Frame2seq: structure-conditioned masked language modeling for protein sequence design

Deniz Akpinaroglu
UCSF
San Francisco, CA 94143
deniz.akpinaroglu@ucsf.edu

Kosuke Seki
UCSF
San Francisco, CA 94143
kosuke.seki@ucsf.edu

Eleanor Zhu
UCSF
San Francisco, CA 94143
tianqi.zhu@ucsf.edu

Tanja Kortemme
UCSF
San Francisco, CA 94143
kortemme@cgl.ucsf.edu

Abstract

Machine learning has revolutionized computational protein design, enabling significant progress in protein backbone generation and sequence design. For protein sequence design, encoder-decoder models have achieved state-of-the-art accuracy, which has translated to experimental success. Here, we introduce Frame2seq, a structure-conditioned masked language model for protein sequence design that, in contrast to autoregressive design methods, generates sequences in a single pass. On the CATH 4.2 test dataset, Frame2seq outperforms the state-of-the-art autoregressive method, ProteinMPNN, achieving 49.1% sequence recovery (2.0% improvement) with over six times faster inference. In addition, Frame2seq accurately estimates the error in its own predictions. To probe the ability of Frame2seq to generate novel designs beyond native-like sequence space, we experimentally test 26 Frame2seq designs for de novo backbones with low identity to the starting sequences. We show that Frame2seq successfully designs soluble (22/26), monomeric, folded, and stable proteins (17/26), including a design with 0% sequence identity to native. The speed and accuracy of Frame2seq will accelerate exploration of novel sequence space across diverse design tasks, including challenging applications such as multi-objective optimization.

1 Introduction

Proteins are molecules that drive cellular processes in all living systems. The ability to design proteins with new functions thus has widespread applications in biotechnology and medicine. Traditionally, computational protein design has relied on physics-based principles and simulations. Recently, machine learning methods learning directly from data have enabled promising advances, such as highly accurate structure prediction with AlphaFold2 [9] and comparable methods [4, 12, 17], as well as inverting these models for protein design [3, 14, 2, 7, 15, 18].

Computational protein design has achieved high experimental success rates in diverse applications, including generation of new protein folds [3, 15] and symmetrical oligomers [15, 16], protein-protein binder design [15], and motif-scaffolding [14, 15]. Essentially all of these methods first generate a protein backbone followed by fixed backbone sequence design as a second step.

ProteinMPNN is a state-of-the-art encoder-decoder model that autoregressively predicts protein sequences given a backbone [5]. ProteinMPNN has been experimentally validated to yield successful

Machine Learning for Structural Biology Workshop, NeurIPS 2023.

monomeric proteins and protein assemblies. Alternative methods built on encoder-only architectures have also been proposed, such as PiFold [6]. PiFold outperforms ProteinMPNN on native sequence recovery but has not been experimentally validated [6].

In this work, we introduce Frame2seq[1], a structure-conditioned masked language model that designs protein sequences. Frame2seq is a bidirectional encoder and achieves fast inference due to single-pass sequence generation. Through in silico and experimental evaluation, we demonstrate that Frame2seq is a state-of-the-art fixed-backbone design method and produces soluble, monomeric, stable proteins. Importantly, we demonstrate that our model pseudo-log-likelihood is highly correlated with prediction success. We also test and experimentally validate the ability of Frame2seq to explore novel sequences with undetectable similarity to the starting protein. We expect this ability to enable particularly challenging applications of sequence design such as multi-objective optimization.

2 Methods

2.1 Datasets

We train Frame2seq on the non-redundant CATH 4.2 [13] data splits of single chain proteins up to length 500 as described by [8]. There are 18024, 608, and 1120 chains with no topological overlap in the training, validation, and test sets, respectively. These splits are clustered for structural diversity and allow to test generalization across many folds [8]. To perform direct comparisons, we benchmark against models trained on the same dataset.

2.2 Structure-conditioned masked language model

Frame2seq is a translation- and rotation-invariant encoder-only model that takes protein backbones as input and generates protein sequences (Figure 1). In a single inference pass, Frame2seq designs complete protein sequences. We preserve the invariance property using invariant point attention (IPA) [9]. We compute rotations and translations from input backbone coordinates to construct coordinate frames as described for AlphaFold2 [9]. To obtain sequence embeddings, we compute phi, psi, and omega torsions, along with absolute position embedding. To obtain pairwise embeddings, we compute inter-residue distances lifted into a radial basis and relative position indices between pairs of residues. Coordinate frames, sequence embeddings, and pairwise embeddings pass through a repeating stack of node and edge update operations followed by a final node update and transition to sequence dimension. IPA layers and transition layers comprise the node updates. We update edges by passing pairwise embeddings and updated sequence embeddings into two layers of MLP. Frame2seq is trained to minimize a categorical cross entropy loss for a native sequence recovery objective.

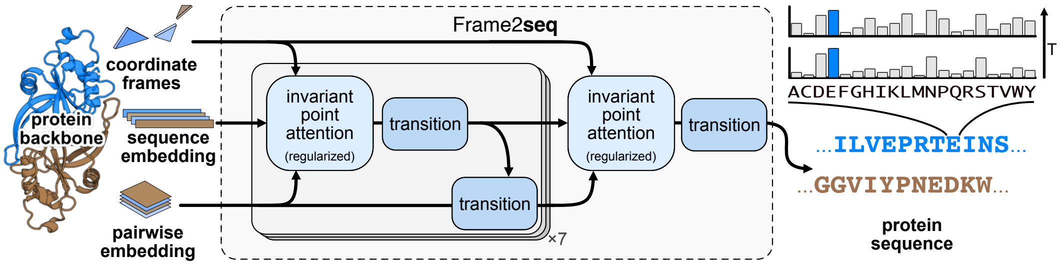

Figure 1: Model architecture. Frame2seq is composed of regularized invariant point attention layers and sequence embedding transitions followed by pairwise embedding updates. Arrows indicate information flow through model components.

[1] Frame2seq code is available at https://github.com/dakpinaroglu/Frame2seq

Table 1: Computational benchmark of model performance and speed over CATH 4.2 held out test dataset targets. Sequence recovery is median of average. AlphaFold2 high confidence rate is percentage of targets that achieve pLDDT > 90. AlphaFold2 success rate is percentage of targets that achieve LDDT-Cα > 90. GPU speed (s) is mean of average.

	Sequence recovery (%) $\uparrow$	AlphaFold2 high confidence rate (%) $\uparrow$		AlphaFold2 success rate (%) $\uparrow$		GPU speed (s) $\downarrow$
		With MSA	Without MSA	With MSA	Without MSA	
Native	-	66.25	2.59	48.48	1.34	-
ProteinMPNN	47.10	**53.21**	**4.82**	**39.02**	**3.21**	2.71
Frame2seq	**49.11**	**53.21**	4.02	38.39	3.04	**0.44**

2.3 Attention regularization for IPA layers

We hypothesize that introducing tolerable difficulty during training allows for better model performance. To increase training difficulty without halting the model's ability to minimize the categorical cross entropy loss, we explore feature dropouts and attention regularization for IPA. We implement regularized IPA layers by randomly masking attention weights between pairs of residues during training (Algorithm 1). We find that a masking rate of 20% is optimal and results in improved model performance (Table S2). We attribute this improvement to rendering the training task sufficiently but not overly more difficult.

3 Results

3.1 Frame2seq recovers native sequence from structure.

We benchmarked Frame2seq against a ProteinMPNN model trained on the CATH 4.2 dataset without backbone noise. We evaluated the performance of both models when provided no ground truth sequence as context and found that Frame2seq was more accurate than ProteinMPNN. Specifically, Frame2seq achieves 49.1% native sequence recovery, outperforming ProteinMPNN by 2.0%. We predicted structures for the native sequences, ProteinMPNN designs, and Frame2seq designs using all 5 AlphaFold2 models with 3 recycles and no templates. We calculated AlphaFold2 high confidence rate (%) and AlphaFold2 success rate (%), both with and without MSAs. We found that ProteinMPNN and Frame2seq achieve similar high confidence and success rates. Both models are outperformed by the native sequences when MSAs are provided, which is expected given the inclusion of native sequences in the AlphaFold2 training set (Figure S2). When averaged over the test dataset targets, Frame2seq inference speed is approximately 6.2 times faster than ProteinMPNN (Table 1). This difference is due to Frame2seq's single-pass and ProteinMPNN's autoregressive sampling formulation.

Incorrect residue predictions of fixed-backbone design methods reveal their underlying compositional bias for certain amino acid types. Compositional bias difference is calculated by subtracting the predicted number of residue type occurrences from true and dividing by the total number. We calculated compositional bias difference for ProteinMPNN and Frame2seq for all amino acid types and found that Frame2seq has overall less bias (Figure 2A).

We next investigated how the performance of ProteinMPNN and Frame2seq depended on residue burial. To measure burial, we computed the average Cβ distance for 8 closest residues (Å) (lower for core residues, higher for surface residues). We found that residue burial has a similar effect on both models, with the expected behavior that core residues are easier to recover than surface residues. Frame2seq's 2.0% native sequence recovery improvement over ProteinMPNN is primarily at the surface residues of the targets (Figure S1).

3.2 Frame2seq accurately estimates the error in its own predictions

We investigated the ability of Frame2seq to discriminate accurate from inaccurate predictions. Towards this goal, we analyzed how well model pseudo-log-likelihood (PLL) correlates with native sequence recovery. For both native and sampled sequences, we computed model PLL and found it to

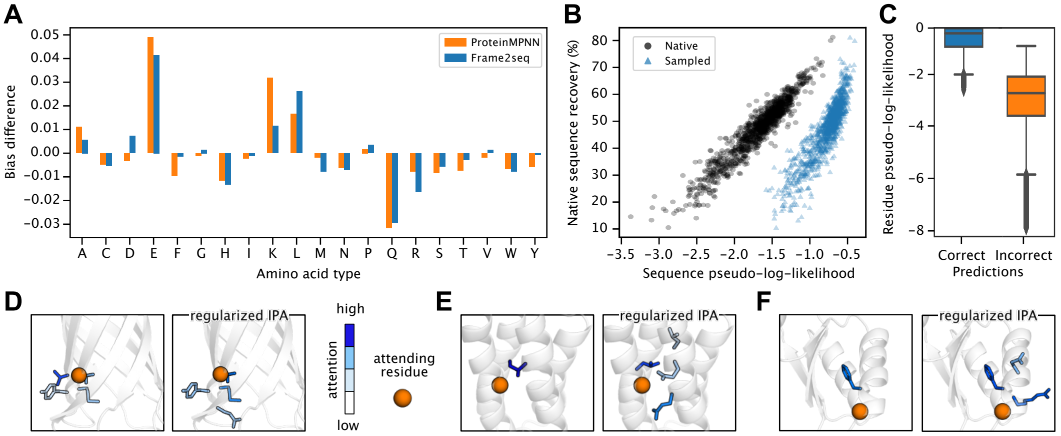

Figure 2: In silico analysis of Frame2seq. (A) Difference in compositional bias between ProteinMPNN and Frame2seq. (B) Correlation between Frame2seq native sequence recovery and model pseudo-log-likelihood averaged over sequence for native (dark blue, Spearman correlation = 0.94) and sampled (light blue, Spearman correlation = 0.89) sequences. (C) Per-residue model pseudo-log-likelihood for correct residue identity (native) and incorrect residue identity (non-native) predictions. Incorrect prediction outliers below 3 times the interquartile range are not shown. (D-F) Frame2seq attends to second-shell interactions, bulky side chains, and side chains pointing towards attending residue with IPA regularization (right). Distribution of attention to residues by the attending residue (orange sphere) with IPA (left) and regularized IPA (right). Attention value increases from white to dark blue. Side chains are only shown for residues with attention. (D) PDBID 6X1K. (E) PDBID 1P68. (F) PDBID 2LV8.

strongly correlate with native sequence recovery (Figure 2B). Frame2seq scores native and sampled sequences favorably when the model predicts accurately. Frame2seq prefers its own predictions over the native sequences, which is expected due to model bias and the possibility of finding alternative sequences that fold into the input structure. We additionally found model PLL to discriminate accurately between native and non-native predictions when computed per position (Figure 2C).

3.3 IPA regularization enforces side chain awareness

With IPA regularization, Frame2seq learns node updates from the context of a restricted local environment. When attention between pairs of residues is randomly dropped, the model attends to alternative residues that provide the most context for sequence prediction (Figure 2D-F). We found that this enforced Frame2seq to allot more attention to residues with bulkier side chains (Figure 2D-F), alternative residues with side chains oriented towards the attending residue (Figure 2E), and longer-range second shell interactions (Figure 2E-F). Training with restricted attention capacity contributes to model's improved performance (Table S2).

3.4 Frame2seq designs stable sequences onto de novo backbones

In silico evaluation of Frame2seq demonstrates its ability to generate realistic protein sequences in principle. However, experimental validation of a design method is essential, because even one or a few incorrect amino acids in a protein, while causing a small difference in sequence recovery, can be detrimental to protein stability. Beyond testing for its ability to output native-like proteins, we used our model to design sequences onto de novo backbones to have low (zero to fifty percent) sequence identity to the native sequence. This is a challenging task for a model that only learns from the native sequences associated with ground truth backbones. However, we found that Frame2seq's generative capabilities extend beyond the naturally occurring sequence space as we successfully designed novel low sequence identity proteins.

3.4.1 Design and characterization of low sequence identity proteins

We generated low sequence identity sequences for the backbones of a de novo Rossmann 2x2 fold (PDBID 2LV8)[10] and the first computationally designed de novo fold, Top7 (PDBID 1QYS)[11]. To achieve low sequence identity to native, we restricted Frame2seq model logits to exclude the true residue identity at each or randomly chosen position(s) before sampling sequences. We then predicted structures for each candidate using all 5 AlphaFold2 models with no MSAs, no templates, and 3 recycles. We filtered our designs by calculating an average pLDDT and computing structural deviation from the true backbones, and selected designs with pLDDT > 89 and predicted backbone heavy atom RMSD < 1.15 Å. We searched our design sequences against non-redundant protein sequences and the PDB database and found minimal to no matches. Frame2seq designs with less than 20% sequence identity to the native yield zero matches, and above 20% identity designs are most significantly matched to the native.

We experimentally evaluated twenty-six sequences (eight and eighteen designed onto the de novo Rossmann and Top7 backbones, respectively). Out of the total 26, 25 of our designs express in *E. coli*, 22 are soluble, 17 are monomeric and folded (Figure 3A). 16 of these experimentally successful designs do not melt at up to 95 °C. We demonstrate experimental success over all sequence identity bins we explored, including 0% (Figure 3B). Figure 3C-E highlights biophysical characterization of a Top7 design with 0% sequence identity to the starting protein (PDBID 1QYS). This novel protein, Top0, is monomeric when assessed by size exclusion chromatography (Figure 3C), folds into the expected secondary structure as measured by circular dichroism (Figure 3D), and does not melt at up to 95 °C (Figure 3E). This design challenge shows that Frame2seq successfully samples unexplored sequence space for de novo backbones while likely maintaining a near-identical structure as assessed by AlphaFold2 predictions (Figure 3F).

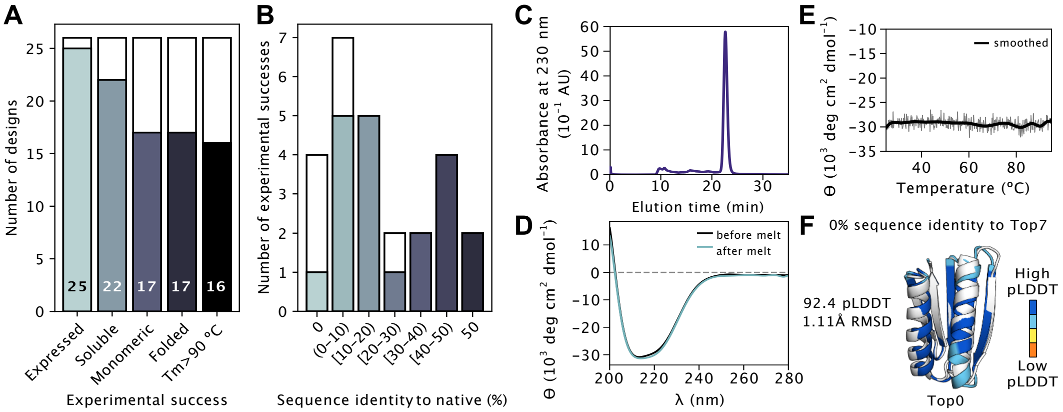

Figure 3: Experimental evaluation of Frame2seq designs. (A) Number of designs for the de novo Rossmann fold (PDBID 2LV8) and Top7 (PDBID 1QYS) backbones that achieve experimental success (filled bars). (B) Number of experimentally successful designs (filled bars) for different sequence identity bins. (C-E) Experimental characterization of a low sequence identity design, Top0, with 0% sequence identity to the native Top7. (C) Size exclusion chromatography profile for the Top0 design. (D) Circular dichroism (CD) spectra of the Top0 design. (E) Changes in CD signal (mean residue ellipticity) of the Top0 design as a function of temperature. (F) AlphaFold2 prediction for the Top0 design (colored by pLDDT) aligned to the 1QYS X-ray structure (gray).

4 Conclusion

Frame2seq is a structure-conditioned masked language model for protein sequence design. Compared to the state-of-the-art experimentally verified alternative, Frame2seq demonstrates improved sequence recovery with significantly reduced inference time. Our implementation of IPA regularization increases local context awareness for sequence design. We show that Frame2seq learns to estimate the error in its own predictions both per-residue and over entire sequences and that model scores are predictive of accuracy. While native sequence recovery is a common measure of performance for fixed-backbone design methods, the true measure of success for such a method would assess its ability to expand over the conditional sequence space exploited by nature without being limited to a single observed sequence. We assess Frame2seq's true success by designing for low sequence identity, down to zero percent. Given de novo backbones as input, Frame2seq successfully designs stable, soluble, monomeric proteins that expand beyond the naturally occurring sequence space. We believe our method establishes a foundation for striking a balance between sufficient recovery of evolutionarily relevant sequence and sufficient divergence from a single observed example. We expect our method will be uniquely useful for multi-objective design applications where satisfaction of multiple functional criteria might require exploring larger sequence spaces to generate successful pareto-optimal solutions.

Acknowledgments and Disclosure of Funding

This work was supported by a National Science Foundation Fellowship (NSF GRFP) to D.A.. T.K. is a Chan Zuckerberg Investigator.

References

[1] Gustaf Ahdritz, Nazim Bouatta, Sachin Kadyan, Qinghui Xia, William Gerecke, Timothy J O'Donnell, Daniel Berenberg, Ian Fisk, Niccolò Zanichelli, Bo Zhang, et al. Openfold: Retraining alphafold2 yields new insights into its learning mechanisms and capacity for generalization. *bioRxiv*, pages 2022–11, 2022.

[2] Namrata Anand and Tudor Achim. Protein structure and sequence generation with equivariant denoising diffusion probabilistic models. *arXiv preprint arXiv:2205.15019*, 2022.

[3] Ivan Anishchenko, Samuel J Pellock, Tamuka M Chidyausiku, Theresa A Ramelot, Sergey Ovchinnikov, Jingzhou Hao, Khushboo Bafna, Christoffer Norn, Alex Kang, Asim K Bera, et al. De novo protein design by deep network hallucination. *Nature*, 600(7889):547–552, 2021.

[4] Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N Kinch, R Dustin Schaeffer, et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, 2021.

[5] Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J Ragotte, Lukas F Milles, Basile IM Wicky, Alexis Courbet, Rob J de Haas, Neville Bethel, et al. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378(6615):49–56, 2022.

[6] Zhangyang Gao, Cheng Tan, and Stan Z Li. Pifold: Toward effective and efficient protein inverse folding. *arXiv preprint arXiv:2209.12643*, 2022.

[7] John Ingraham, Max Baranov, Zak Costello, Vincent Frappier, Ahmed Ismail, Shan Tie, Wujie Wang, Vincent Xue, Fritz Obermeyer, Andrew Beam, et al. Illuminating protein space with a programmable generative model. *BioRxiv*, pages 2022–12, 2022.

[8] John Ingraham, Vikas Garg, Regina Barzilay, and Tommi Jaakkola. Generative models for graph-based protein design. *Advances in neural information processing systems*, 32, 2019.

[9] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[10] 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, 2012.

[11] 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, 2003.

[12] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

[13] Ian Sillitoe, Nicola Bordin, Natalie Dawson, Vaishali P Waman, Paul Ashford, Harry M Scholes, Camilla SM Pang, Laurel Woodridge, Clemens Rauer, Neeladri Sen, et al. Cath: increased structural coverage of functional space. *Nucleic acids research*, 49(D1):D266–D273, 2021.

[14] Jue Wang, Sidney Lisanza, David Juergens, Doug Tischer, Joseph L Watson, Karla M Castro, Robert Ragotte, Amijai Saragovi, Lukas F Milles, Minkyung Baek, et al. Scaffolding protein functional sites using deep learning. *Science*, 377(6604):387–394, 2022.

[15] Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.

[16] BIM Wicky, LF Milles, A Courbet, RJ Ragotte, J Dauparas, E Kinfu, S Tipps, RD Kibler, M Baek, F DiMaio, et al. Hallucinating symmetric protein assemblies. *Science*, 378(6615):56–61, 2022.

[17] Ruidong Wu, Fan Ding, Rui Wang, Rui Shen, Xiwen Zhang, Shitong Luo, Chenpeng Su, Zuofan Wu, Qi Xie, Bonnie Berger, et al. High-resolution de novo structure prediction from primary sequence. *BioRxiv*, pages 2022–07, 2022.

[18] Jason Yim, Brian L Trippe, Valentin De Bortoli, Emile Mathieu, Arnaud Doucet, Regina Barzilay, and Tommi Jaakkola. Se (3) diffusion model with application to protein backbone generation. *arXiv preprint arXiv:2302.02277*, 2023.

5 Supplementary Material

5.1 Model details

5.1.1 Model architecture and training

The IPA components are based on the AlphaFold2 Structure Module [9] PyTorch implementation from OpenFold [1]. Our model contains 8 IPA layers followed by Structure Module transition layers. Unless it is the last IPA layer, the node transitions are followed by edge transitions (7 total). We set the hidden dimension to 128 and the Structure Module transition dimension to 512. Frame2seq contains approximately 10M trainable parameters. Frame2seq models are trained with early stopping, for 144 or 200 epochs over the full dataset, which takes approximately 4 or 6 days on a single NVIDIA A40 GPU. We trained models using the Adam optimizer with the Noam scheduler.

Table S1: Frame2seq hyperparameters.

Parameter	Value	Description
d_{node}	128	Node dimension
d_{edge}	128	Edge dimension
$n_{\text{ipa-layers}}$	8	IPA layers
$n_{\text{ipa-heads}}$	4	IPA attention heads
$d_{\text{ipa-scalar-key}}$	16	IPA scalar key dimension
$d_{\text{ipa-scalar-value}}$	16	IPA scalar value dimension
$d_{\text{ipa-point-key}}$	4	IPA point key dimension
$d_{\text{ipa-point-value}}$	6	IPA point value dimension

The IPA layers encode the sequence embedding, the edge embedding and the coordinate frames, then predict and update the sequence embedding. The edge embeddings are updated by standard message passing with 2 fully connected layers.

We train models with partial native sequence as input node features according to the following mask rate:

•With 75% probability, all amino acids are replaced with a masked residue token.

•With 25% probability, a sampled 0-100% of amino acids are provided for featurization.

5.1.2 Regularized invariant point attention layers

We adopt the invariant point attention from [9] and implement attention regularization via a dropout. Attention between pairs of residues is randomly masked at a 20% rate during training via the dropout in Algorithm 1 line 9.

Algorithm 1 Regularized invariant point attention (IPA)

```
function REGULARIZED IPA($\mathbf{s}_i$, $\mathbf{z}_{ij}$, $T_i$):
    $\mathbf{q}_i^h, \mathbf{k}_i^h, \mathbf{v}_i^h$ = LinearNoBias($\mathbf{s}_i$)
    $\vec{\mathbf{q}}_i^{hp}, \vec{\mathbf{k}}_i^{hp}$ = LinearNoBias($\mathbf{s}_i$)
    $\vec{\mathbf{v}}_i^{hp}$ = LinearNoBias($\mathbf{s}_i$)
    $b_{ij}^h$ = LinearNoBias($\mathbf{z}_{ij}$)
    $w_C = \sqrt{\frac{2}{9N_{\text{query points}}}}$
    $w_L = \sqrt{\frac{1}{3}}$
    $a_{ij}^h = w_L \left( \frac{1}{\sqrt{c}} \mathbf{q}_i^{h\top} \mathbf{k}_j^h + b_{ij}^h - \frac{\gamma^h w_C}{2} \sum_p \left\| \{T_i \circ \vec{\mathbf{q}}_i^{hp} - T_j \circ \vec{\mathbf{k}}_i^{hp} \right\|^2 \right)$
    $a_{ij}^h$ = softmax$_j$(Dropout$_{0.2}$($a_{ij}^h$))
    $\tilde{\mathbf{o}}_i^h = \sum_j a_{ij}^h \mathbf{z}_{ij}$
    $\mathbf{o}_i^h = \sum_j a_{ij}^h \mathbf{v}_j^h$
    $\vec{\mathbf{o}}_i^{hp} = T_i^{-1} \circ \sum_j a_{ij}^h (T_j \circ \vec{\mathbf{v}}_j^{hp})$
    $\tilde{\mathbf{s}}_i$ = Linear$\left( \text{concat}_{h,p}(\tilde{\mathbf{o}}_i^h, \mathbf{o}_i^h, \vec{\mathbf{o}}_i^{hp}, \left\| \vec{\mathbf{o}}_i^{hp} \right\|) \right)$
    return $\tilde{\mathbf{s}}_i$
end function
```

5.1.3 Ablation study

We ablate IPA regularization and edge update operations to study their effects on the performance of Frame2seq.

Table S2: Ablation study. Sequence recovery is median of average.

Ablation	Sequence recovery at 100 epochs (%) ↑	Sequence recovery at final epoch (%) ↑
No IPA regularization	44.31	—
No edge update operations	44.32	—
Baseline	46.53	—
Ensemble	47.79	**49.11**

We ensemble 3 baseline models that are independently trained. We report Frame2seq's performance as an ensemble.

5.2 Further in silico analysis

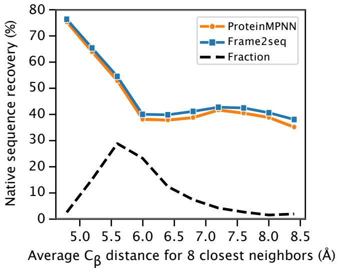

Fig. S1: Native sequence recovery as a function of average Cβ distance for 8 closest neighbors.

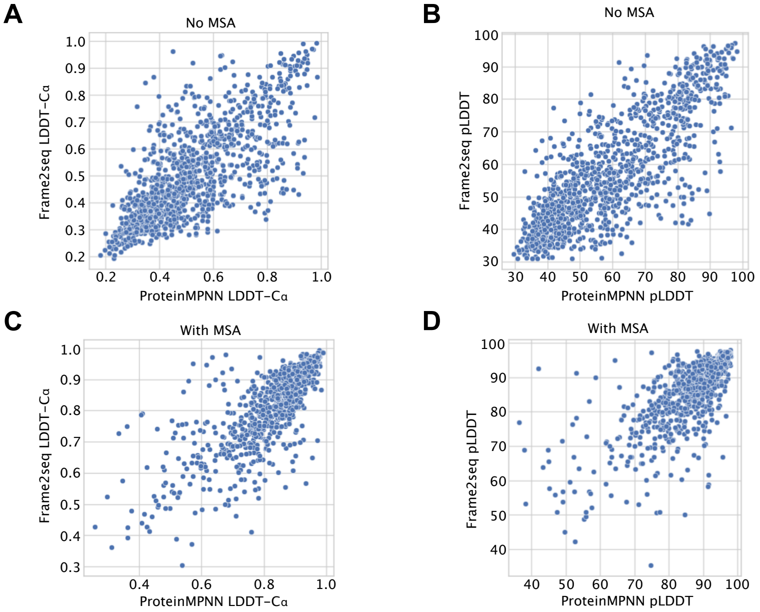

Fig. S2: AlphaFold2 accuracy and confidence for ProteinMPNN and Frame2seq sequences. (A) AlphaFold2 accuracy (LDDT-Cα) for predictions without MSAs. (B) AlphaFold2 confidence (pLDDT) for predictions without MSAs. (C) AlphaFold2 accuracy (LDDT-Cα) for predictions with MSAs. (D) AlphaFold2 confidence (pLDDT) for predictions with MSAs.

Table S3: Computational benchmark of model speed over CATH 4.2 held out test dataset targets. CPU speed (s) and GPU speed (s) is reported as mean of average for ProteinMPNN and Frame2seq inference.

Method	CPU speed (s) ↓	GPU speed (s) ↓
ProteinMPNN	13.70	2.71
Frame2seq	**6.79**	**0.44**

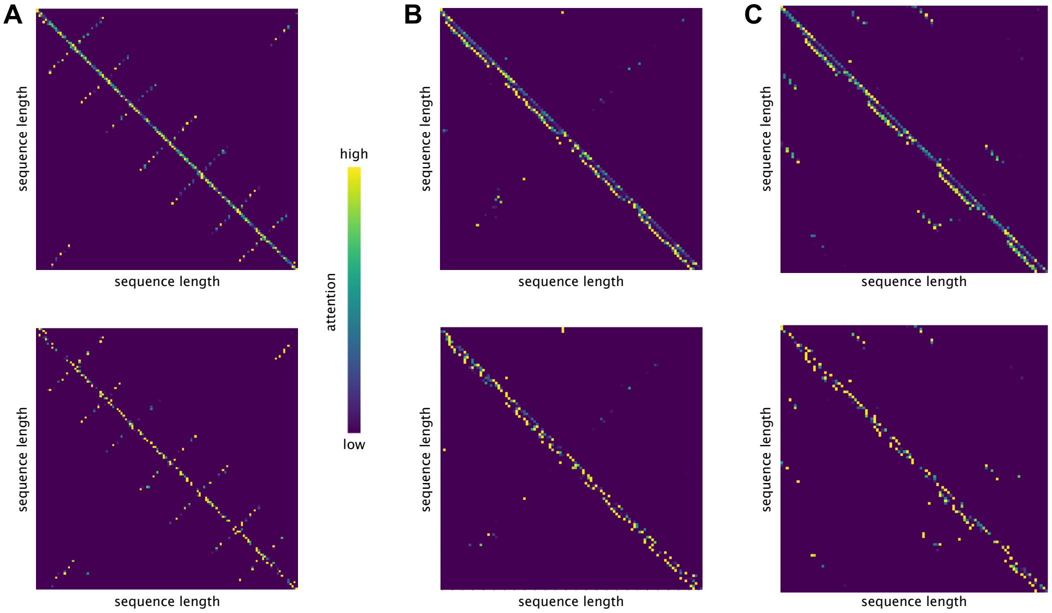

Fig. S3: Effects of regularizing IPA on attention matrices. (A-C) Attention between pairs of residues with IPA (top) and with regularized IPA (bottom). (A) PDBID 6X1K. (B) PDBID 1P68. (C) PDBID 2LV8.

5.3 Experimental methods

5.3.1 Protein expression and purification

Designs were ordered with an N-terminal 6xHis-tag and thrombin cleavage site and were purchased from Twist Biosciences in a pET-28a+ vector. DNA was purified by transforming into DH5a *E. coli* cells, picking a single colony into LB-Kan[50], and miniprepping.

To screen for solubility and expression of the designs, chemically competent BL21(DE3) were transformed with DNA encoding designs using manufacturer's protocol and plated onto LB-Kan[50] plates at 37 °C. A single colony was inoculated into 5 mL of LB-Kan[50] media at 37 °C or 18 °C with 220 rpm shaking overnight. The next day, 1 mL of culture was pelleted and stored at -80 °C, and protein expression was started in the remainder of the culture by adding 1 mM IPTG and growing at 30 °C or 18 °C. Cells were allowed to grow overnight, and were pelleted by centrifugation and stored at -80 °C. Cells were lysed either by using B-PER Bacterial Protein Extraction Reagent or by passing through a microfluidizer and centrifuged to pellet insoluble components. Lysate from uninduced cells, total induced cells, the supernatant of induced cells, and the pellet of induced cells were analyzed by SDS-PAGE to determine the solubility and expression level of protein designs.

For soluble designs, protein expression was scaled up. A single colony from a fresh transformation was inoculated into 5 mL of LB-Kan[50] media at 37 °C with 220 rpm shaking overnight. The next day, the entire overnight culture was added into either 250 mL or 500 mL of TB-Kan[50] with 220 rpm shaking at 37 °C. At OD = 0.4-0.6, protein expression was induced by adding IPTG to a final concentration of 1 mM. Temperature was reduced by transferring the cultures to a preheated 30 °C incubator, or by transferring cultures to a 4 °C cold room for 30 min before being transferred to a precooled 18 °C incubator. Cultures were allowed to grow overnight. Cells were harvested by centrifugation at 5,000 x g for 15 min at 4 °C.

Cell pellets were resuspended in 2 mL of resuspension buffer (30 mM sodium phosphate monobasic pH 7.5, 300 mM sodium chloride, 10 mM imidazole pH 7.5, 1 mg/mL Hen Egg White Lysozyme, and 15 U/mL Benzonase) per g of cell pellet. Cells were allowed to incubate in resuspension buffer for 30 minutes at room temperature with gentle shaking. Cells were additionally lysed either by passing the cell suspension through a microfluidizer three times, or by sonication on ice at 15% amplitude for 2 minutes using 5 s on / 5 s off cycles. Insoluble components were pelleted by centrifugation at 20 k x g for 20 min at 4 °C. Supernatant was decanted and incubated with 2 mL of pre-equilibrated 50% Ni-NTA resin slurry for 1 hr at 4 °C with gentle end-over-end shaking. The resin was washed four times with 20 mL of 30 mM sodium phosphate monobasic pH 7.5, 300 mM sodium chloride, and 20 mM imidazole pH 7.5. Proteins were eluted with 10 mL of 30 mM sodium phosphate pH 7.5, 300 mM sodium chloride, and 250 mM imidazole pH 7.5. Proteins were then dialyzed against at least 100X volumes of 1X PBS three times for at least 8 hours each at 4 °C, except for the 37% sequence identity design to 2LV8, which was dialyzed at room temperature. If large molecular weight contaminants were observed, they were removed using a 30 kDa MWCO Amicon concentrator and collecting the flow-through fraction. Protein concentrations were quantified by using a Bradford assay against a dilution series of a 1 mg/mL BSA standard or by measuring absorbance at 280 nm and calculating concentration using extinction coefficients as provided by ExPasy.

Protein purity was assessed by SDS-PAGE. Samples were prepared by mixing protein, 4X Laemmli buffer, 1 mM DTT, and water to a total volume of 15 μL and then denatured by incubation at 95 °C for 10 min. 10 μL of sample was loaded onto a 4-20% Tris-Glycine gel and run at 180 V for 45 min. Gels were stained with Coomassie dye, destained with water, and imaged using a Bio-Rad Gel-Doc EZ Imager.

5.3.2 Size-exclusion chromatography

Chromatography was done using an Agilent 1200 series HPLC attached to a Superdex S200 10/300 GL column and a UV detector. The UV detector was configured to detect signal at 230 nm using 360 nm as a baseline. 0.1 mg or 0.05 mg of protein with a concentration ranging from 0.2 to 3 mg/mL after purification was injected at a flow rate of 0.8 mL/min. All runs were done isocratically using 1X PBS as running buffer. Sizes were estimated using a linear regression model between elution time and molecular weight using BioRad Gel Filtration Standards.

5.3.3 Circular dichroism

Protein samples were diluted to 0.03 - 0.075 mg/mL in 1X PBS and added to a cuvette with a 1 or 2 mm path length. CD was completed using a Jasco J-710 spectrometer measuring from between 200 – 280 nm at a rate of 50 nm / min. Protein concentrations were adjusted as needed to keep the High-Tension Voltage within the linear range of the instrument. A thermal melt was performed from 25 – 95 °C at a ramp rate of 1 °C / min while measuring signal at 220 nm. Proteins were cooled back to 20 °C and were re-scanned to measure secondary structure after the thermal melt.

5.3.4 Low sequence identity designs

Table S4: Protein sequence for low sequence identity design, Top0.

Sequence identity	Protein sequence
0%	MGSSHHHHHHSSGLVPRGSHMMLRINLIVTQENEKLNFNFEISDREKFAAIIKQIEEIVRALNSEKITVEVESKSREQSQRYSEVMEALMKKENFTNLDIKYNNNKIEITATK

Table S5: DNA insert sequence for low sequence identity design, Top0.

Sequence identity	DNA insert sequence
0%	ATGTTACGTATTAACCTGATCGTGACTCAAGAGAATGAGAAACTGAACTTCAACTTCGAGATTTCTGACCGAGAGAAGTTTGCTGCCATTATTAAGCAGATAGAGGAGATCGTGCGTGCCCTGAATAGTGAAAAGATAACGGTCGAAGTCGAGTCGAAGTCAAGAGAACAATCCCAGCGATACTCAGAGGTGATGGAAGCGCTTATGAAGAAGGAGAATTTCACAAACCTGGACATAAAGTACAATAACAACAAGATCGAAATAACGGCCACCAAGTGA