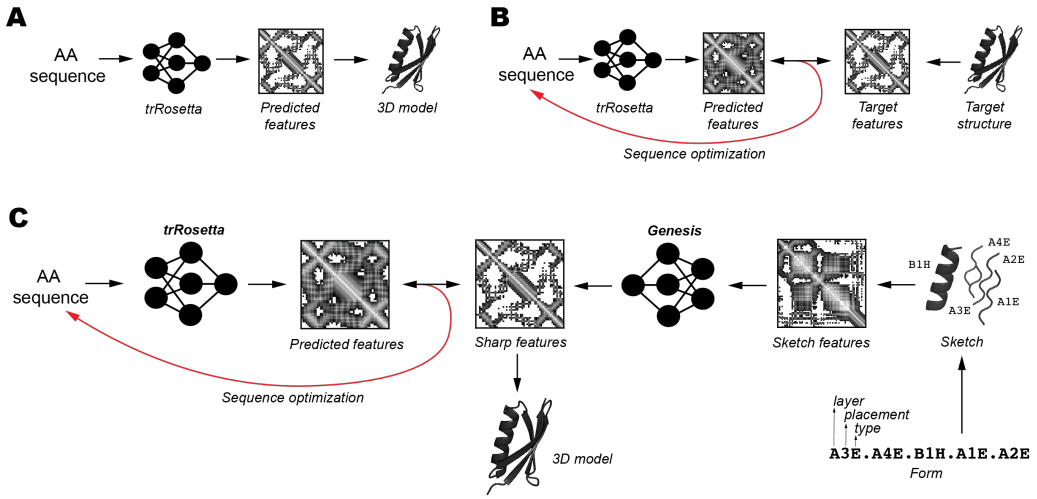

Figure 1: GENESIS neural *de novo* protein design pipeline. A: The trRosetta structure prediction method. B: The trRosetta method used for fixed backbone design maximizing the predicted probabilities towards the given target inter-residue distances and orientations. C: Our developed GENESIS-trRosetta framework for *de novo* protein design. We use the trRosetta design strategy to generate sequences for the refined feature maps produced by the GENESIS neural network from the naively sketched protein folds.

Recent advances within deep neural networks (DNNs) combined with the availability of large-scale protein structure data in the protein data bank (PDB) (Berman et al., 2000) have enabled highly accurate structure prediction from sequence (Yang et al., 2020; Baek et al., 2021; Jumper et al., 2021) (Fig. 1A). Interestingly, the trained structure prediction DNNs can be "reversed" for the protein design task. A good example is the transform-restrained Rosetta (trRosetta) neural network that

was used to hallucinate novel proteins by using a specific loss that maximizes the contrast between random (background) and native distance predictions (Anishchenko et al., 2021). TrRosetta can also be employed for fixed backbone design via backpropagating gradients from the target structure to the sequence, which has the effect of implicitly optimizing over the full sequence and structure landscape (Norn et al., 2021) (Fig. 1B). In the latter case, the method searches for the lowest-energy sequence while maximizing the probability of the target structure relative to all other conformations. Encouragingly, the trRosetta design framework is able to design new sequences for a target structure within minutes on modern graphical processing units (GPUs), enabling multi-state and high-throughput sampling of the design space.

Inspired by these recent advances, this work puts forth the hypothesis that trRosetta can also facilitate tailored *de novo* design, where the shape and secondary structure element composition is controlled (Fig. 1C). To achieve this, we implemented a framework that uses a simple string description of a protein fold (termed "Form") and auto-generates realistic designed proteins. The framework first creates a 3D representation of the Form termed "Sketch". We trained a variational autoencoder (VAE) termed GENESIS to encode the inter-residue distances and orientations of a Sketch to a latent representation, sampled and then decoded close-native distance and orientation probabilities from this representation ready for the trRosetta sequence design task. Our approach circumvents the need to create designable 3D backbones and is, unlike conventional *de novo* design methods, not directly based on energy functions. This allows the *de novo* design process to be fast and efficient biasing the search towards productive sequence spaces.

2 PROTEIN FOLD SKETCHING

We define the overall shape of a protein through a string specifying the SSE types, lengths, and relative positions on a lattice, termed "Form" (Taylor et al., 2008). In a Form, each level or layer of the lattice can be populated by an arbitrary number of SSEs. The layers are equally spaced from each other by 8 Å for β-β layers and 10–11 Å for α-α or helix-β layers (Chothia & Finkelstein, 1990). A Form can be expanded into a 3D representation that we call a "Sketch". A Sketch is a rough 3D approximation of a native protein structure albeit lacking loops, native-like irregularities within SSEs, and AA side chains.

To address these limitations, we employ a DNN to automatically learn to decipher important structural features and incorporate native-like patterns into the Sketches. The DNN takes the form of a VAE that is trained to transform a large dataset of Sketches into their respective native structures. The dataset was built by generating different sets of "Sketches" and mapping them to their native counterparts (Fig. 2A) (Appendix A.3). The Sketches have small idealized SSEs (5 residues for β-strands and 9 residues for α-helices), no sequence information, and dummy backbone residues along the shortest path between end- and starting points of the SSEs representing the loops. The loops were modeled in this way because we do not have information about their potential conformations. The sets encompass many 2- and 3-layer fully β-, fully α- and mixed α/β topologies and capture a large scope of possible protein folds (Appendix A.1). The Sketches do not resemble their native counterparts having a median template modeling (TM)-score of below 0.5 indicating that they are not classified as the same protein fold (Fig. 2B). Importantly, although Sketches can fit onto multiple native structures the majority only map to 1 or 2 conformations (Fig. 2C). Since not all protein domains can be formulated as a Form (e.g., β-barrels), we augment our data set by adding corrupted backbone structures, where the loops are replaced by dummy residues as done on the Sketches (Sup. Fig. 5A). The corrupted backbones add architectural and structural diversity to our data set by retaining tertiary motif dispositions and secondary structure irregularities, respectively (Sup. Fig. 5B).

We split our data set into a series of training and test sets with different structural properties based on the Structural Classification of Proteins — extended (SCOPe) scheme (Andreeva et al., 2020). We then optimize GENESIS for a training set consisting of a particular type of structure and test the validity of its predictions on proteins with increasingly different structural properties (Appendix A.2). Our evaluation procedure quantitatively assesses the extent to which our framework generalizes beyond the distribution induced by a given training set. We argue that any method that facilitates *de novo* design should generalize to unknown subsets ("dark matter") of the protein space.

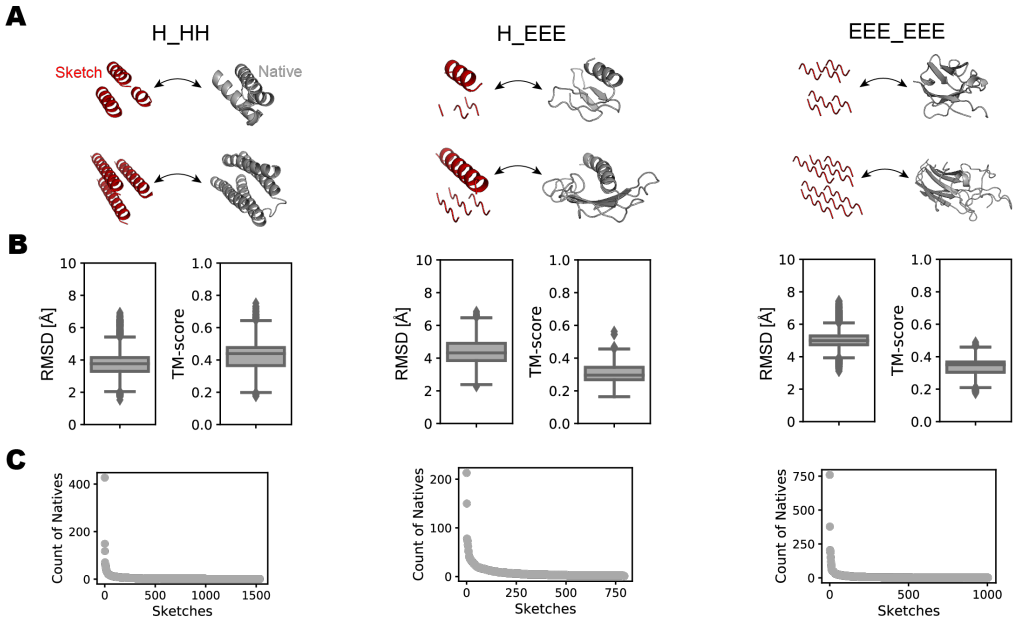

Figure 2: Data generation. A: Examples of Sketches (red) and a corresponding native structure (grey) for the major protein structure classes (HH_H: 3 α-helical bundle, H_EEE: mixed α/β sandwich, EEE_EEE: β sandwich). B: Similarities between the Sketches and their corresponding native structures based on RMSD and the TM-score for major protein structure classes. C: The number of native structures that can be represented by an individual Sketch across the three major protein structure classes.

3 RESULTS

3.1 NEURAL BACKBONE REFINEMENT

We developed a convolutional variational autoencoder (VAE) that operates on inter-residue distance and orientations rather than atomistic coordinates (Fig. 3A) (see Appendix A.4). Importantly, distances and orientations are invariant with respect to translation and rotation which ensures stable and predictable performance in the presence of transformations of the data input under the special Euclidean group. Our VAE is conditioned on the real-valued distances and orientations of the Sketches, and from the latent conditional distribution predicts distance- and orientation probabilities of native-like conformations.

We train the VAE in a supervised manner by minimizing the 1st Wasserstein distance between the true feature maps and the distribution predicted by the VAE. In contrast to the previously utilized cross-entropy loss (Senior et al., 2020; Yang et al., 2020), the Wasserstein distance enables weighting individual errors between the distributions, i.e., penalizing large differences between the true and predicted distributions more than small differences (see Appendix A.5 for details). We follow a standard pre-train—fine-tune regimen (Fig. 3A): (1) We pre-train the VAE on the corrupted backbones with a learning rate set to 1e-3 over 300 epochs. During this phase, the VAE is conditioned on real backbones with loops that have been corrupted with noise. (2) We fine-tune the VAE for 500 epochs on the Sketches. The pre-training slightly improves the performance on the test set when compared to directly training the VAE on the Sketches (Sup. Fig. 6).

We coupled the fine-tuned VAE (called GENESIS) with the trRosetta framework. We use the trRosetta design method to optimize sequences for our generated distance- and orientation probabilities. We subsequently used the generated sequences and constraints from trRosetta to minimize the energy with gradient descent and generate 3D models using PyRosetta (see Appendix A.7). In summary, the GENESIS-trRosetta *de novo* design framework uses a Form to build a Sketch that is

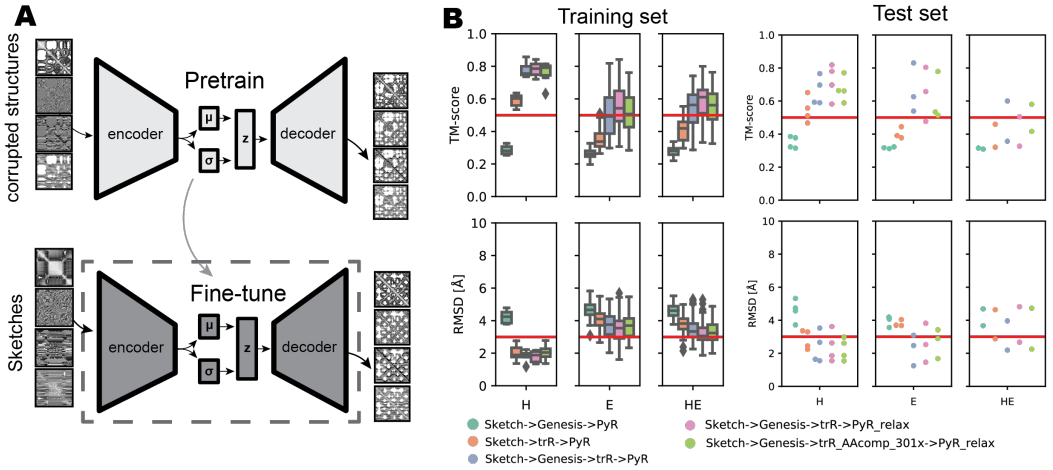

Figure 3: Pipeline and performance. A: General architecture and training scheme of GENESIS. The VAE is first pretrained with corrupted feature maps and subsequently fine-tuned with Sketch feature maps. B: Different pipelines and their performances for the different classes of proteins ("H": fully α-helical, "E": fully β, and "HE": mixed α/β). The number of optimization steps is 101 if not differently indicated. "Sketch" represents the input feature maps from the Sketch, "GENESIS" is the module to optimize the feature maps, "trR" is the trRosetta design for sequence design module and "PyR" is the PyRosetta script to output 3D models from the generated features and sequences. The first pipeline is an ablation of the trRosetta module, where restraints are derived directly from the GENESIS generated feature maps using a poly-Valine AA chain for the 3D model generation. The second pipeline is an ablation of the GENESIS module where the trRosetta module is directly used to optimize the Sketch feature maps. The three subsequent pipelines are variations of the full pipeline, including additional relaxation steps (PyR_relax) and adding an AA composition loss with 301 optimization steps to the trRosetta module (trR_AAcomp_301x).

then refined by GENESIS, designed through trRosetta, and finally assembled and minimized with a full atomistic energy function in PyRosetta.

Our ablation studies showed the importance of both GENESIS and trRosetta modules. First, we removed the trRosetta design module by gathering structural restraints directly from the GENESIS distance- and orientation probabilities and using a poly-Valine AA sequence. We see a low performance with the median TM-score (Zhang & Skolnick, 2005) below 0.5 and the median root-mean-squared-deviations (RMSDs) around 4 Å between the predicted 3D model and the native structure on the training and test set examples across all major classes (Fig. 3B). Second, we removed GENESIS resulting in a framework where trRosetta is challenged to directly design sequences for a given Sketch. On the training examples, we obtained a TM-score median around 0.6 and an RMSD median around 2 Å for the 3-helical bundle architectures, while for the fully-β and mixed α/β architectures the results were rather modest with a median TM-score around 0.4 and the median RMSD around 3.5 Å. The few selected test examples follow the same trend as the train examples: GENESIS alone is not sufficient to build native-like poly-Valine backbones, and simply using trRosetta to design sequences from Sketches resulted in a poor performance with sequences and constraints that do not recapitulate the target fold described by the Sketch. *Thus, our experiments support that, while* GENESIS *cannot solve the backbone design problem by itself, its predicted features can guide sequence generating engines (trRosetta) towards the sequences that are more likely to adopt the specified folds. On the other hand, a naively assembled Sketch lacks native-like features are required by trRosetta to use and design fold-specific sequences.*

We assess the performance of different variations of the GENESIS pipeline. Using the basic framework Sketch $\rightarrow$ GENESIS $\rightarrow$ trRosetta $\rightarrow$ PyRosetta, we achieved a TM-score of 0.8 and a median RMSD of 2 Å for fully α-helical proteins, a median TM-score of 0.55 and median RMSD 3.5 Å for fully-β proteins, and a median TM-score of 0.5 with a median RMSD 4 Å for mixed α/β proteins. We saw an improvement when adding a simple structural relaxation step that favors SSE pairing and

packing after the gradient descent minimization with median TM-scores of approximately 0.8, 0.6, 0.55 and RMSDs of 2 Å, 3 Å, and 3.5 Å for α-, β-, and mixed α/β-proteins from the training set, respectively. We also tested the pipeline using the trRosetta hybrid-design protocol, where, instead of optimizing for a single sequence, the algorithm optimizes for multiple sequences from which a position-specific scoring matrix (PSSM) is generated and used to guide the sequence design task. The results were comparable to the standard pipeline in terms of TM-scores and RMSDs (Sup. Fig. 7).

In order to evaluate the generalization capability of GENESIS, we trained and tested GENESIS on the series of data splits given by SCOPe (see Appendix A.2). The SCOPe database hierarchically classifies proteins based on structural similarities: The top level (Class) divides proteins into major classes: fully-α, fully-β and mixed α/β. The Fold-level arranges the structures according to SSE disposition and connectivity. The Superfamily and Family-levels further classify the structures according fine-grained structural and functional features.

At the Family-, Superfamily- and Fold-level test sets, α-helical proteins reach TM-scores and RMSDs are around 0.6 and 3 Å, respectively, whereas for fully-β and mixed α/β proteins, a degradation in performance is observed (Sup. Fig. 8B). At the Superfamily-level test set, 2 out of 25 test proteins cross the critical 0.5 TM-score threshold for fully-β structures (Sup. Fig. 8B). For mixed α/β proteins, around 3 test proteins are predicted with a TM-score above 0.5 for both Superfamily- and Fold-level test sets (Sup. Fig. 8B). These results indicate that GENESIS shows some ability for generalization across families, while proteins with different connectivities and structural features are more challenging to (re-)generate successfully.

3.2 CONDITIONED SAMPLING AND DESIGN OF NATIVE PROTEIN TOPOLOGIES

To showcase the GENESIS-trRosetta *de novo* design framework, we conditionally sample 5 different topologies. We sample a 2-layer mixed α/β Ubiquitin-like fold, where 4 strands are packed against a helix, and a 3-layer mixed α/β Rossmann fold with a central 4 stranded β-sheet and 2 exterior packing helices on both sides. We additionally challenge the framework by generating 2 different 2-layer β-sandwiches, an Immunoglobulin (Ig)-like fold and a Jelly-roll fold. Finally, we design sequences that adopt the Top7 fold (Kuhlman et al., 2003), a novel fold not observed in the natural repertoire and representing a challenging generalization test for our method.

As we do not have prior knowledge about SSEs and loop lengths, we first sampled 20–30 combinations and generated a small set of sequences for these. We then used AlphaFold (AF) Jumper et al. (2021) to predict a structure for the initial set of sequences and realigned the AF onto the GENESIS models and selected the combinations achieving a TM-score above 0.5. Using the top combinations, we designed up to 10,000 sequences and predict AF models for each of them. We assessed the quality of the designed sequences and models by comparing the TM-score between the GENESIS (Gen) and AF model (TM-score(Gen,AF)) and the AFs' median confidence score (median plDDT) (Fig. 4A). We saw that with a stringent threshold of TM-score(Gen,AF) > 0.6 and median plDDT > 70 around 10% or more of the designed sequences and models were selected as a success (red area in Fig. 4A). These results indicate that the GENESIS-trRosetta framework is able to successfully yield backbone conformations that can harbor sequences with strongly encoded native structural features.

3.3 PROBING THE "DARK MATTER" OF PROTEIN FOLDS

The ultimate goal of *de novo* protein design methods is to generate protein folds non-existent in nature. We asked the question if our framework based on DNNs and trained on natural derived structure data is capable of (besides the Top7 fold) generalizing outside the distribution of natural folds e.g., if our framework is able to generate out-of-distribution. To this end, we sought to sample protein folds not included in the training set, nor observed in nature. Previously, Taylor and colleagues (Taylor et al., 2009) computationally analyzed possibly unexplored regions of the 3-layer mixed α/β fold space through Cα-traces that obey constraints of natural protein structures, such as handedness of connection and loop-crossing. We further reduced the set by discarding Cα-traces that have mixed secondary structure types on the same layer, disembodied SSEs (unpacked) or nearly crossing loops. We selected three distinct folds to design with the GENESIS-trRosetta method.

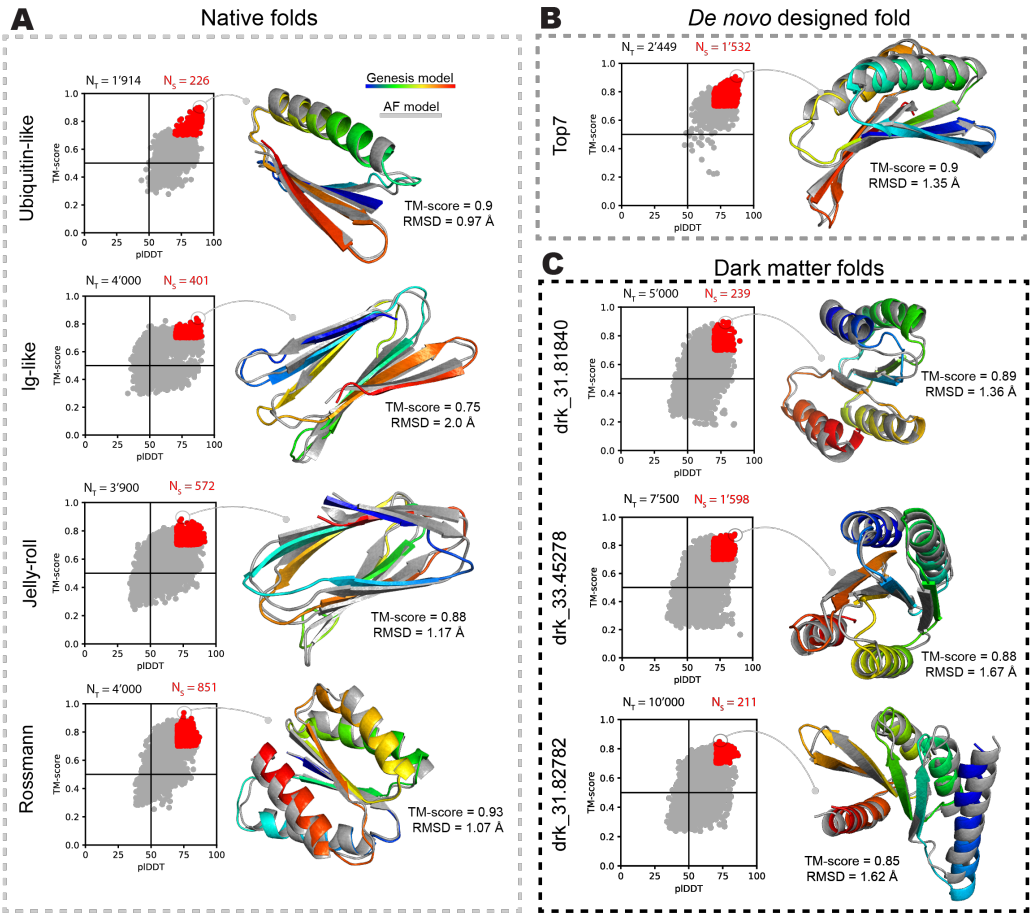

Figure 4: Computational design results. The TM-scores between the GENESIS models and the AlphaFold models versus the median predicted lDDT (plDDT) confidence metric. Each point represents a designed model and the red area represents the selection according to the selection criteria of TM-score > 0.6 and a plDDT > 70. An example for each is given on the right side. A: Native folds. B: A previously *de novo* designed fold. C: Novel folds not existent or discovered in the natural repertoire (dark matter).

We used the GENESIS-trRosetta framework to sample sequences using as input different Forms varying in the loop and SSE lengths and selected the top models (by TM-score between the GENESIS model and the Cα-trace) to up-sample sequences. All sequences were then fed into AF and the predictions aligned to the initial GENESIS model. By assessing the sequence-to-structure qualities of the designed models via the TM-score(Gen,AF) and the AFs' median plDDT (Fig. 4B,C), we observe that the GENESIS-trRosetta framework is capable of resolving blurry feature maps towards native-like distributions even for novel folds (Fig. 4B,C).

4 DISCUSSION

We show that a specialized VAE termed GENESIS is able to encode representations of idealized protein folds and decode native-like conformations. By basing ourselves on distance- and orientation representations, we are able to alleviate the need of generating designable protein backbones in 3D space with fold-specific restraints and energy functions, and thereby also bypassing the need for designable backbones. We couple GENESIS to the trRosetta design engine to generate multiple sequences for the sampled distance- and orientation representations for a set of known and novel folds.

Our results demonstrate that the GENESIS-trRosetta framework is capable of designing new proteins adopting known folds and novel folds non-existent in nature. By changing secondary structure and loop lengths the overall size can be adjusted and different conformations sampled. The generalization capability of GENESIS is significant and can guide the trRosetta-Rosetta hybrid design method to sequences with strong fold signatures. Using AlphaFold as an orthogonal test shows that many of these sequences adopt the intended target shape. Additionally, our framework is considerably fast, within minutes to generate a sequence and a 3D model for a given target protein shape even on a CPU. This demonstrates the usage of DNNs can leverage the automated generation of proteins normally only accessible through large-scale simulations.

Our work opens exciting new horizons for *de novo* protein design where control over the shape is desired. For example, our method could be harvested to generate custom protein backbones such that they fit onto non-canonically structured protein interfaces. Often nanomaterials exhibit highly regular patterns, and could therefore be engaged by secondary structures that are placed respecting the regularity constraints. Another example where our method could be used is for the design of larger molecular assemblies that are constructed from smaller protein domains. Often, the overall shape of the assembly is controlled by the shape of the individual subunits. Hence, we expect that the versatility and speed of the GENESIS-trRosetta method together with other potential deep neural network tools for protein design and engineering to explore the protein universe should be broadly useful.

ACKNOWLEDGMENTS

B.E.C. is a grantee from the European Research Council (Starting grant - 716058), the Swiss National Science Foundation, and the Biltema Foundation. Parts of the computational simulations were performed at the CSCS - Swiss National Supercomputing Centre through a grant obtained by B.E.C.. Z.H. is supported by a grant from the National Center of Competence in Research in Chemical Biology. MB is supported in part by the ERC Consolidator Grant No. 724228 (LEMAN). JS is supported by the UKRI CDT in AI for Healthcare http://ai4health.io (Grant No. P/S023283/1).

REFERENCES

Antonina Andreeva, Eugene Kulesha, Julian Gough, and Alexey G Murzin. The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. *Nucleic Acids Research*, 48(D1):D376–D382, January 2020. ISSN 0305-1048. doi: 10.1093/nar/gkz1064. URL https://doi.org/10.1093/nar/gkz1064.

Ivan Anishchenko, Samuel J. Pellock, Tamuka M. Chidyausiku, Theresa A. Ramelot, Sergey Ovchinnikov, Jingzhou Hao, Khushboo Bafna, Christoffer Norn, Alex Kang, Asim K. Bera, Frank DiMaio, Lauren Carter, Cameron M. Chow, Gaetano T. Montelione, and David Baker. De novo protein design by deep network hallucination. *Nature*, 600(7889):547–552, December 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-04184-w. URL https://www.nature.com/articles/s41586-021-04184-w. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 7889 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Machine learning;Protein design;Protein folding Subject_term_id: machine-learning;protein-design;protein-folding.

Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N. Kinch, R. Dustin Schaeffer, Claudia Millán, Hahnbeom Park, Carson Adams, Caleb R. Glassman, Andy DeGiovanni, Jose H. Pereira, Andria V. Rodrigues, Alberdina A. van Dijk, Ana C. Ebrecht, Diederik J. Opperman, Theo Sagmeister, Christoph Buhlheller, Tea Pavkov-Keller, Manoj K. Rathinaswamy, Udit Dalwadi, Calvin K. Yip, John E. Burke, K. Christopher Garcia, Nick V. Grishin, Paul D. Adams, Randy J. Read, and David Baker. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, August 2021. doi: 10.1126/science.abj8754. URL https://www.science.org/doi/abs/10.1126/science.abj8754. Publisher: American Association for the Advancement of Science.

H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, January 2000. ISSN 0305-1048. doi: 10.1093/nar/28.1.235.

Jaume Bonet, Sarah Wehrle, Karen Schriever, Che Yang, Anne Billet, Fabian Sesterhenn, Andreas Scheck, Freyr Sverrisson, Barbora Veselkova, Sabrina Vollers, Roxanne Lourman, Mélanie Villard, Stéphane Rosset, Thomas Krey, and Bruno E. Correia. Rosetta FunFolDes – A general framework for the computational design of functional proteins. *PLOS Computational Biology*, 14(11):e1006623, November 2018. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1006623. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006623. Publisher: Public Library of Science.

Aaron Chevalier, Daniel-Adriano Silva, Gabriel J. Rocklin, Derrick R. Hicks, Renan Vergara, Patience Murapa, Steffen M. Bernard, Lu Zhang, Kwok-Ho Lam, Guorui Yao, Christopher D. Bahl, Shin-Ichiro Miyashita, Inna Goreshnik, James T. Fuller, Merika T. Koday, Cody M. Jenkins, Tom Colvin, Lauren Carter, Alan Bohn, Cassie M. Bryan, D. Alejandro Fernández-Velasco, Lance Stewart, Min Dong, Xuhui Huang, Rongsheng Jin, Ian A. Wilson, Deborah H. Fuller, and David Baker. Massively parallel de novo protein design for targeted therapeutics. *Nature*, 550(7674):74–79, October 2017. ISSN 1476-4687. doi: 10.1038/nature23912. URL https://www.nature.com/articles/nature23912. Number: 7674 Publisher: Nature Publishing Group.

Cyrus Chothia and Alexei V. Finkelstein. The Classification and Origins of Protein Folding Patterns. *Annual Review of Biochemistry*, 59(1):1007–1035, 1990. doi: 10.1146/annurev.bi.59.070190.005043. URL https://doi.org/10.1146/annurev.bi.59.070190.005043.

Jiayi Dou, Anastassia A. Vorobieva, William Sheffler, Lindsey A. Doyle, Hahnbeom Park, Matthew J. Bick, Binchen Mao, Glenna W. Foight, Min Yen Lee, Lauren A. Gagnon, Lauren Carter, Banumathi Sankaran, Sergey Ovchinnikov, Enrique Marcos, Po-Ssu Huang, Joshua C. Vaughan, Barry L. Stoddard, and David Baker. De novo design of a fluorescence-activating β-barrel. *Nature*, 561(7724):485–491, September 2018. ISSN 1476-4687. doi: 10.1038/s41586-018-0509-0. URL https://www.nature.com/articles/s41586-018-0509-0. Number: 7724 Publisher: Nature Publishing Group.

Jeremy L. England and Eugene I. Shakhnovich. Structural determinant of protein designability. *Physical Review Letters*, 90(21):218101, May 2003. ISSN 0031-9007. doi: 10.1103/PhysRevLett.90.218101.

Naomi K. Fox, Steven E. Brenner, and John-Marc Chandonia. SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Research*, 42(D1):D304–D309, January 2014. ISSN 0305-1048. doi: 10.1093/nar/gkt1240. URL https://doi.org/10.1093/nar/gkt1240.

S. Govindarajan and R. A. Goldstein. Why are some proteins structures so common? *Proceedings of the National Academy USA*, 93(8):3341–3345, April 1996. ISSN 0027-8424. doi: 10.1073/pnas.93.8.3341.

Alastair Grant, David Lee, and Christine Orengo. Progress towards mapping the universe of protein folds. *Genome Biology*, 5(5):107, April 2004. ISSN 1474-760X. doi: 10.1186/gb-2004-5-5-107. URL https://doi.org/10.1186/gb-2004-5-5-107.

Gevorg Grigoryan and William F. Degrado. Probing designability via a generalized model of helical bundle geometry. *Journal of Molecular Biology*, 405(4):1079–1100, January 2011. ISSN 1089-8638. doi: 10.1016/j.jmb.2010.08.058.

R. Helling, H. Li, R. Mélin, J. Miller, N. Wingreen, C. Zeng, and C. Tang. The designability of protein structures. *Journal of Molecular Graphics & Modelling*, 19(1):157–167, 2001. ISSN 1093-3263. doi: 10.1016/s1093-3263(00)00137-6.

Po-Ssu Huang, Scott E. Boyken, and David Baker. The coming of age of de novo protein design. *Nature*, 537(7620):320–327, September 2016a. ISSN 1476-4687. doi: 10.1038/nature19946. URL https://www.nature.com/articles/nature19946. Number: 7620 Publisher: Nature Publishing Group.

Po-Ssu Huang, Kaspar Feldmeier, Fabio Parmeggiani, D. Alejandro Fernandez Velasco, Birte Höcker, and David Baker. De novo design of a four-fold symmetric TIM-barrel protein with

atomic-level accuracy. *Nature Chemical Biology*, 12(1):29–34, January 2016b. ISSN 1552-4469. doi: 10.1038/nchembio.1966. URL https://www.nature.com/articles/nchembio.1966. Number: 1 Publisher: Nature Publishing Group.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, August 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2. URL https://www.nature.com/articles/s41586-021-03819-2. Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 7873 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Computational biophysics;Machine learning;Protein structure predictions;Structural biology Subject_term_id: computational-biophysics;machine-learning;protein-structure-predictions;structural-biology.

Wolfgang Kabsch and Christian Sander. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, 1983. ISSN 1097-0282. doi: 10.1002/bip.360221211. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/bip.360221211.

Nobuyasu Koga, Rie Tatsumi-Koga, Gaohua Liu, Rong Xiao, Thomas B. Acton, Gaetano T. Montelione, and David Baker. Principles for designing ideal protein structures. *Nature*, 491(7423):222–227, November 2012. ISSN 1476-4687. doi: 10.1038/nature11600. URL https://www.nature.com/articles/nature11600. Number: 7423 Publisher: Nature Publishing Group.

Rachel Kolodny, Leonid Pereyaslavets, Abraham O. Samson, and Michael Levitt. On the universe of protein folds. *Annual Review of Biophysics*, 42:559–582, 2013. ISSN 1936-1238. doi: 10.1146/annurev-biophys-083012-130432.

Brian Kuhlman, Gautam Dantas, Gregory C. Ireton, Gabriele Varani, Barry L. Stoddard, and David Baker. Design of a Novel Globular Protein Fold with Atomic-Level Accuracy. *Science*, 302(5649):1364–1368, November 2003. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1089427. URL https://science.sciencemag.org/content/302/5649/1364. Publisher: American Association for the Advancement of Science Section: Research Article.

Hao Li, Robert Helling, Chao Tang, and Ned Wingreen. Emergence of Preferred Structures in a Simple Model of Protein Folding. *Science*, 273(5275):666–669, August 1996. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.273.5275.666. URL https://science.sciencemag.org/content/273/5275/666.

Yu-Ru Lin, Nobuyasu Koga, Rie Tatsumi-Koga, Gaohua Liu, Amanda F. Clouser, Gaetano T. Montelione, and David Baker. Control over overall shape and size in de novo designed proteins. *Proceedings of the National Academy of Sciences USA*, 112(40):E5478–E5485, October 2015. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1509508112. URL https://www.pnas.org/content/112/40/E5478. ISBN: 9781509508112 Publisher: National Academy of Sciences Section: PNAS Plus.

Enrique Marcos and Daniel-Adriano Silva. Essentials of de novo protein design: Methods and applications. *WIREs Computational Molecular Science*, 8(6):e1374, 2018. ISSN 1759-0884. doi: 10.1002/wcms.1374. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/wcms.1374. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wcms.1374.

Enrique Marcos, Benjamin Basanta, Tamuka M. Chidyausiku, Yuefeng Tang, Gustav Oberdorfer, Gaohua Liu, G. V. T. Swapna, Rongjin Guan, Daniel-Adriano Silva, Jiayi Dou, Jose Henrique Pereira, Rong Xiao, Banumathi Sankaran, Peter H. Zwart, Gaetano T. Montelione, and David Baker. Principles for designing proteins with cavities formed by curved beta-sheets. *Science*,

January 2017. doi: 10.1126/science.aah7389. URL https://www.science.org/doi/abs/10.1126/science.aah7389. Publisher: American Association for the Advancement of Science.

Enrique Marcos, Tamuka M. Chidyausiku, Andrew C. McShan, Thomas Evangelidis, Santrupti Nerli, Lauren Carter, Lucas G. Nivón, Audrey Davis, Gustav Oberdorfer, Konstantinos Tripsianes, Nikolaos G. Sgourakis, and David Baker. De novo design of a non-local β-sheet protein with high stability and accuracy. *Nature Structural & Molecular Biology*, 25(11):1028–1034, November 2018. ISSN 1545-9985. doi: 10.1038/s41594-018-0141-6. URL https://www.nature.com/articles/s41594-018-0141-6. Number: 11 Publisher: Nature Publishing Group.

Christoffer Norn, Basile I. M. Wicky, David Juergens, Sirui Liu, David Kim, Doug Tischer, Brian Koepnick, Ivan Anishchenko, Foldit Players, David Baker, and Sergey Ovchinnikov. Protein sequence design by conformational landscape optimization. *Proceedings of the National Academy of Sciences USA*, 118(11), March 2021. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2017228118. URL https://www.pnas.org/content/118/11/e2017228118. Publisher: National Academy of Sciences Section: Physical Sciences.

Victor M. Panaretos and Yoav Zemel. Statistical Aspects of Wasserstein Distances. *Annual Review of Statistics and Its Application*, 6(1):405–431, March 2019. ISSN 2326-8298, 2326-831X. doi: 10.1146/annurev-statistics-030718-104938. URL http://arxiv.org/abs/1806.05500. arXiv: 1806.05500.

Gabriel J. Rocklin, Tamuka M. Chidyausiku, Inna Goreshnik, Alex Ford, Scott Houliston, Alexander Lemak, Lauren Carter, Rashmi Ravichandran, Vikram K. Mulligan, Aaron Chevalier, Cheryl H. Arrowsmith, and David Baker. Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science*, 357(6347):168–175, July 2017. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aan0693. URL https://science.sciencemag.org/content/357/6347/168. Publisher: American Association for the Advancement of Science Section: Research Article.

Carol A. Rohl, Charlie E. M. Strauss, Kira M. S. Misura, and David Baker. Protein structure prediction using Rosetta. *Methods in Enzymology*, 383:66–93, 2004. ISSN 0076-6879. doi: 10.1016/S0076-6879(04)83004-0.

Andrew W. Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander W. R. Nelson, Alex Bridgland, Hugo Penedones, Stig Petersen, Karen Simonyan, Steve Crossan, Pushmeet Kohli, David T. Jones, David Silver, Koray Kavukcuoglu, and Demis Hassabis. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710, January 2020. ISSN 1476-4687. doi: 10.1038/s41586-019-1923-7. URL https://www.nature.com/articles/s41586-019-1923-7. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 7792 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Machine learning;Protein structure predictions Subject_term_id: machine-learning;protein-structure-predictions.

Fabian Sesterhenn, Che Yang, Jaume Bonet, Johannes T. Cramer, Xiaolin Wen, Yimeng Wang, Chi-I. Chiang, Luciano A. Abriata, Iga Kucharska, Giacomo Castoro, Sabrina S. Vollers, Marie Galloux, Elie Dheilly, Stéphane Rosset, Patricia Corthésy, Sandrine Georgeon, Mélanie Villard, Charles-Adrien Richard, Delphyne Descamps, Teresa Delgado, Elisa Oricchio, Marie-Anne Rameix-Welti, Vicente Más, Sean Ervin, Jean-François Eléouët, Sabine Riffault, John T. Bates, Jean-Philippe Julien, Yuxing Li, Theodore Jardetzky, Thomas Krey, and Bruno E. Correia. De novo protein design enables the precise induction of RSV-neutralizing antibodies. *Science (New York, N.Y.)*, 368(6492):eaay5051, May 2020. ISSN 1095-9203. doi: 10.1126/science.aay5051.

K. T. Simons, C. Kooperberg, E. Huang, and D. Baker. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *Journal of Molecular Biology*, 268(1):209–225, April 1997. ISSN 0022-2836. doi: 10.1006/jmbi.1997.0959.

William R. Taylor, Gail J. Bartlett, Vijayalakshmi Chelliah, Daniel Klose, Kuang Lin, Tom Sheldon, and Inge Jonassen. Prediction of protein structure from ideal forms. *Proteins: Structure, Function, and Bioinformatics*, 70(4):1610–1619, 2008. ISSN 1097-0134. doi: https://doi.org/10.1002/prot.21913. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.21913. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.21913.

William R. Taylor, Vijayalakshmi Chelliah, Siv Midtun Hollup, James T. MacDonald, and Inge Jonassen. Probing the "Dark Matter" of Protein Fold Space. *Structure*, 17(9):1244–1252, September 2009. ISSN 0969-2126. doi: 10.1016/j.str.2009.07.012. URL https://www.cell.com/structure/abstract/S0969-2126(09)00295-0. Publisher: Elsevier.

Andrew R. Thomson, Christopher W. Wood, Antony J. Burton, Gail J. Bartlett, Richard B. Sessions, R. Leo Brady, and Derek N. Woolfson. Computational design of water-soluble α-helical barrels. *Science*, 346(6208):485–488, October 2014. ISSN 1095-9203. doi: 10.1126/science.1257452.

Anastassia A. Vorobieva, Paul White, Binyong Liang, Jim E. Horne, Asim K. Bera, Cameron M. Chow, Stacey Gerben, Sinduja Marx, Alex Kang, Alyssa Q. Stiving, Sophie R. Harvey, Dagan C. Marx, G. Nasir Khan, Karen G. Fleming, Vicki H. Wysocki, David J. Brockwell, Lukas K. Tamm, Sheena E. Radford, and David Baker. De novo design of transmembrane beta-barrels. *Science*, February 2021. doi: 10.1126/science.abc8182. URL https://www.science.org/doi/abs/10.1126/science.abc8182. Publisher: American Association for the Advancement of Science.

Ned S. Wingreen, Hao Li, and Chao Tang. Designability and thermal stability of protein structures. *Polymer*, 45(2):699–705, January 2004. ISSN 0032-3861. doi: 10.1016/j.polymer.2003.10.062. URL https://www.sciencedirect.com/science/article/pii/S0032386103009911.

Che Yang, Fabian Sesterhenn, Jaume Bonet, Eva A. van Aalen, Leo Scheller, Luciano A. Abriata, Johannes T. Cramer, Xiaolin Wen, Stéphane Rosset, Sandrine Georgeon, Theodore Jardetzky, Thomas Krey, Martin Fussenegger, Maarten Merkx, and Bruno E. Correia. Bottom-up de novo design of functional proteins with complex structural features. *Nature Chemical Biology*, 17(4):492–500, April 2021. ISSN 1552-4469. doi: 10.1038/s41589-020-00699-x. URL https://www.nature.com/articles/s41589-020-00699-x. Number: 4 Publisher: Nature Publishing Group.

Jianyi Yang, Ivan Anishchenko, Hahnbeom Park, Zhenling Peng, Sergey Ovchinnikov, and David Baker. Improved protein structure prediction using predicted interresidue orientations. *Proceedings of the National Academy of Sciences USA*, 117(3):1496–1503, January 2020. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1914677117. URL https://www.pnas.org/content/117/3/1496. Publisher: National Academy of Sciences Section: Biological Sciences.

Jian Zhang, Fan Zheng, and Gevorg Grigoryan. Design and designability of protein-based assemblies. *Current Opinion in Structural Biology*, 27:79–86, August 2014. ISSN 0959-440X. doi: 10.1016/j.sbi.2014.05.009. URL https://www.sciencedirect.com/science/article/pii/S0959440X14000633.

Yang Zhang and Jeffrey Skolnick. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, 33(7):2302–2309, 2005. ISSN 1362-4962. doi: 10.1093/nar/gki524.

Jianfu Zhou and Gevorg Grigoryan. Rapid search for tertiary fragments reveals protein sequence–structure relationships. *Protein Science*, 24(4):508–524, 2015. ISSN 1469-896X. doi: https://doi.org/10.1002/pro.2610. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/pro.2610. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pro.2610.

Jianfu Zhou and Gevorg Grigoryan. A C++ library for protein sub-structure search. *bioRxiv*, pp. 2020.04.26.062612, April 2020. doi: 10.1101/2020.04.26.062612. URL https://www.biorxiv.org/content/10.1101/2020.04.26.062612v1. Publisher: Cold Spring Harbor Laboratory Section: New Results.

A APPENDIX

A.1 DATASET GENERATION

We created two distinct datasets from the scope (v2.07 stable) Fox et al. (2014) domains of medium sizes (40–128 residues).

1. The pre-training data set was created by corrupting existing protein structures by removing the loops based on the DSSP (hydrogen bond estimation algorithm) Kabsch & Sander (1983) assignments. We remodel the loops as done in a Sketch, e.g. we add dummy residues (Nitrogen (N), Carbon (C), Carbon (Cα), Oxygen (O) backbone atoms with randomized torsion angles) along the shortest path between the two endpoints Cα atoms of the consecutive SSEs. We add as many dummy residues as in the native structure, hence the corrupted structure has the same length as its native counterpart. This procedure leaves the native SSE dispositions that may incorporate important native structural features for the pre-training. In total, we created 40,726 pairs.
2. We developed a program that creates small fold Sketches obeying simple topological rules such as non-crossing loops and loop distance restraints from the architecture types: EE_EEE, EEE_EEE, H_EEE, H_EEEE, H_EEE_H, HH_EE, HH_EEE, HH_EE_H, HHH, HHH_EE (where "_" represents a layer separation and E: β-strand and H: α-helix). We searched the SCOPe domains for partial structural matches within 3 Å RMSD using MASTER Zhou & Grigoryan (2015; 2020) for each of the generated Sketches. Importantly, a Sketch can partially match onto a native domain. We crop the overlapping regions of the native domain at the first and the last residue of the matching Sketch. SSEs within the cropped domain that do not map to SSEs in the Sketch are assigned as loops. Furthermore, we remove domains larger than 128 residues and identical matches for the mapping to the same Sketch. This resulted in a total of 35,435 Sketch - native domain pairs.

A.2 DATA SPLITS

Within SCOPe, protein structures are hierarchically classified into groups where the "Class" groups proteins based on secondary structure content and organization (fully-α, fully-β, mixed-α/β), "Fold" divides them based on SSE disposition and connectivity, "Superfamily" is based on structural features and "Family" contains the structures with similar sequences.

We pick protein families that represent compact structures with small loops for our family test set. The test set includes the SCOPe families b.1.22.1, b.11.1.6, b.69.2.3, b.70.2.1, b.82.1.22, b.114.1.1, a.7.2.0, a.7.2.1, a.7.8.2, a.7.12.1, a.8.11.1, a.24.10.3, a.24.13.1, a.60.9.0, a.160.1.2, c.2.1.7, c.25.1.2, c.118.1.0, c.93.1.0, c.56.5.6, d.110.4.3, e.51.1.1, c.97.1.5, d.17.1.5, d.58.3.2, d.58.10.0, d.58.23.1, d.92.1.13, d.230.1.1, d.240.1.0. We generate higher-level test sets (Fold and Superfamily) by removing all corresponding groups from the picked structures in the Family test set, e.g. for the Family b.1.22.1 the Superfamily is b.1.22 and the Fold is b.1.. Importantly, identical structures and Sketches in the training set were removed in order to avoid any biases during testing.

A.3 DATA ENCODING

The coordinates of the Sketches and their native counterparts are encoded into a total of four 2D distance- and orientation feature maps as done by trRosetta. Briefly, the first feature map is all-against-all Cβ distances. The second feature map is the dihedral "ω" that measures the rotation along the virtual axis of two connecting Cβ residues. The distances and ω angles are symmetric, e.g. measuring from residue i to residue j will give the same result as measuring from residue j to residue i. The third and fourth feature map are the "θ" dihedrals and the "ϕ" angles specifying the direction of Cβ of residue i with respect to residue j. Both, θ and ϕ are asymmetric metrics. Together the four feature maps fully define a protein backbone in 3D space.

While we use real valued feature maps as input to GENESIS, we bin the true feature maps according to the trRosetta scheme. The distances from 2 to 20 Å are binned into 36 equally spaced segments (0.5 Å each) and a 37$^{\text{th}}$ bin to indicate that pairs are not in contact. The dihedral (ω, θ) and angular (ϕ) features are binned into 15° segments yielding 24, 24, and 12 with an additional bin indicating

no contact, respectively. Therefore, we have encoded the true feature maps into tensors of shape 128x128x1x37 for the distances, 128x128x1x25 for the dihedrals and 128x128x1x13 for the angles. Thus, at each "pixel" (each residue pair) we have an additional dimension that can be seen as a Dirac distribution with a score of one for the bin with the distance and zero everywhere else.

A.4 GENESIS ARCHITECTURE

The VAE includes an encoder, a decoder, and a loss function. The input Sketchs' feature maps (real-valued) of shapes 128x128x4 are processed by the encoder that is a sequence of four convolutional blocks. A single block includes a 2D convolution, an instance norm and ELU activation followed by a 40% dropout. From the compressed data representation, two multilayer perceptrons (MLPs) are used to predict predict means and covariances vectors (of sizes 128). Using the reparametrization trick, a latent variable z from a normally distributed $p(z|x)$ is sampled. The decoder $q(y|z)$ passes z through three blocks of 2D transposed convolution, instance norm, ELU activation and 40% dropout to create a decompressed representation. The final layer of the decoder branches into four different output heads. Each head is a convolutional block with a final softmax activation over each pixel yielding distance outputs of shape 128x128x1x37, two dihedral outputs of sizes 128x128x1x25 and an angular output of shape 128x128x1x13.

A.5 LOSS FUNCTION

Our loss function is composed of five individual losses (four reconstruction losses, and a loss on the latent space).

We use the Wasserstein distance (for details see Panaretos & Zemel (2019)) as reconstruction loss. Let us define x P and y Q and the corresponding densities as p and q, respectively. We assume that $(x, y) \in \mathbb{R}^d$. Additionally, let us denote $\mathcal{J}(P, Q)$ all joint distributions $\mathcal{J}$ for (x, y) that have marginals P and Q. Then the general Wasserstein distance can be written as

$$W_p(\mathcal{P}, \mathcal{Q}) = \left(\inf_{J \in \mathcal{J}(\mathcal{P}, \mathcal{Q})} \int \|x - y\|^p \, dJ(x, y) \right)^{1/p} \tag{1}$$

In the discrete case, when P and Q are distributions $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$ the formulation becomes

$$W_p(\mathcal{P}, \mathcal{Q}) = \left(\sum_{i=1}^{n} \|x_i - y_i\|^p \right)^{1/p} \tag{2}$$

In the case of 1D discrete distributions $(p = 1)$, the 1-Wasserstein (W_1) distance is also called Earth mover's distance (EMD) and is efficiently computable. The main advantage of the 1-Wasserstein distance compared to other measures such as the binary cross-entropy and the Kullback-Leibler (KL) divergence is that it takes into account the metric space. This means that larger deviations from the predicted to the true distributions are more penalized while small errors are less penalized.

We define the reconstruction loss as the sum over the 1-Wasserstein distances between the predicted distributions $(\hat{D})$ and the true distributions (D) of each pixel normalized by the length of the protein (N_{AA}). Each pixel is defined as (i, j) where $i = 1, \ldots, n_w$ and $j = 1, \ldots, n_h$ with n_w being the width and n_h the height.

$$L_{\text{rec}} = \frac{1}{N_{\text{AA}}} \sum_{i=1}^{n_w} \sum_{j=1}^{n_h} W_1(D_{i,j}, \hat{D}_{i,j}) \tag{3}$$

Note that the true distribution is modeled as a Dirac distribution supported by the true values, whereas the predicted distribution $(\hat{D})$ is parametrized by the VAE decoder. We additionally use the Kullback-Leibler (KL) divergence on the latent space normalized by the length of the protein to penalize latent vectors not following a Normal distribution KLD $= \frac{1}{N_{\text{AA}}} \text{KL}(p(z|x) \| p(z))$, with $p(z)$ Normal$(0, 1)$. Thus the final loss is defined as

$$L_{\text{tot}} = L^{\text{d}}_{\text{rec}} + L^{\omega}_{\text{rec}} + L^{\theta}_{\text{rec}} + L^{\phi}_{\text{rec}} + \frac{1}{N_{\text{AA}}} \text{KL}(p(z|x) \| p(z)). \tag{4}$$

A.6 TRAINING REGIMEN

We use a batch size of 64 Sketches, the Adam optimizer with a starting learning rate of 0.001 for the pre-training and fine-tuning. We reduce the learning rate with a step of 0.97 at each epoch during pre-training and every second epoch during fine-tuning.

A.7 MODELING AND DESIGN

The trRosetta design framework is utilized to design a set of 1,000 sequences matching GENESIS refined maps. A position-specific scoring matrix (PSSM) is generated from the library of sequences and used within the PyRosetta protocol. In a first stage, the PyRosetta protocol first generates a coarse-grained model using gradient descent with the optimized restraints and a single sequence from trRosetta design. In a second stage we remove the restraints assuming that the generated coarse grain model has adopted the target shape. We further optimize the coarse grain structure with a full atom protocol. We use the Rosetta FastDesign with layer and PSSM sequence constraints during the design task and topological secondary structure energy bonuses during the relaxation task. In this way, the full atom protocol improves the quality of the final sequence and structure model.

A.8 SUPPLEMENTARY FIGURES

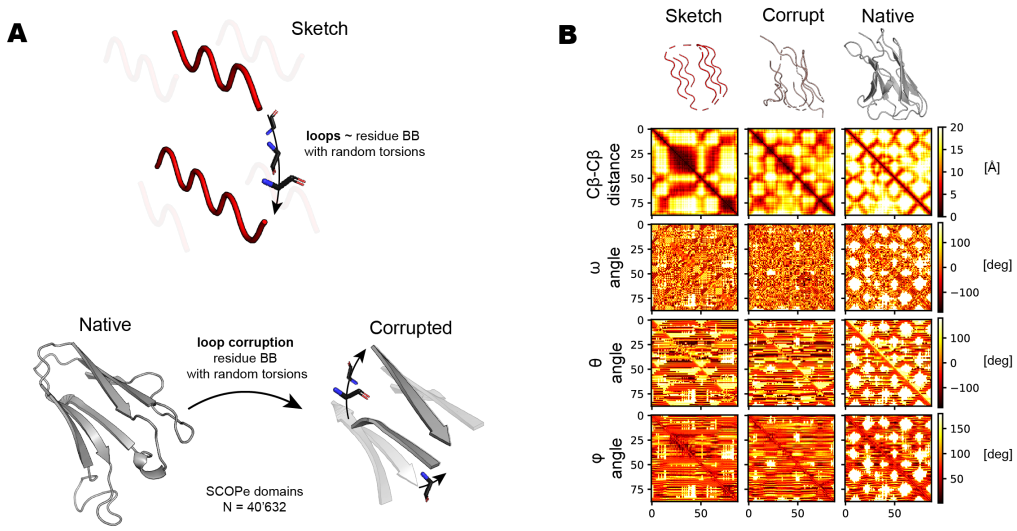

Figure 5: Data engineering. A: Loops on the Sketch and corrupted structure are approximated by adding backbone residue atoms with random backbone dihedral angles along the shortest path between two consecutive SSEs. B: Comparison of the different feature maps (Sketch, corrupted Structures, and native Structure).

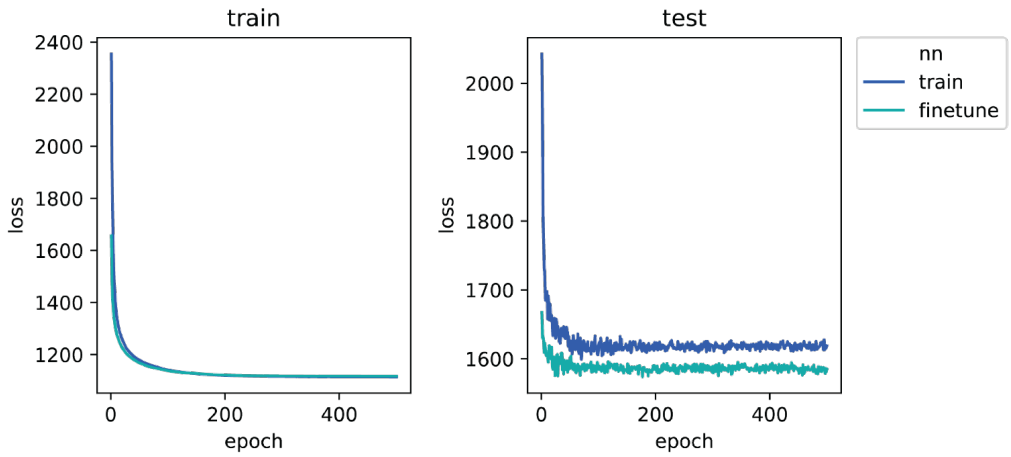

Figure 6: Comparison of the loss function between training only on the Sketches (train) and pre-training on the corrupted Structures followed by fine-tuning on Sketches (fine-tune).

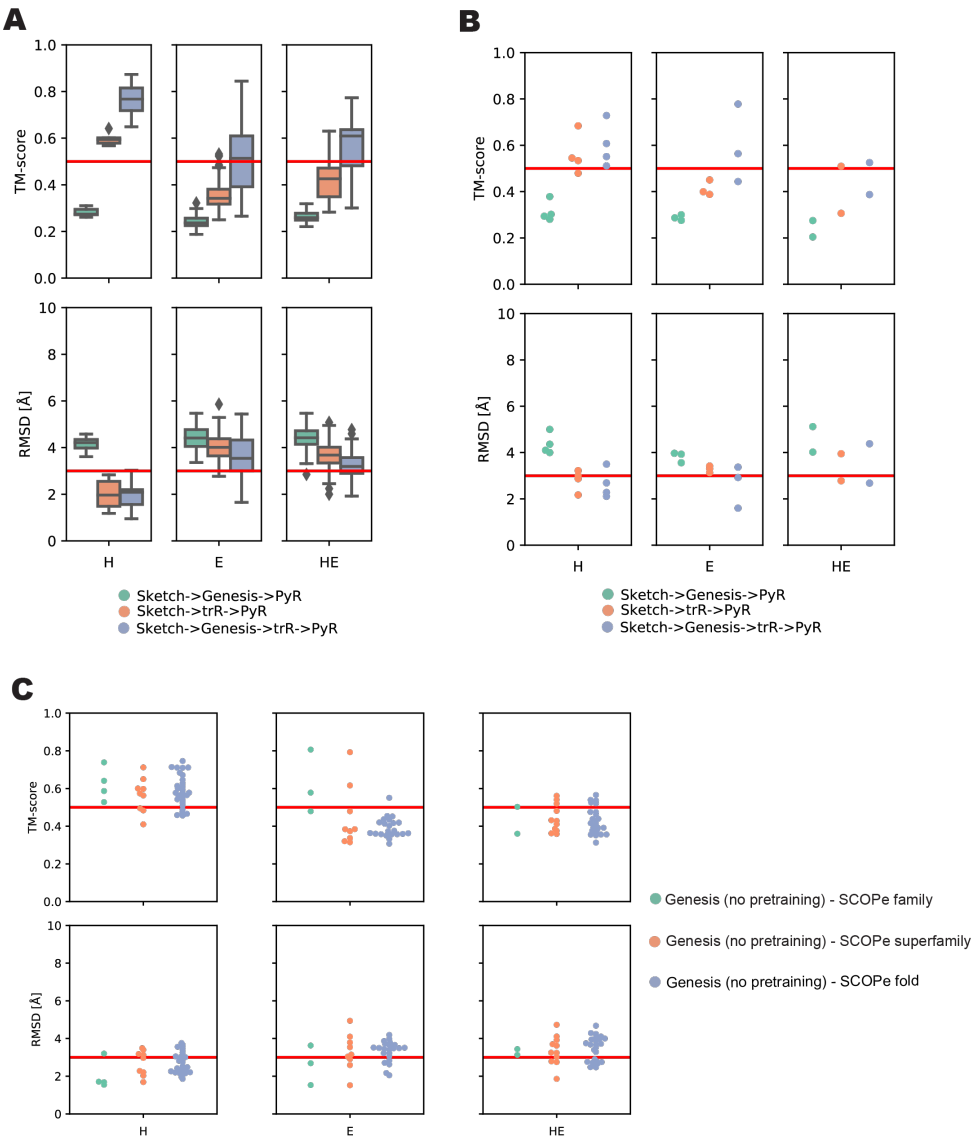

Figure 7: GENESIS-trRosetta with PSSM design. A: Training set: different pipelines and their performances for the different classes of proteins ("H": fully α helical, "E": fully β, and "HE": mixed α/β) using the hybrid trRosetta design approach. B: Test set performances over different protein classes using the hybrid trRosetta design approach. C: Performance of the GENESIS pipeline using the hybrid trRosetta design across different difficulty levels according to the SCOPe structure classification

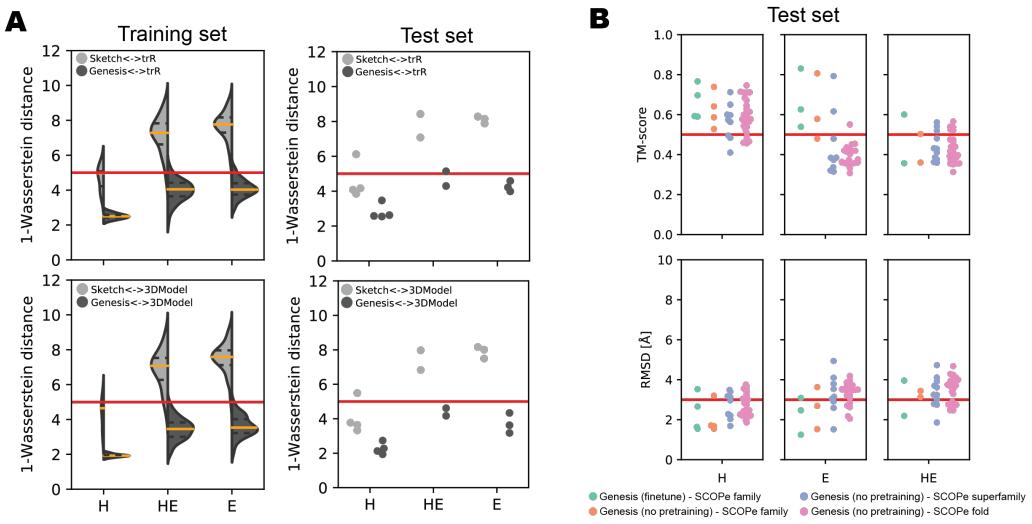

Figure 8: A: Comparison between the Sketch maps - trRosetta (trR) / 3D model (3DModel) maps and the GENESIS refined maps - trRosetta (trR) / 3D model (3DModel) using the first Wasserstein distance metric. B: Performance of the standard GENESIS pipeline across different difficulty levels according to the SCOPe structure classification.