Supporting Information

Fold2Seq: A Joint Sequence(1D)-Fold(3D) Embedding-based Generative Model for Protein Design

1 3D Extension of the Sinusoidal Positional Encoding

We use a simple extension of the sinusoidal encoding described in the original transformer model (Vaswani et al., 2017) to encode each position in our Structure Encoder.

$$\begin{aligned} \mathrm{PE}(x, y, z, 2i) &= \sin(x/10000^{2i/h}) + \sin(y/10000^{2i/h}) + \sin(z/10000^{2i/h}) \\ \mathrm{PE}(x, y, z, 2i+1) &= \cos(x/10000^{2i/h}) + \cos(y/10000^{2i/h}) + \cos(z/10000^{2i/h}) \end{aligned} \tag{1}$$

2 Comparison between Two Training Strategies

In this section, we compare the performance between one-stage training and two-stage training strategies. In the one-stage strategy, we train our model through the 5 loss terms in Eq (4) together. While in the two-stage strategy, we first train our model using L_1 and then train using L_2.
We first compare the learning curves. As the $\mathbf{RE}_f$ loss represents the quality of the model, we plot the $\mathbf{RE}_f$ loss vs epochs on both training and validation sets.

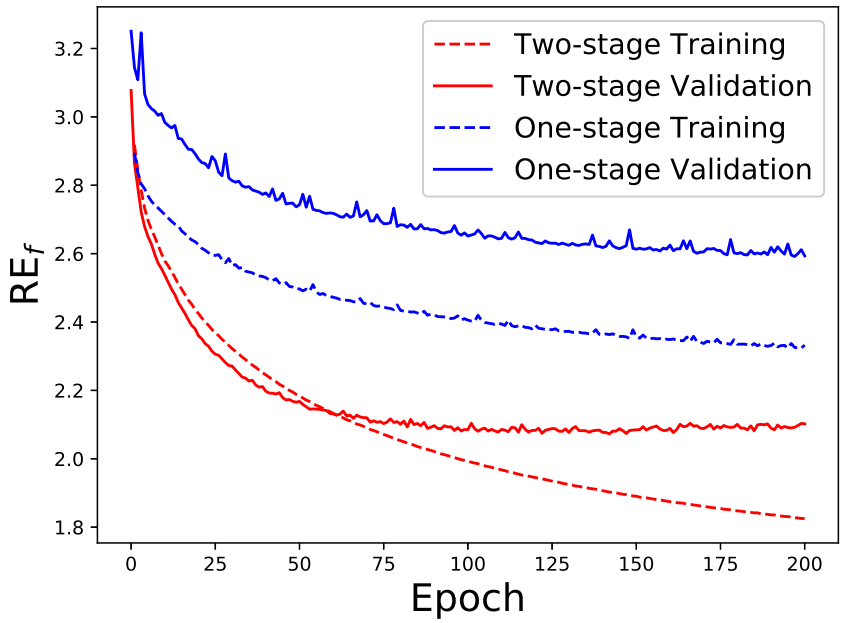

Figure S1: The fold2seq loss($\mathbf{RE}_f$) curves of two training strategies on training and validation set.

As shown in Fig S1, the two-stage strategy significantly outperforms the one-stage strategy. To further demonstrate this point, we calculate the per-residue perplexity and the average sequence recovery rate on

Table S1: Performance of two training strategies assessed in the sequence domain.

(a) Avg. $ppl_{\text{fold}}(i)$ (std. dev.) (%).

Model	ID Test	OD Test
Uniform	20.0	20.0
Natural	18.0	18.0
One-stage strategy	13.1 (4.3)	15.3 (3.2)
Two-stage strategy	**9.0** (5.3)	12.0 (2.4)

(b) Avg. $sr_{\text{fold}}(i)$ (std. dev.) (%).

Model	ID Test	OD Test
Random across two folds	12.8 (7.9)	12.8 (7.9)
One-stage strategy	22.2 (4.3)	20.3 (3.2)
Two-stage strategy	**27.2 (6.3)**	**25.2 (3.2)**
Random within same fold	39.1 (9.4)	39.1 (9.4)

the two test sets. As shown in Table S1, the same conclusion can be drawn. These results validate our design choice of a two-stage training strategy.

3 Dataset Statistics

The statistics of our various datasets are given below.

- Training set includes 45995 proteins belonging to a total of 971 folds.
- Validation set includes 4159 proteins belonging to a total of 185 folds.
- In-distribution (ID) test set includes 1131 proteins belonging to a total of 181 folds.
- Out-of-distribution (OD) test set includes 203 proteins belonging to a total of 27 folds.

4 Sequence Identity Measurement

Sequence identity is measured through the Needleman–Wunsch algorithm (Needleman & Wunsch, 1970) with the Blossum62 scoring matrix.

5 Structure Level Performance Metrics

While we reported fold level performance metrics in the main paper, we also report the corresponding structure level metrics below. Fold2Seq outperforms all other methods except Graph_trans in $ppl_{\text{structure}}(i)$ and $sr_{\text{structure}}(i)$. Note that Graph_trans has an inherent advantage here because it uses the entire structure, where as Fold2Seq only uses high level fold information. Similar to fold based metrics, Fold2Seq performs better in the Missing Residues experiment and can also handle NMR Structural Ensembles.

Table S2: Performance of different methods assessed in the sequence domain.

(a) Avg. $ppl_{\text{structure}}(i)$ (std. dev.)

Model	ID Test	OD Test
Uniform	20.0	20.0
Natural	18.0	18.0
cVAE	14.8	16.3
gcWGAN	13.5	15.2
Graph_trans	**7.3**	**10.3**
Fold2Seq	8.1	11.9

(b) Avg. $sr_{\text{structure}}(i)$ (std. dev.) (%).

Model	ID Test	OD Test
Random across two folds	12.8 (7.94)	12.8 (7.94)
cVAE	17.7 (7.34)	15.3 (5.34)
gcWGAN	17.5 (6.35)	14.1 (3.45)
RosettaDesign	20.3 (5.13)	20.2 (2.98)
Graph_trans	**29.3 (4.3)**	**27.4 (3.2)**
Fold2Seq	27.1 (6.31)	24.1 (2.64)
Random within same fold	39.1 (9.35)	39.1 (9.35)

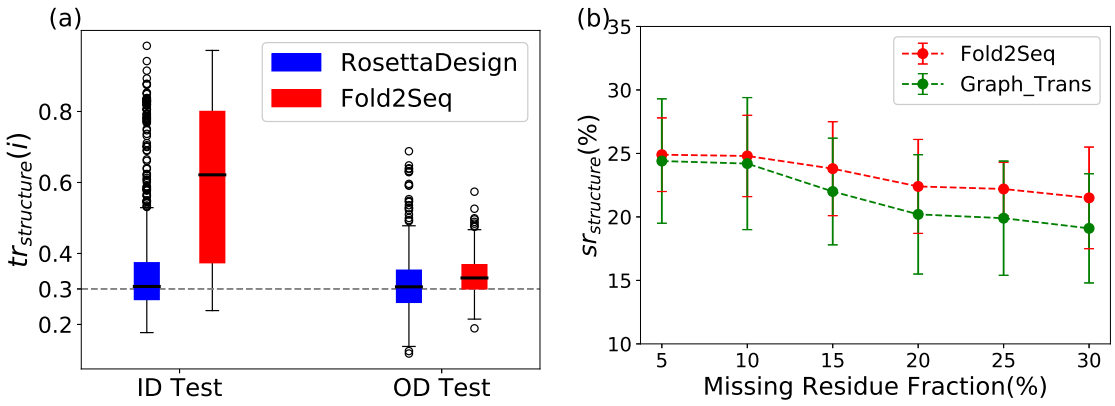

Figure S2: (a). $tr_{\text{structure}}(i)$ distributions of RosettaDesign and Fold2Seq. (c). Avg. $sr_{\text{structure}}$ (%) for the OD test set with a string of missing residues.

Table S3: Avg. $sr_{\text{structure}}(i)$ (std. dev.) (%). for NMR ensemble.

NMR Input	ID	OD
Single	22.7 (3.4)	20.9 (4.2)
Ensemble	24.1 (4.6)	22.3 (3.1)

6 Generalizability Analysis of Performances

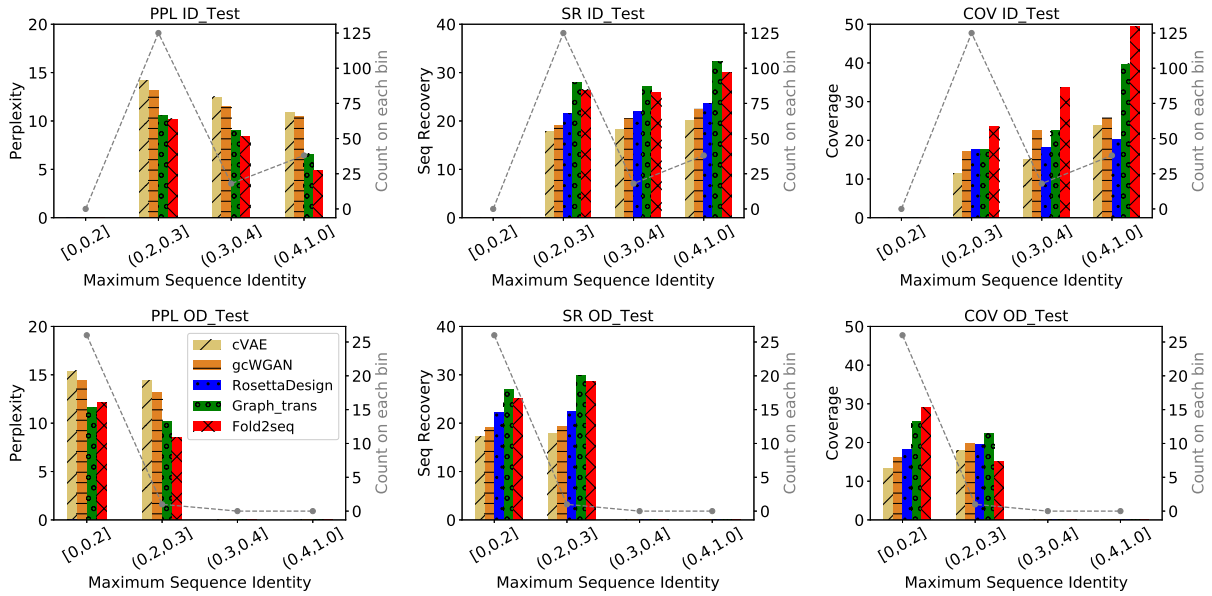

Figure S3: The ppl_{fold}, sr_{fold} and cov_{fold} performances of different models in three ontologies, over 4 bins of increasing sequence-identity ranges. Low sequence identity indicates low similarity between sequences in a test fold and the training set. Sequence statistics over the bins (gray dots connected in dashed lines) are also provided.

7 Comparison of Folded Structures

In this section, we show some representative folded structures whose sequences are designed by RosettaDesign and Fold2Seq. The folded structures were predicted using iTasser, a state of the art program for protein structure prediction. Figure S4 shows some structures where Fold2Seq performs better than RosettaDesign and Figure S5 shows some structures where RosettaDesign is better.

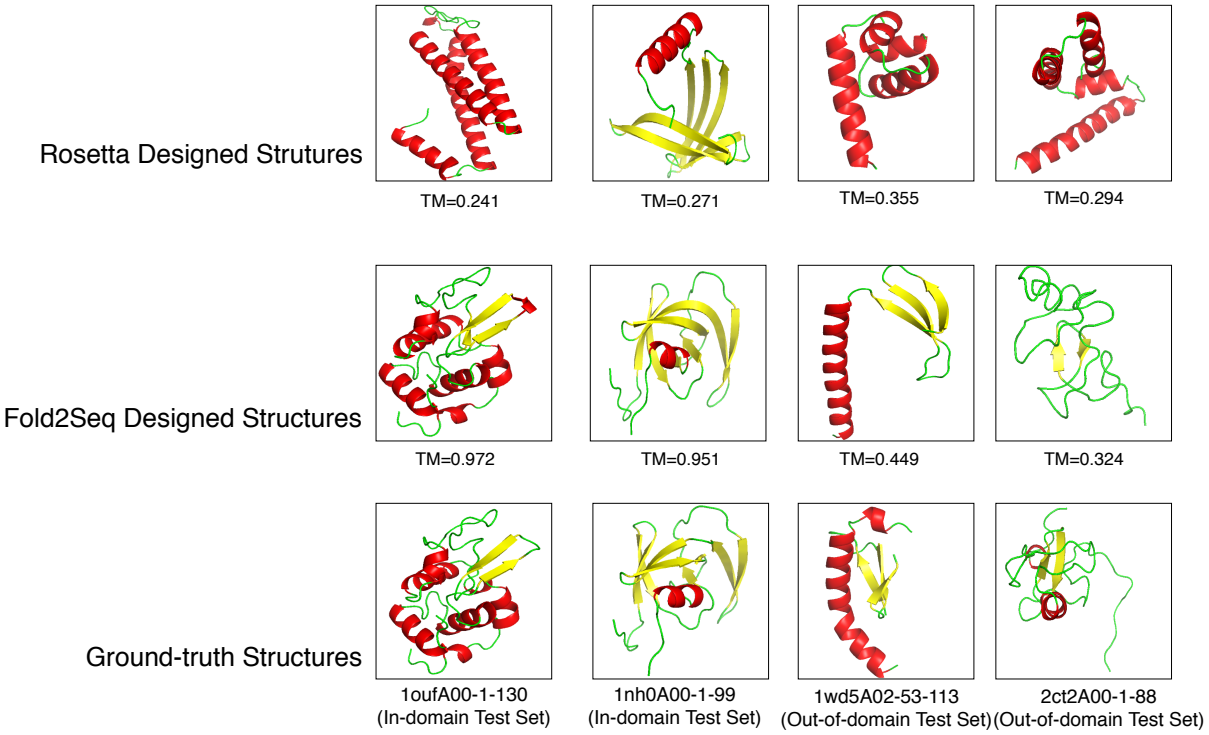

Figure S4: The native and designed structures in the folds with $\Delta tr_{\text{fold}} > 0$. The IDs at the bottom are the CATH domain names of each structure.

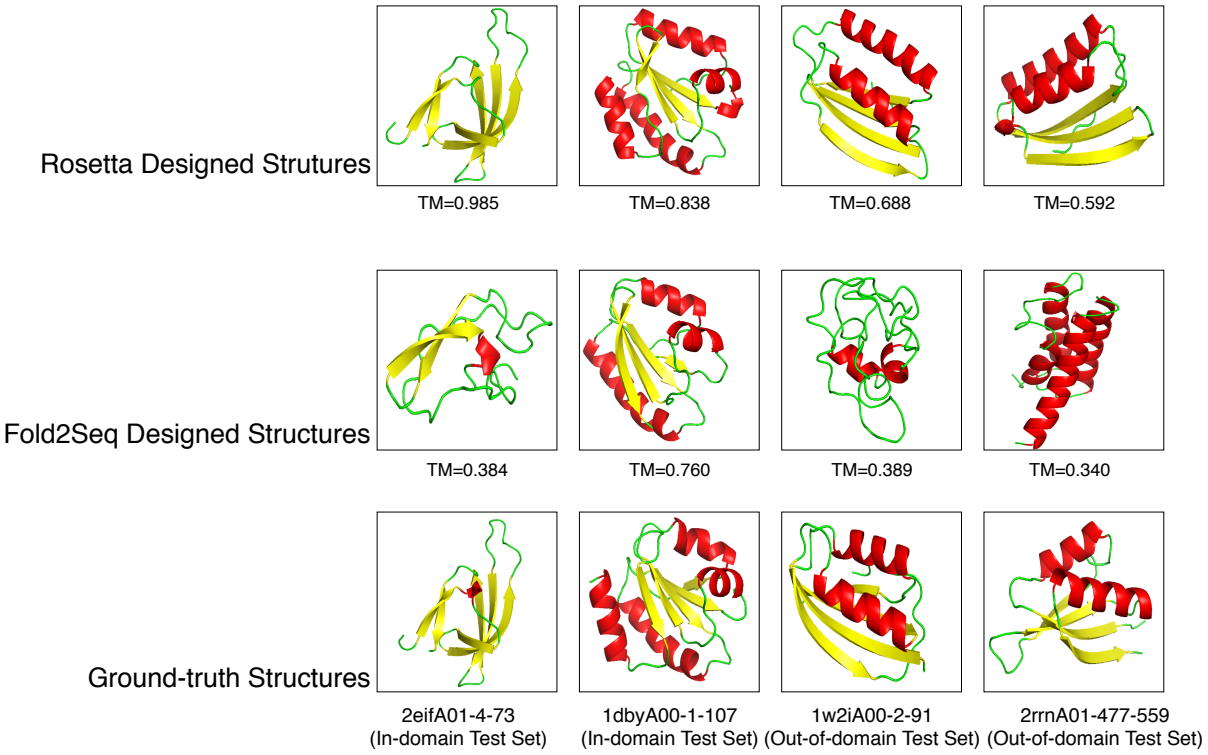

Figure S5: The native and designed structures in the folds with $\Delta tr_{\text{fold}} < 0$. The IDs at the bottom are the CATH domain names of each structure.

8 t-SNE Visualization of Fold/Structure Embeddings

We use t-SNE to visualize the fold embeddings $\boldsymbol{h}$ from Fold2Seq and Graph_trans for the proteins in the OD test set (see Figure S6). We show that Fold2Seq captures the similarity and diversity within the fold space better and that the embeddings from proteins belonging to the same fold are better clustered in Fold2Seq.

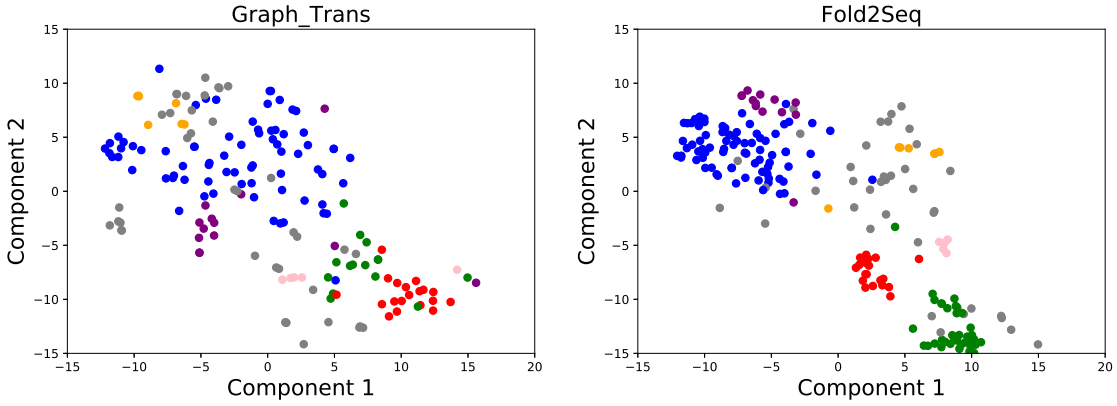

Figure S6: The t-SNE visualization of the averaged structure (fold) latent embeddings $\boldsymbol{h}$ by two methods on the OD test set. Each protein is colored by its fold category. Same color indicates the same fold, except that gray points represent outliers, which is defined by its fold having < 5 proteins in the test set.

9 Ablation Study

We performed an ablation study to delineate the contributions from different components of the algorithm. The details of the different ablations are given below.

- cVAE: We use cVAE (Greener et al., 2018) as baseline with 1D string fold representation and MLP-based VAE.
- Trans_string_$\mathbf{RE}_f$: We replace the MLP-based VAE in cVAE with transformer autoencoder model. The loss is $L = \mathbf{RE}_f$.
- Trans_voxel_$\mathbf{RE}_f$: We replace the 1D string fold representation in "Trans_string_$\mathbf{RE}_f$" with 3D voxel representation. We also add the convolutional residual block and 3D positional encoding. The loss is $L = \mathbf{RE}_f$.
- +$\mathbf{RE}_s$+**CS**: We add the sequence encoder, together with the reconstruction loss and the cosine similarity loss to the previous loss: $L = \lambda_1\mathbf{RE}_f + \lambda_2\mathbf{RE}_s - \lambda_5\mathbf{CS}$.
- +2**FC**: We add the two **FC** losses. $L = \lambda_1\mathbf{RE}_f + \lambda_2\mathbf{RE}_s + \lambda_3\mathbf{FC}_f + \lambda_4\mathbf{FC}_s - \lambda_5\mathbf{CS}$.
- +**CY** (Fold2Seq): We add the cyclic loss into the former model with the final loss $L = \lambda_1\mathbf{RE}_f + \lambda_2\mathbf{RE}_s + \lambda_3\mathbf{FC}_f + \lambda_4\mathbf{FC}_s + \lambda_5(\mathbf{CY} - \mathbf{CS})$.

The key results of the ablation study are summarized in Table 3(a) and Section 4. Overall, the string to voxel change, the addition of 2 **FC** losses and the cyclic loss gives us a significant performance boost.

References

Joe G Greener, Lewis Moffat, and David T Jones. Design of metalloproteins and novel protein folds using variational autoencoders. *Scientific reports*, 8(1):1–12, 2018.

Saul B Needleman and Christian D Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453, 1970.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.