HHS Public Access

Author manuscript

Nat Biotechnol. Author manuscript; available in PMC 2023 August 20.

Published in final edited form as:

Nat Biotechnol. 2022 November ; 40(11): 1617–1623. doi:10.1038/s41587-022-01432-w.

Single-sequence protein structure prediction using language models and deep learning

Ratul Chowdhury[a,*], **Nazim Bouatta**[a,*], **Surojit Biswas**[b,c,*], **Charlotte Rochereau**[d], **Christina Floristean**[e], **Gustaf Ahdritz**[f], **Joanna Zhang**[e], **George M. Church**[a,b], **Peter K. Sorger**[a,g,†], **Mohammed AlQuraishi**[a,f,†]

[a]Laboratory of Systems Pharmacology, Program in Therapeutic Science, Harvard Medical School, Boston, MA, USA.

[b]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA

[c]Nabla Bio. Inc.

[d]Integrated Program in Cellular, Molecular, and Biomedical Studies, Columbia University, New York, NY, USA.

[e]Department of Computer Science, Columbia University, New York, NY, USA.

[f]Department of Systems Biology, Columbia University, New York, NY, USA.

[g]Department of Systems Biology, Harvard Medical School, Boston, MA, USA.

Abstract

AlphaFold2 and related computational systems predict protein structure using deep learning and co-evolutionary relationships encoded in multiple sequence alignments (MSAs). Despite dramatic increases in prediction accuracy achieved by these systems, challenges remain in: (i) prediction of orphan and rapidly evolving proteins for which an MSA cannot be generated, (ii) rapid exploration of designed structures, and (iii) understanding the rules governing spontaneous polypeptide folding in solution. Here we report development of an end-to-end differentiable recurrent geometric network (RGN) that uses a protein language model (AminoBERT) to learn latent structural information from unaligned proteins. A linked geometric module compactly represents C_α backbone geometry. On average RGN2 outperforms AlphaFold2 and RoseTTAFold on orphan proteins and classes of designed proteins while achieving up to a 10^6-fold reduction

[†]*Co-corresponding authors*, Address correspondence to: Mohammed AlQuraishi, Columbia University Irving Medical Center, New York Presbyterian Hospital Building, PH-18, 201G, 622 West 168th St. New York, NY 10032, ma4129@cumc.columbia.edu.
[*]These authors contributed equally

Author Contributions: R.C., N.B., S.B., and M.A. conceived of and designed the study. R.C. developed the refinement module and performed all analyses. N.B. developed the geometry module and trained RGN2 models. S.B. developed and trained the AminoBERT protein language model and helped integrate its embeddings within RGN2. C.R. trained several RGN2 models and performed RF predictions. C.F prepared the docker image and helped package the standalone software along with a Python-based user interface (notebook) for generating RGN2 predictions. G.A. performed MSAs to identify orphans. J.Z. helped C.F. in preparation of the RGN2 prediction notebook. P.K.S. and G.M.C. supervised the research and provided funding. N.B., S.B., M.A. and P.K.S wrote the manuscript and all authors discussed the results and edited the final version.

Competing interests: M.A. is a member of the SAB of FL2021–002, a Foresite Labs company, and consults for Interline Therapeutics. P.K.S. is a member of the SAB or Board of Directors of Glencoe Software, Applied Biomath, RareCyte and NanoString and an advisor to Merck and Montai Health. A full list of G.M.C.'s tech transfer, advisory roles, 559 and funding sources can be found on the lab's website: http://arep.med.harvard.edu/gmc/tech.html. SB is employed by and holds equity in Nabla Bio, Inc.

in compute time. These findings demonstrate the practical and theoretical strengths of protein language models relative to MSAs in structure prediction.

INTRODUCTION

Predicting 3D protein structure from amino acid sequence is a grand challenge in biophysics of practical and theoretical importance. Progress has long relied on physics-based methods that estimate energy landscapes and dynamically fold proteins within these landscapes[1–4]. A decade ago, the focus shifted to extracting residue-residue contacts from co-evolutionary relationships embedded in multiple sequence alignments (MSAs)[5] (Supplementary Figure 1). Algorithms such as the first AlphaFold[6] and trRosetta[7] use deep neural networks to generate distograms able to guide classic physics-based folding engines. These algorithms perform substantially better than algorithms based on physical energy models alone. More recently, the superior performance of AlphaFold2[8] in folding a wide range of protein targets that were part of the recent CASP14 prediction challenge shows that when MSAs are available, machine learning (ML)-based methods can predict protein structure with sufficient accuracy to complement X-ray crystallography, cryoEM, and NMR as a practical means to determine structures of interest.

Predicting the structures of single sequences using ML nonetheless remains a challenge: the requirement in AlphaFold2 for co-evolutionary information from MSAs makes it less performative with proteins that lack sequence homologs, currently estimated at ~20% of all metagenomic protein sequences[9] and ~11% of eukaryotic and viral proteins[10]. Protein design and studies quantifying the effects of sequence variation on function[11] or immunogenicity[12] also require single-sequence structure prediction. More fundamentally, the physical process of polypeptide folding in solution is driven solely by the chemical properties of that chain and its interaction with solvent (excluding, for the moment, proteins that require folding co-factors). An algorithm that predicts structure directly from a single sequence is—like energy-based folding engines[1–4]—closer to the real physical process than an algorithm that uses MSAs. We speculate that ML algorithms able to fold proteins from single sequences will ultimately provide new understanding of protein biophysics.

Structure prediction algorithms that are fast and low-cost are of great practical value because they make efficient exploration of sequence space possible, particularly in design applications[13]. State of the art MSA-based predictions for large numbers of long proteins can incur substantial costs when performed on the cloud. Reducing this cost would enable many practical applications in enzymology, therapeutics and chemical engineering including designing new functions[14–16], raising thermostability[17], altering pH sensitivity[18], and increasing compatibility with organic solvents[19]. Efficient and accurate structure prediction is also valuable in the case of orphan proteins, many of which are thought to play a role in taxonomically restricted and lineage-specific adaptations. OSP24, for example, is an orphan virulence factor for the wheat pathogen *F. graminearum* that controls host immunity by regulating proteasomal degradation of a conserved signal transduction kinase[20]. It is one of many orphan genes found in fungi, plants insects and other organisms[21] for which MSAs are not available.

We have previously described an end-to-end differentiable, ML-based recurrent geometric network (hereafter RGN1)[22] that predicts protein structure from position-specific scoring matrices (PSSMs) derived from MSAs; related end-to-end approaches have since been reported[23–25]. RGN1 PSSM-structure relationships are parameterized as torsion angles between adjacent residues making it possible to sequentially position the protein backbone in 3D space (backbone geometry comprises the arrangement of N, C_α, and C' atoms for each amino acid). All RGN1 components are differentiable and the system can therefore be optimized from end to end to minimize prediction error (as measured by distance-based root mean squared deviation; dRMSD). While RGN1 does not rely on the co-evolutionary information used to generate MSAs, a requirement for PSSMs necessitates that multiple homologous sequences be available.

Here, we describe a new end-to-end differentiable system, RGN2 (Figure 1), that predicts protein structure from single protein sequences by using a protein language model (AminoBERT). Language models were first developed as a means to extract semantic information from a sequence of words (a key requirement for natural language processing; NLP) [26]. In the context of proteins, AminoBERT aims to capture the latent information in a string of amino acids that implicitly specifies protein structure. RGN2 also makes use of a natural way of describing polypeptide geometry that is rotationally- and translationally-invariant at the level of the polypeptide as a whole. This involves using the Frenet-Serret formulas to embed a reference frame at each C_α carbon; the backbone is then easily constructed by a series of transformations. In this paper we describe the implementation and training of AminoBERT, the use of Frenet-Serret formulas in RGN2, and a performance assessment for natural and designed proteins with no significant sequence homologs. We find that on average the GDT_TS achieved by RGN2 is higher than AlphaFold2 (AF2)[8] and RoseTTAFold (RF)[27] even though AF2/RF can achieve higher absolute GDT_TS scores than RGN2 on naturally occurring orphan proteins without known homologs and *de novo* designed proteins. While RGN2 is not as performant as MSA-based methods for proteins that permit use of MSAs, RGN2 is up to six orders of magnitude faster, enabling efficient exploration of sequence and structure landscapes.

RESULTS

RGN2 and AminoBERT models

RGN2 involves two primary innovations relative to RGN1 and other ML-based structure prediction approaches. First, it uses amino acid sequence itself as the primary input as opposed to a PSSM, making it possible to predict structure from a single sequence. In the absence of a PSSM or MSA, latent information on the relationship between protein sequence (as a whole) and 3D structure is captured using a protein language model we term AminoBERT. Second, rather than describe the geometry of protein backbones as a sequence of torsion angles, RGN2 uses a simpler and more powerful approach based on the Frenet-Serret formulas; these formulas describe motion along a curve using the reference frame of the curve itself. This approach to protein geometry is inherently translationally- and rotationally-invariant, a key property of polypeptides in solution. We refined structures predicted by RGN2 using a Rosetta-based protocol[28] that imputes the backbone and

side-chain atoms. Refinement is first performed in torsion space to optimize side-chain conformations and eliminate clashes and then in Cartesian space using quasi-Newton-based energy minimization. These refinement steps are non-differentiable but improve the quality of predicted structures.

Language models were originally developed for natural language processing and operate on a simple but powerful principle: they acquire linguistic understanding by learning to fill in missing words in a sentence, akin to a sentence completion task in standardized tests. By performing this task across large text corpora, language models develop powerful reasoning capabilities. The Bidirectional Encoder Representations from Transformers (BERT) model[29] instantiated this principle using Transformers, a class of neural networks in which attention is the primary component of the learning system[30]. In a Transformer, each token in the input sentence can "attend" to all other tokens through the exchange of activation patterns corresponding to the intermediate outputs of neurons in the neural network. In AminoBERT we utilize the same approach, substituting protein sequences for sentences and using amino acid residues as tokens.

To generate the AminoBERT language model we trained a 12-layer Transformer using ~250 million natural protein sequences obtained from the UniParc sequence database[31]. To enhance the capture of information in full protein sequences we introduced two training objectives not part of BERT or previously reported protein language models[26,32–36]. First, 2–8 contiguous residues were masked simultaneously in each sequence (similar to the ProtTrans[37] language model) making the reconstruction task harder and emphasizing learning from global rather than local context. Second, *chunk permutation* was used to swap contiguous protein segments; chunk permutations preserve local sequence information but disrupt global coherence. Training AminoBERT to identify these permutations is another way of encouraging the Transformer to discover information from the protein sequence as whole. The AminoBERT module of RGN2 is trained independently of the geometry module in a self-supervised manner without fine-tuning (see Methods for details).

In RGN2 we parameterized backbone geometry using the discrete version of the Frenet-Serret formulas for one-dimensional curves[38]. In this parameterization, each residue is represented by its C_α atom and an oriented reference frame centered on that atom. Local residue geometry was described by a single rotation matrix relating the preceding frame to the current one, which is the geometrical object that RGN2 predicts at each residue position. This rotationally- and translationally-invariant parameterization has two advantages over our previous use of torsion angles in RGN1. First, it ensured that specifying a single biophysical parameter, namely the sequential $C_\alpha - C_\alpha$ distance of ~3.8Å (which corresponds to a trans conformation) results in only physically-realizable local geometries. This overcomes a limitation of RGN1, which yielded chemically unrealistic values for some torsion angles. Second, it reduced by ~10-fold the computational cost of chain extension calculations, which often dominates RGN training and inference times (see Methods).

RGN2 training was performed using both the ProteinNet12 dataset[39] and a smaller dataset comprised solely of single protein domains derived from the ASTRAL SCOPe dataset

(v1.75)[40]. Since we observed no detectable difference between the two, all results in this paper derive from the smaller dataset as it required less training time.

Predicting structures of proteins with no homologs

To assess how well RGN2 predicts the structures of orphan proteins having no known sequence homologs (Supplementary Figures 2 and 3), we compared it to AlphaFold2 (AF2)[8], and RoseTTAFold (RF)[27], currently the best publicly available methods. In addition to UniRef30, we used two other complementary databases (PDB70 and MGnify) to prepare a list of 77 proteins with the following properties: (i) they are at least 20 residues long (ii) they are orphans (*i.e.,* MSA depth = 1) across all three datasets simultaneously and (iii) they have solved structures in the Protein Databank (PDB)[41] (see Methods for orphan test set construction details). We note that more than 85% of these sequences are included in the training sets for AF2 and RF, which may result in an overestimate of the accuracy of these methods. We predicted the structures of orphan proteins using all methods and assessed accuracy with respect to experimentally-determined structures (Figure 2A) using dRMSD and GDT_TS (the global distance test, which roughly captures the fraction of the structure that is correctly predicted). We found that RGN2 outperformed AF2 and RF on both metrics in 46% and 52% of cases, respectively (these correspond to the top-left quadrant in Figure 2B). In 35% and 32% of cases, AF2 and RF outperformed RGN2 on both metrics, respectively; split results were obtained in the remaining cases. When we computed differences in error metrics obtained for different prediction methods, we found that RGN2 outperformed AF2 and RF by an average ΔdRMSD of 0.83Å and 1.16Å and ΔGDT_TS of 4.75 and 4.91 units, respectively. When the same analysis was applied only to structures that had been determined by X-ray crystallography (i.e., 80% of the targets shown in light gray in Figure 2B) RGN2 exhibited a similar improvement over AF2 and RF: an average ΔdRMSD of 0.81Å and 1.07Å and ΔGDT_TS of 4.32 and 4.79 units, respectively.

To investigate the structural basis for these differences in performance we applied the DSSP algorithm[42] to determine the fraction of each secondary structure element (helical – alpha, 5 and 3/10, beta-strand, and bridge, and unstructured loops, bends, and hydrogen bonded turns) in PDB structures for the orphan protein test set (Figure 3A). We found that RGN2 outperformed all other methods on proteins rich in single helices and bends or hydrogen-bonded turns interspersed with helices, while other methods—AF2 in particular—better predicted targets with high fractions of beta-strand and beta-bridges (such as hairpins). Performance on the remaining ~19% targets was split between RGN2 and competing methods (Figure 2B). We also examined performance as a function of protein length and found that RGN2 generally outperformed AF2 on longer helical proteins. One possible explanation for these findings is that the Frenet-Serret geometry used by RGN2 is based on two local parameters (curvature and torsion) and these parameters have fixed values for helices. Thus, RGN2 has an intrinsic ability to learn helical patterns.

In Figure 3B–D we show examples of structures for which RGN2 outperformed AF2. For example, PDB structures 5FKP and 2KWZ (97% and 73% alpha helical, respectively) have a polypeptide bend (Figure 3B) and a short alpha helix (Figure 3C) held in place by hydrogen-bonded turns, respectively. RGN2 correctly predicts the challenging, less-

structured bends and turns in these proteins, yielding 4.2 and 3.4-point gains in GDT_TS and ΔdRMSD > 1.1Å over AF2, respectively. A longer protein 6E5N have an alpha-helical bundle connected by bends (Figure 3D). AF2 accurately predicted the majority of helical domains in 6E5N, but had dihedral errors in the hydrogen-bonded turns and bends; in contrast, RGN2 correctly predicted these unstructured polypeptide stretches between helices. This contributed to an 8.7-point increase in GDT_TS and 1.62Å decrease in dRMSD, respectively.

Predicting the structures of de novo (designed) proteins

We evaluated the accuracy of RGN2 on a test set of 149 synthetic proteins that were originally designed *de novo* using computationally parametrized energy functions such as Rosetta and Amber; these proteins are expected to be well-suited to prediction by RosettaFold (RF). Many of these proteins are intended to have applications in therapeutic development such as novel antimicrobial peptides. This test set comprises all known designed proteins that are not part of the AF2 training set, as ascertained by PDB deposition date and filtered to have an *organism* annotation of *synthetic construct*. This filter helps to eliminate ambiguous *de novo* protein entries (*e.g.,* 7NBI) which are synthesized single point mutants of known proteins. As before, we assessed prediction accuracy using dRMSD and GDT_TS. We found that RGN2 outperformed AF2 and RF on both metrics in 44% and 43% of cases, respectively (Figure 4). On average, RGN2 outperformed AF2 and RF on these targets with dRMSD and GDT_TS gains of 0.46Å and 3.62, and of 0.71Å and 4.14, respectively (Figure 4). The same analysis applied only to structures determined by X-ray crystallography (i.e., 66% of the targets shown in light gray in Figure 4B) yield similar improvements in RGN2 relative to AF2 and RF: an average ΔdRMSD of 0.42Å and 0.65Å and ΔGDT_TS of 3.58 and 4.07 units, respectively. We conclude that RGN2 can better predict sequence-structure relationships for helical regions of *de novo* protein space than all competing methods (Figure 5) but that beta sheet prediction from single sequences remains a challenge.

As an illustration of how RGN2 improves on and complements AF2 predictions, we show in Figure 5B the structure of an antimicrobial peptide (PDB Accession Code: 2L96). This is a largely helical protein comprising two short helices connected by a hydrogen-bonded turn and an N-terminal loop. Similar to orphans with similar secondary structural composition, RGN2 predicts this target more accurately than AF2 (ΔdRMSD = –0.90Å and GDT_TS = +3.31 ΔGDT_TS). In Figure 5C and 5D we show predicted structures of two different alpha-beta targets, both connected by hydrogen-bonded turns (PDB Accession Codes: 5UP5 and 7KBQ). For these targets both RGN2 and AF2 are equally accurate in capturing the ordered secondary structured elements (TM-Scores with PDB > 0.7). RGN2 is only marginally more accurate than AF2 on a global level (ΔdRMSD < −1.5Å), but hydrogen-bonded turns are better recapitulated by RGN2 and consequently result in a higher ΔGDT_TS score. Similar observations suggest that future hybrid methods using both a language model and MSAs may outperform either method alone.

Contact prediction precision

We performed a comparative contact prediction analysis between RGN2 and ESM-1b first on our newly revised set of 124 *de novo* protein targets (i.e., those > 20 aa long with no homolog across PDB70, MGnify, and UniRef90 datasets; Table 1a)) and our set of designed proteins (Table 1b). These tables show the percentage precision of top L/2, L/5, and L/10 contacts. We note that ESM-1b outperforms RGN2 on the beta-rich contacts, but for alpha rich contacts, RGN2 remains marginally ahead. We note that gains in contact prediction accuracy do not necessarily translate to improved tertiary structure prediction[43].

RGN2 prediction speed

Rapid prediction of protein structure is essential for tasks such as protein design and analysis of allelic variation or disease mutations. By virtue of being end-to-end differentiable, RGN2 predicts unrefined structures using fast neural network operations and does not require physics-based conformational sampling to assemble a folded chain. Because it operates directly on single sequences, RGN2 also avoids expensive MSA calculations. To quantify these benefits, we compared the speed of RGN2 and other methods on orphan and *de novo* proteins datasets of varying lengths (breaking down computation time by prediction stage; Table 2). In MSA-based methods, MSA generation scaled linearly with MSA depth (*i.e.,* the number of homologous sequences used) whereas distogram prediction (by trRosetta) scaled quadratically with protein length. AF2 predictions scale cubically with protein length. In contrast, RGN2 scales linearly with protein length and both template-free and MSA-free implementations of AF2 and RF were $>10^5$ fold slower than RGN2. In the absence of post-prediction refinement, RGN2 is up to 10^6-fold faster, even for relatively short proteins. Adding physics-based refinement increased compute cost for all methods, but even so RGN2 remains the fastest available method. Of interest, even when MSA generation is discounted, neural network-based inference for AF2 and RF remains much slower than RGN2, inclusive of post-prediction refinement. This gap will only widen for design tasks involving longer proteins, whose chemical synthesis is increasingly becoming feasible[44]. Thus, fast prediction is an important benefit of using a protein language model such as AminoBERT.

DISCUSSION

RGN2 represents one of the first attempts to use ML to predict protein structure from a single sequence. This is computationally efficient and has many advantages in the case of orphan and designed proteins for which generation of multiple sequence alignment is often not possible. RGN2 accomplishes this by fusing a protein language model (AminoBERT) with a simple and intuitive approach to parameterizing C_α backbone geometry based on the Frenet-Serret formulas. Whereas most recent advances in ML-based structure prediction have relied on MSAs[5] to learn latent information about folding, AminoBERT learns this information from proteins without alignment. Training in this case involves sequences with masked residues and block permutations. We speculate that the latent space of the language model also captures recurrent evolutionary relationships[45]. The use of Frenet-Serret formulas in RGN2 addresses the requirement that proteins exhibit translational and rotational invariance. From a practical standpoint, the speed and accuracy of RGN2 shows that language models are effective at learning structural information from primary

sequence while having the ability to extrapolate beyond known proteins, thereby enabling effective prediction of orphan and designed proteins. Nonetheless, methods that utilize MSA information (when it is available) often outperform RGN2, most notably AlphaFold2 when assessed on proteins in the “Free Modeling” category of CASP14 (Supplementary Figure 4). Thus, language models are not a substitute for MSAs but rather a complementary way to get at the latent rules governing protein folding. We speculate that folding systems that use both language models and MSAs will be more performative than systems using one approach alone.

Transformers and their embodiment of local and distant attention is a key feature of language models such as AminoBERT. Very large Transformer-based models trained on hundreds of millions and potentially billions of protein sequences are increasingly available[26,35,36] and the scaling previously observed in natural language applications[46] makes it likely that the performance of RGN2 and similar methods will continue to improve and become broadly performative over intrinsically disordered proteins and cyclic peptides as well (Supplementary Figure 5). AlphaFold2 also exploits attention mechanisms based on Transformers to capture the latent information in MSAs. Similarly, the self-supervised MSA Transformer[47] uses a related attention strategy that attends to both positions and sequences in an MSA, and achieves state-of-the-art contact prediction accuracy. Architectures merging language models and MSAs are also likely to benefit from augmentation from high-confidence structures found in the newly reported AlphaFold Database.[8]. Finally, training on experimental data is almost certain to be invaluable in selected applications requiring high accuracy within members of multi-protein families, such as predicting structural variation within kinases or G protein-coupled receptors.

We consider RGN2 to be a first step in the development of methods able to compute sequence-to-structure maps without a requirement for explicit evolutionary information. One limitation of the current RGN2 implementation to be addressed by future systems is that the immediate output of the recurrent geometric network only constrains local dependencies between C_α atoms (curvature and torsion angles) resulting in sequential reconstruction of backbone geometry. Allowing the network to reason directly on arbitrary pairwise dependencies throughout the structure, and using a better inductive prior than immediate contact may further improve the quality of model predictions. A second limitation is that refinement in RGN2 is not part of an end-to-end implementation; refinement via a 3D rotationally- and translationally-equivariant neural network would be more efficient and likely yield better quality structures. Currently, Rosetta-based refinements results in 28% higher GDT_TS and 16% lower dRMSD values, on average, relative to predictions from RGN2 alone, as evaluated using all 213 orphan and *de novo* targets described in this study (Supplementary Figure 6).

It has been known since Anfinsen’s refolding experiments that single polypeptide chains contain the information needed to specify fold[48]. The demonstration that a language model can learn information on structure directly from protein sequences and then guide accurate prediction of an unaligned protein suggests that RGN2 behaves in a manner that is more similar to the physical process of protein folding than MSA-based methods. Transformers can learn structural encodings present in both local and distant features of a sequence, which

is reflective of the role played by local residues in the molten globule stage and distant residues in the 3D protein fold. Moreover, language models learned by deep neural networks are readily formulated in a maximum entropy framework.[49] The physical process of protein folding is also entropically driven, potentially suggesting a means to compare the two. A fusion of biophysical and learning-based perspectives may ultimately prove the key to direct sequence-to-structure prediction from single polypeptides at experimental accuracy and for understanding folding energetics and dynamics.

METHODS

AminoBERT summary

AminoBERT is a 12-layer Transformer where each layer is composed of 12 attention heads. It is trained to distill protein sequence semantics from ~260 million natural protein sequences obtained from the UniParc sequence database[31] (downloaded May 19, 2019).

During training each sequence is fed to AminoBERT according to the following algorithm:

1. With probability 0.3 select sequence for chunk permutation, and with probability 0.7 select sequence for masked language modeling.
2. If sequence was selected for chunk permutation, then:

   With probability 0.35 chunk permute, else (with probability 0.65) leave the sequence unmodified.
3. Else if the sequence was selected for masked language modeling, then:

With probability 0.3 introduce 0.15 × sequence_length masks into the sequence with clumping, else (with probability 0.7) introduce the same number of masks into the sequence randomly across the length of the sequence (standard masked language modeling).

The loss for an individual sequence (seq) is given by:

$$Loss(seq) = I[seq\ is\ chunk\ permuted] \times chunk\_permutation\_loss(seq) + (1 - I[seq\ is\ globally\ perturbed] \times masked\_lm\_loss(seq)$$

where I[x] is the indicator of the event x, and returns 1 if x is true, and 0 if x is false. *Chunk_permutation_loss(seq)* is a standard cross entropy loss reflecting the classification accuracy of predicting whether seq has been chunk permuted. Finally, *masked_lm_loss(seq)* is the standard masked language modeling loss as previously described in Devlin *et al.*[29]. Note, that mask clumping does not affect how the loss is calculated.

Chunk permutation is performed by first sampling an integer x uniformly between 2 and 10, inclusively. The sequence is then randomly split into x equal-sized fragments, which are subsequently shuffled and rejoined.

Mask clumping is performed as follows:

1. Sample an integer *clump_size* ~ Poisson (2.5) + 1

2. Let *n_mask* = 0.15 × *sequence_length*. Randomly select *n_mask*/*clump_size* positions in the sequence around which to introduce a set of *clump_size* contiguous masks

AminoBERT architecture

Each multi-headed attention layer in AminoBERT contains 12 attention heads, each with hidden size 768. The output dimension of the feed-forward unit at the end of each attention layer is 3072. As done in BERT[29], we prepend a [CLS] token at the beginning of each sequence, for which an encoding is maintained through all layers of the AminoBERT Transformer. Each sequence was padded or otherwise clipped to length 1024 (including the [CLS] token).

For chunk permutation classification, the final hidden vector of the [CLS] token is fed through another feed forward layer of output dimension 768, followed by a final feed forward layer of output dimension 2, which are the logits corresponding to whether the sequence is chunk permuted or not. Masked language modeling loss calculations are set up as described in Devlin *et al.* [29].

AminoBERT training procedure

AminoBERT was trained with batch size 3072 for 1,100,000 steps, which is approximately 13 epochs over the 260 million sequence corpus. For our optimizer we used Adam with a learning rate of 1e-4, $\beta 1 = 0.9$, $\beta 2 = 0.999$, epsilon=1e-6, L2 weight decay of 0.01, learning rate warmup over the first 20,000 steps, and linear decay of the learning rate. We used a dropout probability of 0.1 on all layers, and used GELU activations as done for BERT. Training was performed on a 512 core TPU pod for approximately one week.

Geometry module

The geometry of the protein backbone as summarized by the C_α trace can be thought of as a one-dimensional discrete open curve, characterized by a bond and torsion angle at each residue. Following *Niemi et al.*[38], the starting point for describing such discrete curves is to assign a frame, a triplet of orthonormal vectors, to each C_α atom. If we denote by r_i the vector characterizing the position of a C_α atom at the i-th vertex, we could then define a unit tangent vector along an edge connecting two consecutives C_α atoms

$$t_i = \frac{r_{i+1} - r_i}{|r_{i+1} - r_i|}$$

For assigning frames to each i-th C_α atom, we need two extra vectors, the binormal and normal vectors defined as follows:

$$b_i = \frac{t_{i-1} \times t_i}{|t_{i-1} \times t_i|}$$

$$n_i = b_i \times t_i$$

While for a protein (in a given orientation) the tangent vector is uniquely defined, the normal and binormal vectors are arbitrary. Indeed, when assigning frames to each residue we could take any arbitrary orthogonal basis on the normal plane to the tangent vector. Such arbitrariness does not affect our strategy of predicting 3D structures starting from bond and torsion angles.

To derive the equivalent of the Frenet-Serret formulas—which describe the geometry of continuous and differentiable one-dimensional curves—for the discrete case, we need to relate two consecutive frames along the protein backbone in terms of rotation matrices

$$\begin{pmatrix} n_{i+1} \\ b_{i+1} \\ t_{i+1} \end{pmatrix} = \mathscr{R}_{i+1,i} \begin{pmatrix} n_i \\ b_i \\ t_i \end{pmatrix}$$

In three-dimensions, rotation matrices are in general parametrized in term of three Euler angles. However, in our case the rotation matrices relating two consecutive frames are fully characterized by only two angles, a bond angle ψ and a torsion angle θ, as the third Euler angle vanishes, reflecting the following condition $b_{i+1} \cdot t_i = 0$. We can now write the equivalent of the Frenet-Serret formulas for the discrete case

$$\begin{pmatrix} n_{i+1} \\ b_{i+1} \\ t_{i+1} \end{pmatrix} = \begin{pmatrix} \cos\psi\cos\theta & \cos\psi\sin\theta & -\sin\psi \\ -\sin\theta & \cos\theta & 0 \\ \sin\psi\cos\theta & \sin\psi\sin\theta & \cos\psi \end{pmatrix} \begin{pmatrix} n_i \\ b_i \\ t_i \end{pmatrix}$$

The bond and torsion angles are defined by the following relations

$$\cos\ \psi_{i+1,i} = t_{i+1} \cdot t_i$$

$$\cos\ \theta_{i+1,i} = b_{i+1} \cdot b_i$$

We now turn to backbone reconstruction starting from bond and torsion angles. First, using tangent vectors along the backbone edges, we can reconstruct all C_α atom positions, and thus the full protein backbone in the C_α trace, by using the following relation:

$$r_k = \sum_{i=0}^{k-1} |r_{i+1} - r_i| \cdot t_i$$

where $|r_{i+1} - r_i|$ is the length of the virtual bonds connecting two consecutive C_α atoms. In most cases, the average virtual bond length is ~ 3.8 Å, which corresponds to trans conformations. In terms of the familiar torsion angles ϕ, ψ, and ω, those conformations are achieved for $\omega \sim \pi$. For cis conformations, mainly involving proline residues, the virtual bond length is ~ 3.0 Å (and it corresponds to $\omega \sim 0$). In RGN2, for backbone reconstruction, we impose the condition that the virtual bond length is strictly equal to 3.8 Å, and for reconstructing the backbone we use the following relation:

$$r_k = \sum_{i=0}^{k-1} 3.8 \times t_i$$

The intuition behind the previous equation is the idea of a moving observer along the protein backbone. We could think of the tangent vector t_i as the velocity of the observer along a given edge, and the constant virtual bond length as the effective time spent for travelling along the edge. The only freedom allowed for such observer is to abruptly change the direction of the velocity vector at each vertex.

The model outputs bond and torsion angles. By centering the first C_α atom of the protein backbone at the origin of our coordinate system, we sequentially reconstruct all the C_α atom coordinates using the following relation:

$$\begin{pmatrix} n_{i+1} \\ b_{i+1} \\ t_{i+1} \\ r_{i+1} \end{pmatrix} = \begin{pmatrix} & & & 0 \\ & \mathscr{R}_{i+1,i} & & 0 \\ & & & 0 \\ 0 & 0 & 3.8 & 1 \end{pmatrix} \begin{pmatrix} n_i \\ b_i \\ t_i \\ r_i \end{pmatrix}$$

Data preparation for comparison with trRosetta

Performance of RGN2 was compared against trRosetta across two sets of non-homologous proteins: (i) 129 orphans from the Uniclust30 database[50], and (ii) 35 *de novo* proteins by Xu *et al.*[51]. Both sets were filtered to ensure no overlap with the training sets of RGN2 and trRosetta. While RGN2 is trained on the ASTRAL SCOPe (v1.75) dataset[40], trRosetta was trained on a set of 15,051 single chain proteins (released before May 1, 2018).

Structure prediction with trRosetta, AF2, and RF

Conventional trRosetta-based structure prediction involves first feeding the input sequence through a deep MSA generation step. For orphans and *de novo* proteins without any sequence homologs, the MSA only includes the original query sequence. Next, the MSA is used by the trRosetta neural network to predict a distogram (and orientogram) that captures inter-residue (C_α-C_α and C_β-C_β) distances and orientations. This information is subsequently utilized by a final Rosetta-based refinement module. This module first threads a naïve sequence of polyalanines of length equaling the target protein that maximally obeys the distance and orientation constraints. After side-chain imputation that reflects the original sequence, multiple steps including clash elimination, rotamer repacking, and energy minimization are performed to identify the lowest energy structure.

AF2 and RF predictions did not require MSAs since our target proteins don't have homologs and so we made our predictions using their respective official Google Colab notebooks.

Structure refinement in RGN2

Raw predictions from RGN2 contain a single C_α trace of the target protein. After performing a local internal coordinate building step to generate the backbone and side-chain atoms corresponding to the target sequence, we use Rosetta-based refinement to finetune the structure. This refinement comprises hybrid optimization of side chains using

five invocations of energy minimization in torsional space followed by a single step of quasi-Newton all-atom minimization in Cartesian space (using the *FastRelax* protocol of RosettaScripts[52]). An optional *CartesianSampler*[52] mover step can be added to further correct local strain density in the predicted model. The six-step *FastRelax* protocol is repeated for 300 cycles for each target. Finally, 100 cycles of coarse-grained, fast minimization using *MinMover*[52] is applied to obtain the predicted structure.

RGN2 is available freely as a standalone tool from https://drive.google.com/file/d/1FIU6UZrhmc44YVCMLAzCHXWMHkSpoCkt/view?usp=sharing. Users can make structure predictions using a Python-based web user interface by uploading the protein sequence in fasta format.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements:

We gratefully acknowledge the support of the NVIDIA Corporation for the donation of GPUs used for this research. This work is supported by the DARPA PANACEA program grant HR0011-19-2-0022 and NCI grant U54-CA225088 to PKS. We also acknowledge support from the TensorFlow Research Cloud (TFRC) for graciously providing the TPU resources used for training AminoBERT.

REFERENCES

1. Yang J. & Zhang Y. I-TASSER server: New development for protein structure and function predictions. Nucleic Acids Res. (2015) doi:10.1093/nar/gkv342.
2. Wang J, Wang W, Kollman PA & Case DA Automatic atom type and bond type perception in molecular mechanical calculations. J. Mol. Graph. Model. 25, 247–260 (2006). [PubMed: 16458552]
3. Hess B, Kutzner C, Van Der Spoel D. & Lindahl E. GRGMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. J. Chem. Theory Comput. 4, 435–447 (2008). [PubMed: 26620784]
4. Alford RF et al. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. J. Chem. Theory Comput. (2017) doi:10.1021/acs.jctc.7b00125.
5. AlQuraishi M. Machine learning in protein structure prediction | Elsevier Enhanced Reader. 65, 1–8 (2021).
6. Senior AW et al. Improved protein structure prediction using potentials from deep learning. | Nature | 577, (1923).
7. Yang J. et al. Improved protein structure prediction using predicted interresidue orientations. Proc. Natl. Acad. Sci. U. S. A. 117, 1496–1503 (2020). [PubMed: 31896580]
8. Jumper J. et al. Highly accurate protein structure prediction with AlphaFold. Nat. 2021 1–11 (2021) doi:10.1038/s41586-021-03819-2.
9. Pearson WR An introduction to sequence similarity ('homology') searching. Curr. Protoc. Bioinforma. (2013) doi:10.1002/0471250953.bi0301s42.
10. Perdigão N. et al. Unexpected features of the dark proteome. Proc. Natl. Acad. Sci. U. S. A. (2015) doi:10.1073/pnas.1508380112.
11. Price ND et al. A wellness study of 108 individuals using personal, dense, dynamic data clouds. Nat. Biotechnol. (2017) doi:10.1038/nbt.3870.
12. Stittrich AB et al. Genomic architecture of inflammatory bowel disease in five families with multiple affected individuals. Hum. Genome Var. (2016) doi:10.1038/hgv.2015.60.

13. Huang X, Pearce R. & Zhang Y. EvoEF2: accurate and fast energy function for computational protein design. doi:10.1093/bioinformatics/btz740.
14. Jiang L. et al. De novo computational design of retro-aldol enzymes. Science (2008) doi:10.1126/science.1152692.
15. Renata H, Wang ZJ & Arnold FH Expanding the enzyme universe: Accessing non-natural reactions by mechanism-guided directed evolution. Angewandte Chemie - International Edition (2015) doi:10.1002/anie.201409470.
16. Richter F, Leaver-Fay A, Khare SD, Bjelic S. & Baker D. De novo enzyme design using Rosetta3. PLoS ONE (2011) doi:10.1371/journal.pone.0019230.
17. Steiner K. & Schwab H. Recent advances in rational approaches for enzyme engineering. Comput. Struct. Biotechnol. J. (2012) doi:10.5936/csbj.201209010.
18. Sáez-Jiménez V. et al. Improving the pH-stability of versatile peroxidase by comparative structural analysis with a naturally-stable manganese peroxidase. PLoS ONE (2015) doi:10.1371/journal.pone.0140984.
19. Park HJ, Joo JC, Park K, Kim YH & Yoo YJ Prediction of the solvent affecting site and the computational design of stable Candida antarctica lipase B in a hydrophilic organic solvent. J. Biotechnol. (2013) doi:10.1016/j.jbiotec.2012.11.006.
20. Jiang C. et al. An orphan protein of Fusarium graminearum modulates host immunity by mediating proteasomal degradation of TaSnRK1α. Nat. Commun. 11, (2020).
21. Tautz D. & Domazet-Lošo T. The evolutionary origin of orphan genes. Nature Reviews Genetics vol. 12 692–702 (2011).
22. AlQuraishi M. End-to-End Differentiable Learning of Protein Structure. Cell Syst. 8, 292–301.e3 (2019). [PubMed: 31005579]
23. Ingraham J, Riesselman A, Sander C. & Marks D. Learning protein structure with a differentiable simulator. in 7th International Conference on Learning Representations, ICLR 2019 (2019).
24. Li J. Universal Transforming Geometric Network. (2019).
25. Kandathil SM, Greener JG, Lau AM & Jones DT Ultrafast end-to-end protein structure prediction enables high-throughput exploration of uncharacterised proteins. bioRxiv 2020.11.27.401232 (2021) doi:10.1101/2020.11.27.401232.
26. Rives A. et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc. Natl. Acad. Sci. U. S. A. 118, (2021).
27. Baek M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science 10, eabj8754 (2021).
28. Conway P, Tyka MD, DiMaio F, Konerding DE & Baker D. Relaxation of backbone bond geometry improves protein energy landscape modeling. Protein Sci. (2014) doi:10.1002/pro.2389.
29. Devlin J, Chang MW, Lee K. & Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. in NAACL HLT 2019 – 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference vol. 1 4171–4186 (2019).
30. Vaswani A. et al. Attention Is All You Need. arXiv (2017).
31. Leinonen R. et al. UniProt archive. Bioinformatics (2004) doi:10.1093/bioinformatics/bth191.
32. Meier J. et al. Language models enable zero-shot prediction of the effects of mutations on protein function. bioRxiv 2021.07.09.450648 (2021) doi:10.1101/2021.07.09.450648.
33. Elnaggar A. et al. CodeTrans: Towards Cracking the Language of Silicone's Code Through Self-Supervised Deep Learning and High Performance Computing. bioRxiv 14, (2021).
34. Alley E, Khimulya G, Biswas S, AlQuraishi M. & Church G. Unified rational protein engineering with sequence-only deep representation learning. bioRxiv 589333 (2019) doi:10.1101/589333.
35. Heinzinger M. et al. Modeling the language of life - Deep learning protein sequences. bioRxiv (2019) doi:10.1101/614313.
36. Madani A. et al. ProGen: Language modeling for protein generation. bioRxiv (2020) doi:10.1101/2020.03.07.982272.

37. Elnaggar A. et al. ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Learning. http://biorxiv.org/lookup/doi/10.1101/2020.07.12.199554 (2020) doi:10.1101/2020.07.12.199554.
38. Hu S, Lundgren M. & Niemi AJ Discrete Frenet frame, inflection point solitons, and curve visualization with applications to folded proteins. Phys. Rev. E - Stat. Nonlinear Soft Matter Phys. 83, (2011).
39. AlQuraishi M. ProteinNet: A standardized data set for machine learning of protein structure. BMC Bioinformatics (2019) doi:10.1186/s12859-019-2932-0.
40. Fox NK, Brenner SE & Chandonia JM SCOPe: Structural Classification of Proteins - Extended, integrating SCOP and ASTRAL data and classification of new structures. Nucleic Acids Res. (2014) doi:10.1093/nar/gkt1240.
41. Burley SK et al. RCSB Protein Data Bank: Powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. Nucleic Acids Res. 49, D437–D451 (2021). [PubMed: 33211854]
42. Touw WG et al. A series of PDB-related databanks for everyday needs. Nucleic Acids Res. (2015) doi:10.1093/nar/gku1028.
43. Outeiral C, Nissley DA & Deane CM Current structure predictors are not learning the physics of protein folding. Bioinformatics 38, 1881–1887 (2022). [PubMed: 35099504]
44. Hartrampf N. et al. Synthesis of proteins by automated flow chemistry. Science 368, 980–987 (2020). [PubMed: 32467387]
45. Rao R, Meier J, Sercu T, Ovchinnikov S. & Rives A. Transformer protein language models are unsupervised structure learners. bioRxiv (2020) doi:10.1101/2020.12.15.422761.
46. Kaplan J. et al. Scaling laws for neural language models. arXiv (2020).
47. Rao R. et al. MSA Transformer. (2021) doi:10.1101/2021.02.12.430858.
48. Anfinsen CB, Haber E, Sela M. & White FH The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. Proc. Natl. Acad. Sci. U. S. A. 47, 1309–1314 (1961). [PubMed: 13683522]
49. Mikolov T. et al. Strategies for Training Large Scale Neural Network Language Models.
50. Mirdita M. et al. Uniclust databases of clustered and deeply annotated protein sequences and alignments. Nucleic Acids Res. (2017) doi:10.1093/nar/gkw1081.
51. Xu J, McPartlon M. & Li J. Improved protein structure prediction by deep learning irrespective of co-evolution information. bioRxiv (2020) doi:10.1101/2020.10.12.336859.
52. Fleishman SJ et al. Rosettascripts: A scripting language interface to the Rosetta Macromolecular modeling suite. PLoS ONE (2011) doi:10.1371/journal.pone.0020161.

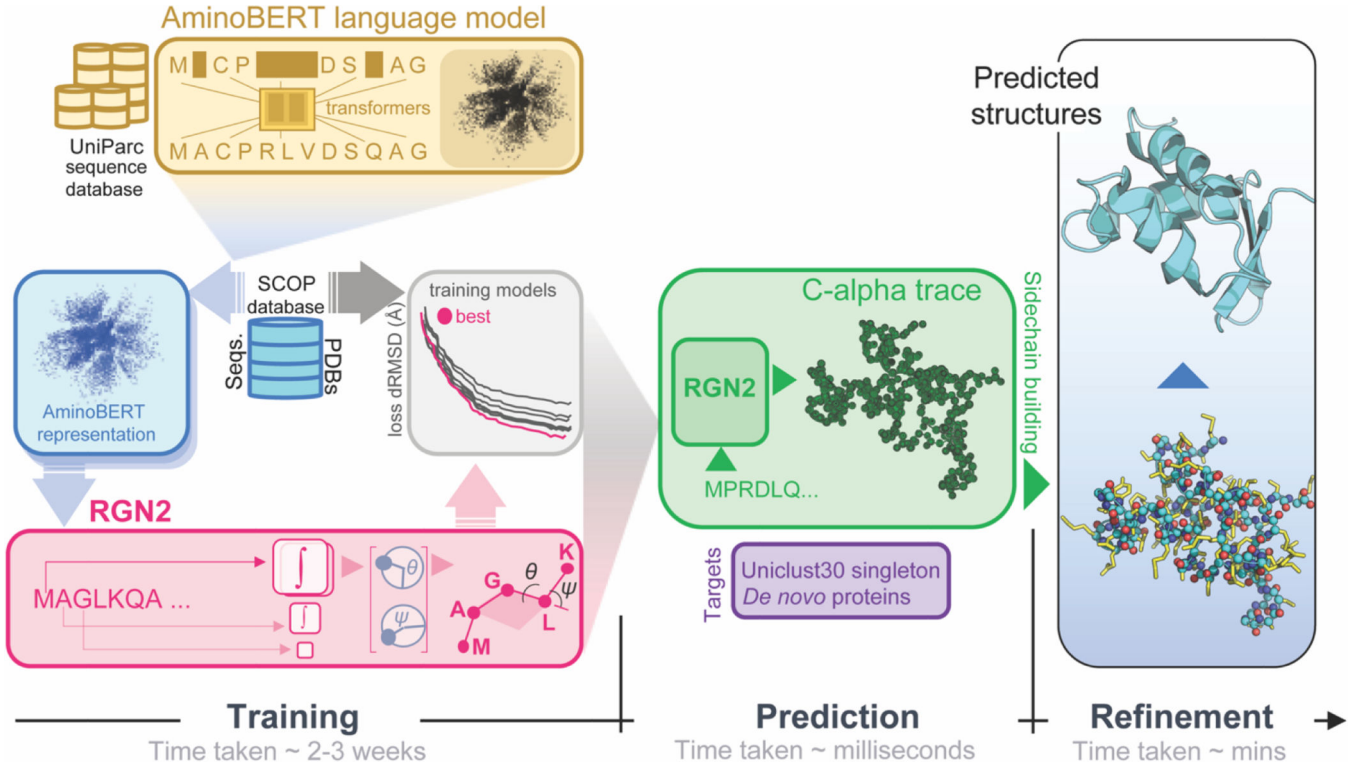

Figure 1. Organization and application of RGN2.

RGN2 combines a Transformer-based protein language model (AminoBERT; yellow) with a recurrent geometric network that utilizes Frenet-Serret frames to generate the backbone structure of a protein (green). Placement of side chain atoms and refinement of hydrogen-bonded networks are subsequently performed using the Rosetta energy function (blue).

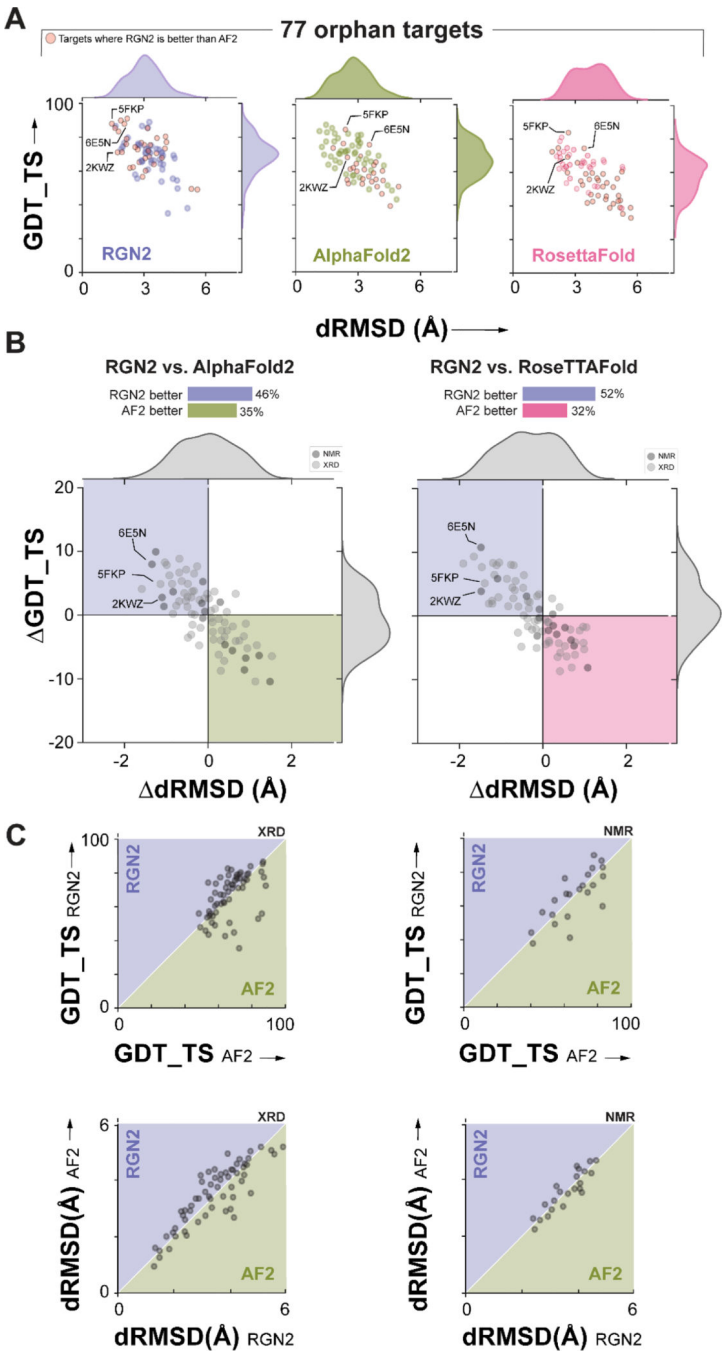

Figure 2. Prediction performance on orphan proteins.

(A) Absolute performance metrics for RGN2 (purple), AF2 (green), and RF (pink) across 77 orphan proteins lacking known homologs. **(B)** Differences in prediction accuracy between RGN2 and AF2 / RF are shown for the 77 orphan proteins, using dRMSD and GDT_TS as metrics. Points in top-left quadrant correspond to targets with negative ΔdRMSD and positive ΔGDT_TS, *i.e.,* where RGN2 outperforms the competing method on both metrics, and vice-versa for the bottom-right quadrant. The other two quadrant (white) indicate targets where there is no clear winner as the two metrics disagree. The structures of 20% of the

targets were determined experimentally using NMR and are denoted with dark gray markers while the remaining 80% of targets were determined using X-ray crystallography (XRD). **(C)** Head-to-head comparisons of absolute GDT_TS and dRMSD scores for RGN2 and AF2 are shown broken down by experimental method (NMR and XRD). RGN2 outperforms AF2 for proteins in the upper purple triangle while AF2 outperforms RGN2 for targets in the lower green triangle.

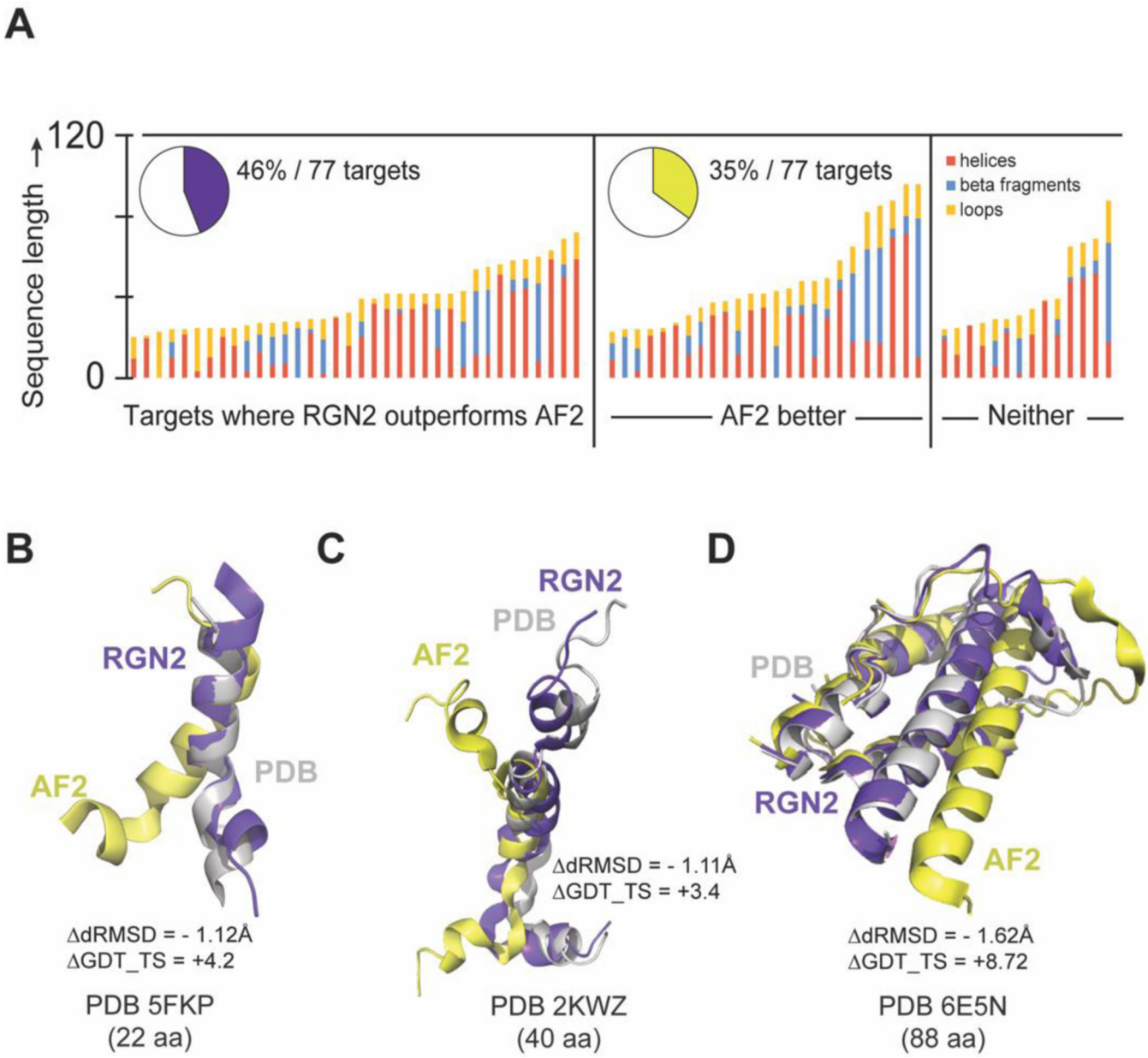

Figure 3. Comparing RGN2 and AF2 structure predictions for orphan proteins

(A) Stacked bar charts show the relative fractions of secondary structure elements in orphan proteins broken down by these categories: RGN2 outperforms AF2, AF2 outperforms RGN2, and no clear winner. Bar height indicates protein length. **(B-D)** Alpha helical targets of different lengths (5FKP, 2KWZ, and 6E5N) which contain bends or hydrogen-bonded turns between helical domains tend to be better predicted by RGN2 than AF2.

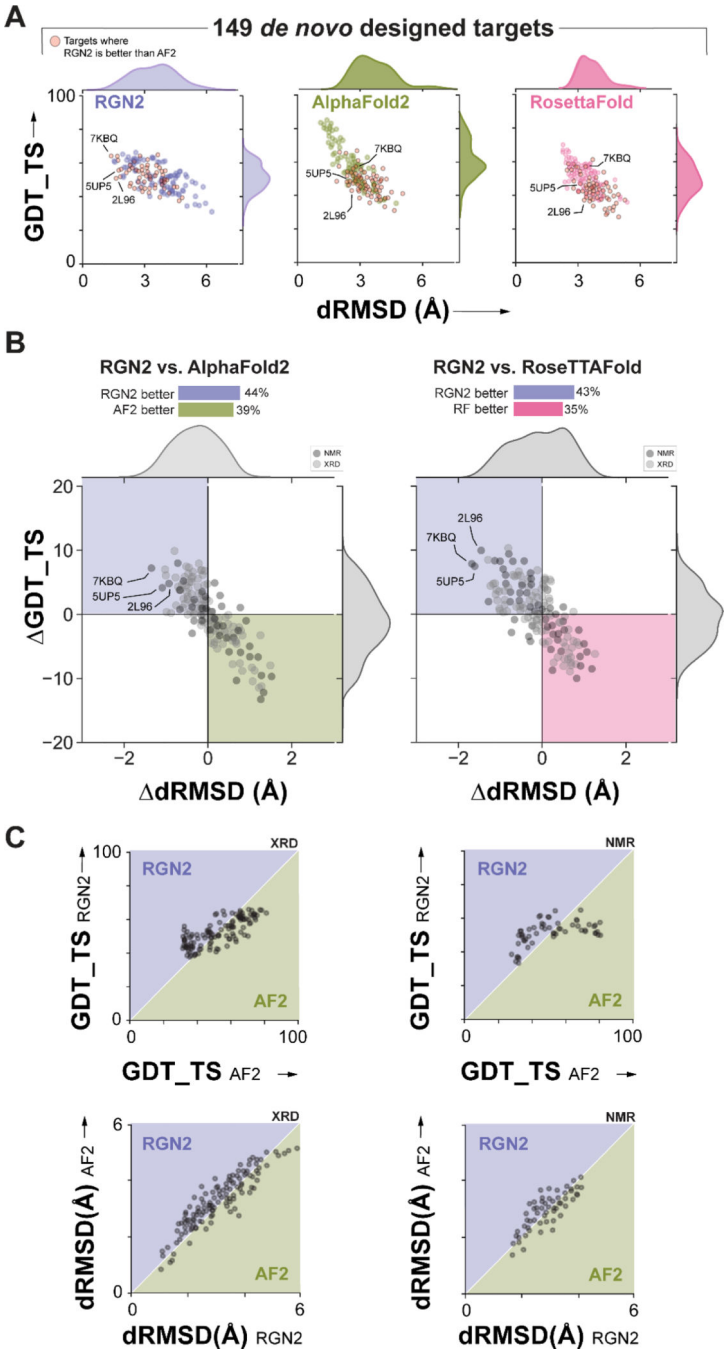

Figure 4. Prediction performance on Designed Proteins

(A) Absolute performance metrics for RGN2 (purple), AF2 (green), and RosettaFold (pink) across 149 *de novo* designed proteins. **(B)** Differences in prediction accuracy between RGN2 and AF2 are shown for these 149 proteins using dRMSD and GDT_TS as metrics. Points in top-left quadrant correspond to targets with negative ΔdRMSD and positive ΔGDT_TS, *i.e.,* where RGN2 outperforms the competing method on both metrics, and vice-versa for the bottom-right quadrant. The other two quadrant (white) indicate targets where there is no clear winner as the two metrics disagree. The structures of 34% of the

targets were determined experimentally using NMR and are denoted with dark gray markers while the remaining 64% of targets were determined using X-ray crystallography. (XRD). **(C)** Head-to-head comparisons of absolute GDT_TS and dRMSD scores for RGN2 and AF2 are shown broken down by experimental method (NMR and XRD). RGN2 outperforms AF2 for proteins in the upper purple triangle while AF2 outperforms RGN2 for targets in the lower green triangle.

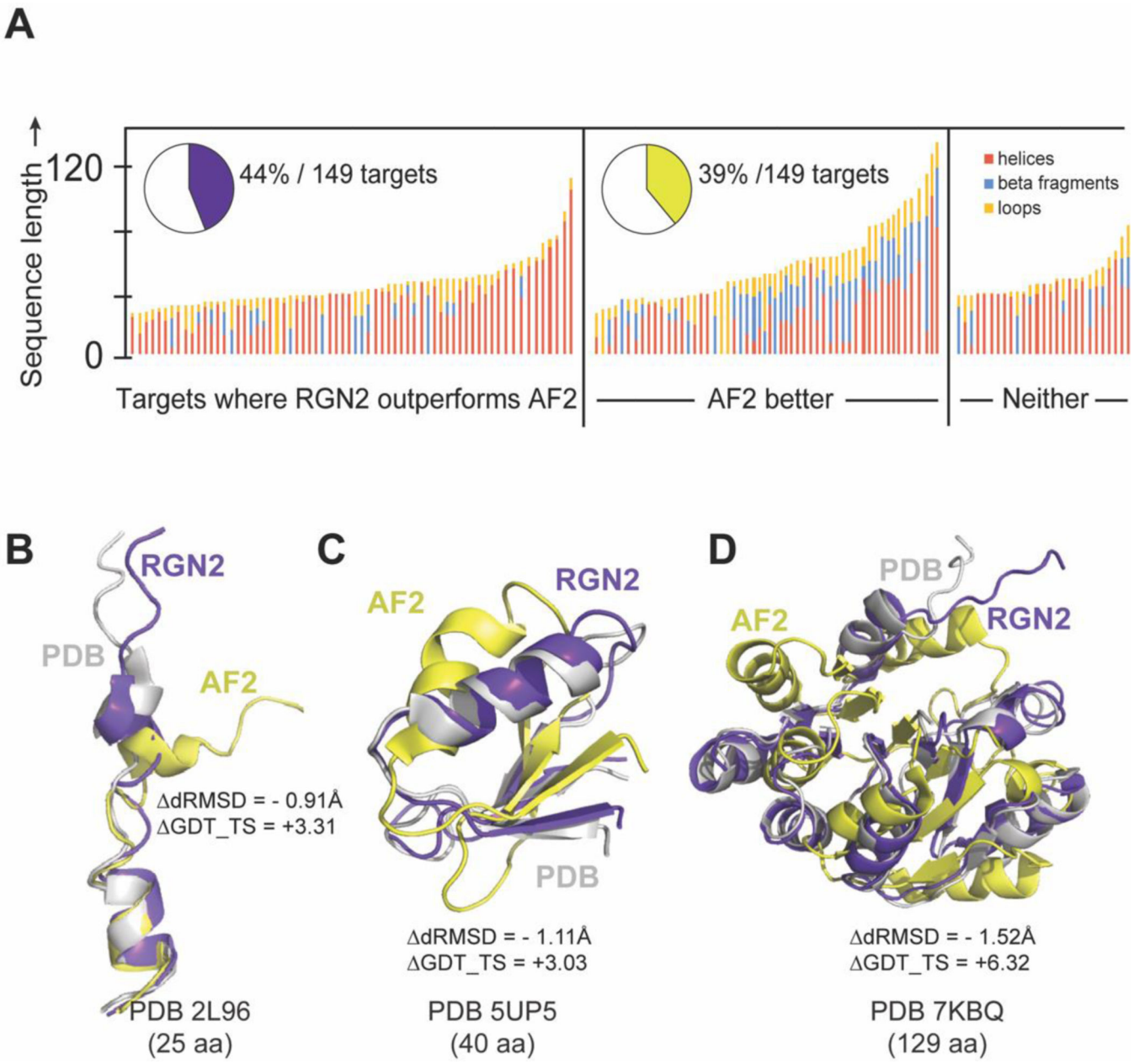

Figure 5. Comparing RGN2 and AF2 structure predictions for designed proteins.

(A) Stacked bar chart shows 149 *de novo* designed proteins. Bar height indicates protein length. **(B)** Overlaid ribbon diagrams of PDB entries (with increasing protein length) 2L96, 5UP5, and 7KBQ (white) and RGN2 (purple) and AF2 (yellow) predicted structures are visually depicted to show how RGN2 outperforms AF2 for each of these cases.

Table 1.

A quantitative comparison of average TM-Scores and precision of top L/2, L/5, and L/10 contacts and contacts within α-helical and β-type folds across 77 orphan proteins, and 149 *de novo* proteins, using ESM-1b and RGN2.

Targets	Method	Top L/x			Structural Classes	
		L/2	L/5	L/10	α-helix	β-type
77 orphans	RGN2	29.3	52.3	64.4	86.5	20.3
	ESM-1b	30.1	51.8	69.6	84.1	49.5
149 *de novo*	RGN2	35.6	55.1	61.8	87.9	23.3
	ESM-1b	29.3	54.5	68.4	84.1	39.2

Table 2.

Comparison of prediction times between RGN2 and AF2, RF, and trRosetta across 330 targets spanning our orphan and *de novo* protein datasets. RGN2 predictions were performed in batches with maximum permissible batch size set to 128 targets. The trRosetta MSA generation step was not used since none of the targets had known homologous proteins.

Protein length (L) bins (# residues)	Total targets	Mean protein length (# residues)	Mean trRosetta prediction time per structure (s)		Mean AF2 prediction time (s)	Mean RF prediction time (s)	Mean RGN2 prediction time (ms)	Mean RGN2 prediction + refinement time (s)
			Distogram	3D-Structure				
$0 < L \leq 100$	184	37.5	1768	1004	831.5	412.6	2.7	145.2
$100 < L \leq 200$	93	148.7	2791	1927	851.6	408.3	2.2	177.4
$200 < L \leq 300$	28	258.3	2877	1752	828.4	492.7	3.1	182.3
$300 < L \leq 400$	17	333.4	3647	2140	825.6	501.6	5.7	200.5
$400 > L$	8	460.5	4012	3011	841.6	498.6	5.9	229.2