A MORE ANALYSIS

Module-wise training Regarding to the module-wise training strategy, we employ a stagewise fashion for sequence refinement. This is represented by the equation $\boldsymbol{s}^{(l)} = f_{\phi^{(l)}}(\boldsymbol{s}^{(l-1)}; \boldsymbol{z}^{(l-1)})$, where $\boldsymbol{s}^{(l)}$ represents the refined sequence generated by the l-th refinement module $f_{\phi^{(l)}}$, and $\boldsymbol{z}^{(l-1)}$ denotes the pretrained embedding extracted from $\boldsymbol{s}^{(l-1)}$. The module-wise training could significantly reduce the computational cost and helpful for alleviating oversmoothing:

1. Stantard: For a given protein, during each epoch, the parameters of layers l and $l+1$ are updated, resulting in a change in $\boldsymbol{s}^{(l)}$. Due to the large size of the pretrained model (ESM2-650M), performing a forward pass to obtain the updated version of $\boldsymbol{z}^{(l)}$ consumes significant GPU memory and increases the time cost. Consequently, conducting end-to-end training would significantly increase the computational cost.
2. Ours: By fixing $\theta^{(l)}$ before training $\theta^{(l+1)}$, we can ensure that $\boldsymbol{s}^{(l)}$ is retrained without the need for updating $\boldsymbol{z}^{(l)}$. This approach allows us to initialize $\boldsymbol{s}^{(l)}$ in the first epoch and reuse it in subsequent epochs, effectively avoiding redundant forward passes of the large pretrained model.
3. The overall KWDesign is $l * 10$-layer GNN model, where each refienment module consists of 10 GNN layers. It is well known that GNN models suffer from the issue of oversmoothness when training in an end-to-end fashion. To overcome this, we adopt module-wise training to ensure that the pre-placed module serves as a good initialization for the subsequent module. As discovered by (Pina & Vilaplana, 2023), the layer-wise training promotes the node features to be uncorrelated at each single layer to alleviate oversmoothing.

Distribution comparison Fig. 4 shows the confidence distributions of positive and negative residues generated by PiFold and KW-Design on the CATH4.2 test set. Positive residues tend towards a confidence of 1.0, while negative residues have mostly below 0.6 confidence, indicated by different colors. Our results demonstrate that KW-Design produces positive residues with higher confidence compared to PiFold, while also reducing the number of negative residues. This suggests that KW-Design can convert low-confidence positive residues to high-confidence ones and correct negative residues as positive ones.

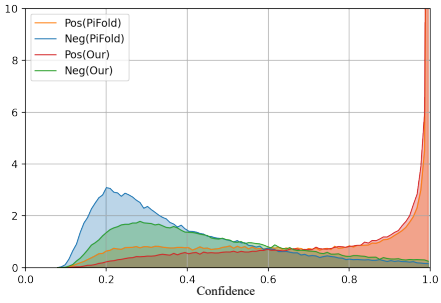
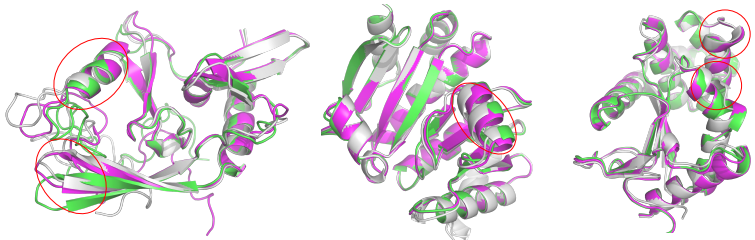


Figure 4: Confidence distributions.

Compare structures In Fig.5, we use ESMFold(Lin et al., 2022) to generate protein structures from designed sequences and compare the designed proteins of PiFold and KW-Design against the reference ones. We observe that the designed structures of KW-Design are more similar to the reference ones than those of PiFold. Specifically, KW-Design achieves 15.9%, 35.3%, and 60% improvement in structural mean squared error (MSE) on the 1a73, 1a81, and 1ac1 proteins, respectively. These results demonstrate that KW-Design can generate proteins that are structurally more similar to the reference ones compared to PiFold.



(a) PDBID: 1a73
PiFold MSE: 0.63, Our: 0.50

(b) PDBID: 1a81
PiFold MSE: 0.34, Our: 0.26

(c) PDBID: 1ac1
PiFold MSE: 0.75, Our: 0.30

Figure 5: Comparing the designed proteins. The green structures are reference ones, while the gray and purple structures are designed by PiFold and KW-Design, respectively. We use red circles to highlight the regions where KW-Design produces more similar structures to the reference ones than PiFold.

B DISCUSSION ABOUT DIVERSITY

TS45 In addition to designing single- and multi-chain proteins, we also include a set of *de novo* proteins collected from the CASP15 competition to provide a more realistic assessment (Senior et al., 2019; Kinch et al., 2021). The Critical Assessment of Protein Structure Prediction (CASP15), which took place from May through August 2022, was held after the release dates of CATH4.3 (July 1, 2019) and PDB (August 2, 2021). In CASP15, diverse protein targets are introduced, including FM (Free Modeling), TBM (Template-Based Modeling), TBM-easy, and TBM-hard proteins. There are 18 FM, 25+2 TBM (including 20 TBM-eazy, 5 TMB-hard, 2 FM/TBM). The FM targets have no homology to any known protein structure, making them particularly suitable for *de novo* protein design. The TBM targets have some homology to known protein structures, while the TBM-easy targets are relatively easy TBM targets. The TBM-hard targets are more difficult TBM targets, with lower levels of sequence identity to known structures. We download the public TS-domains structures from CASP15 which consists of 45 structures, namely TS45. We use TS45 as a benchmark for *de novo* protein design, as the structures are less similar to known structures and were not determined prior to the construction of the training sets.

Diversity Definition To improve the success rate of protein design, it is important to explore a set of protein sequences rather than placing a bet on a single sequence. In this case, generating diverse sequences is crucial for exploring the reasonable protein sequence space. We define the pairwise diversity (Jain et al., 2022) as $D_{ij} = \frac{\sum_{l=1}^{n} \mathbb{1}_{r_{i,l} \neq r_{j,l}}}{n}$, where $r_{i,l}$ indicates the l-th residue of the i-th designed sequence. The overall diversity score is

$$\texttt{Div} = \sum_{i,j} \frac{D_{i,j}}{m^2} \tag{14}$$

where $i, j \in \{1, 2, 3, \cdots, m\}$ and m is the number of totally designed sequences. By default, we set $m = 10$. However, measuring diversity alone without combining it with other metrics may be misleading. For example, a high diversity indicates a low recovery rate, more likely to result in a low structural similarity.

Experiments & Analysis We benchmark the diversity on TS45 dataset using models pre-trained on CATH4.3. As discovered by previous research (Hsu et al., 2022; Dauparas et al., 2022), the sampling temperature affects diversity. Denote the temperature as T, the predicted probability vector is $\boldsymbol{p} \in \mathbb{R}^{n,20}$, we sample new sequences from the distribution of $\texttt{Multinomial}(\texttt{softmax}(\boldsymbol{p}/T))$. We vary the temperature from 0.0 to 0.5 and plot the trends of recovery and diversity in Fig. 6. Under the same sampling temperature, high recovery leads to decreased diversity. **However, at the same level of recovery, stronger models have higher diversity.** This phenomenon can be attributed to the fact that stronger models, such as KWDesign, PiFold, and ProteinMPNN, exhibit relatively high confidence levels for a larger number of residues. Even after applying smoothing to the probability distribution, the recovery rate remains consistently high, while a noticeable improvement is observed in terms of diversity.

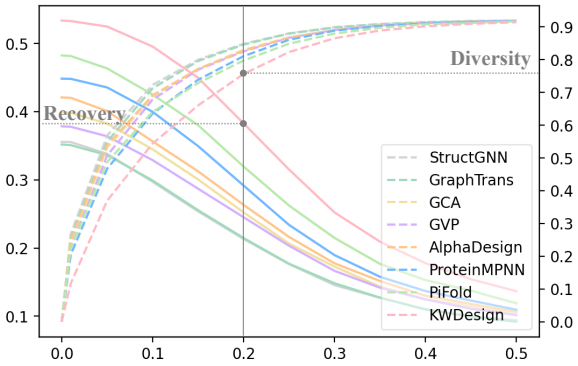


Figure 6: The trends of recovery and diversity.

C MORE COMPARISON

Compare to LMDesign In parallel with our work, we are observing another exciting project called LMDesign (Zheng et al., 2023), which was recently published at ICML as an oral presentation. LMDesign aims to use the pre-trained ESM model to improve protein design. However, there are several differences between our knowledge-Design and LMDesign.

- **More comprehensive**: We enhance protein design by fusing multimodal knowledge from pre-trained models, including both structural and sequential information, while LMDesign only uses single-modal information. Our experiments demonstrate that combining these modalities leads to nontrivial improvements, as shown in Table 5
- **More efficient**: We introduce the memory-retrieval mechanism to save more than 50% of the training time, while LMDesign does not use this mechanism.
- **Novel modules**: We introduce confidence-aware recycling techniques as well as virtual MSA to boost the model performance.
- **More effective**: Overall, our model outperforms LMDesign by 5.12% on the CATH4.2 dataset.

Table 6: Results comparison on the CATH dataset. All baselines are reproduced under the same code framework, except ones marked with †. We copy results of GVP-large and ESM-IF from their manuscripts (Hsu et al., 2022). The **best** and <u>suboptimal</u> results are labeled with bold and underline.

Model	Perplexity ↓			Recovery % ↑			CATH version	
	Short	Single-chain	All	Short	Single-chain	All	4.2	4.3
StructGNN	8.29	8.74	6.40	29.44	28.26	35.91	✓	
GraphTrans	8.39	8.83	6.63	28.14	28.46	35.82	✓	
GCA	7.09	7.49	6.05	32.62	31.10	37.64	✓	
GVP	7.23	7.84	5.36	30.60	28.95	39.47	✓	
GVP-large†	7.68	<u>6.12</u>	6.17	32.6	39.4	39.2		✓
AlphaDesign	7.32	7.63	6.30	34.16	32.66	41.31	✓	
ESM-IF†	8.18	6.33	6.44	31.3	38.5	38.3		✓
ProteinMPNN	6.21	6.68	4.61	36.35	34.43	45.96	✓	
PiFold	<u>6.04</u>	6.31	4.55	<u>39.84</u>	38.53	51.66	✓	
LMDesign	6.77	6.46	<u>4.52</u>	37.88	<u>42.47</u>	<u>55.65</u>	✓	
Knowledge-Design (Ours)	**5.48**	**5.16**	**3.46**	**44.66**	**45.45**	**60.77**	✓	

sc-TM The structural similarity is the ultimate standard for measuring the quality of the designed sequence. However, the structures of designed protein sequences needed to be predicted by other algorithms, such as AlphaFold (Jumper et al., 2021), OmegaFold (Wu et al., 2022) and ESMFold (Lin et al., 2022). The protein folding algorithm itself has a certain inductive bias and will cause some prediction errors, which will affect the evaluation. To overcome the inductive bias, we introduce the self-consistent TM-score (sc-TM) metric:

$$\texttt{sc-TM} = \texttt{TMScore}(f(\hat{\mathcal{S}}), f(\mathcal{S})) \tag{15}$$

where f is the protein folding algorithm and $\texttt{TMscore}(\cdot, \cdot)$ is a widely used metric (Zhang & Skolnick, 2005) for measuring protein structure similarity. Since the structures of the designed sequence and reference sequence are predicted by the same protein folding algorithm, the model's inductive bias is expected to be canceled out when calculating the TM-score. This approach results in a more robust metric, called the sc-TM, that is less affected by the inductive bias of the protein folding algorithm.

sc-TM Results On the CASP15 dataset, we investigate the recovery and sc-TM metrics as the temperature increases. To compute sc-TM, we utilize AlphaFold to predict protein structures from sequences. According to Fig.7 and Fig.6, we observe that a slight increase in temperature from 0 to 0.1 is beneficial in significantly enhancing diversity while maintaining good recovery and sc-TM. Specifically, when the sampling temperature is set to 0.0 and 0.1, KWDesign outperforms the baselines, achieving the highest sc-TMScore.

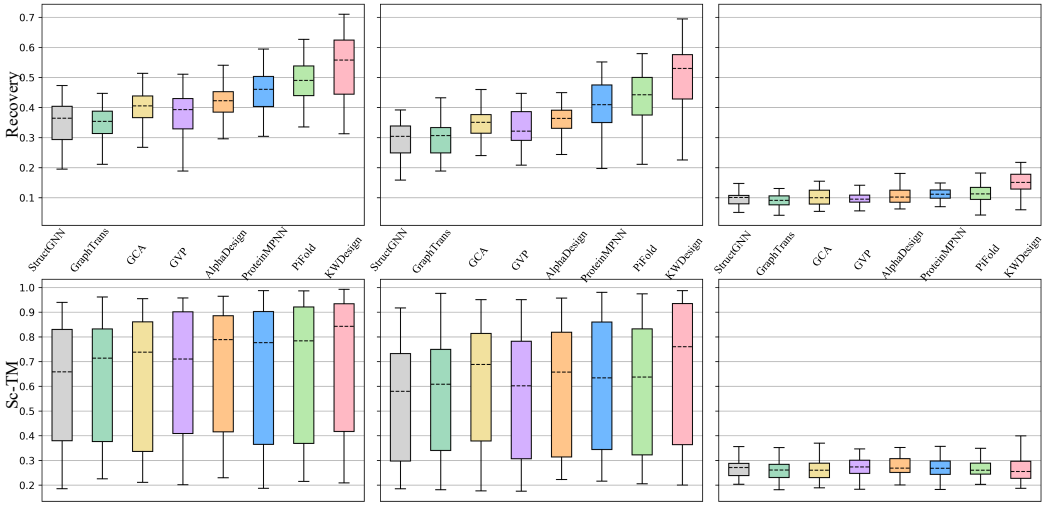


Figure 7: The statistics of recovery and sc-TM with increasing temperature.

Do pretrained language models prefer to correct disordered regions? Thanks to the insightful comments of Reviewer 4EsM, we investigate the preferred regions that pretrained language models tend to correction. Expeirments follow five steps:

1. Select PDBs containing diverse protein structures, run the pretrained PiFold to get the designed sequence S_{pi}
2. Further use the pretrained KWDesign to refine S_{pi}, resulting in the optimized sequence S_{kw}. All the used PiFold and KWDesign are pretrained on CATH4.3.
3. For each residue, we record whether PiFold has successfully recovered the residue type ($good_{pi}$) or not (bad_{pi}). Similarly, we record $good_{kw}$ and bad_{kw}.
4. We list four states for each residues and set a different color for each state: $good_{pi} \rightarrow bad_{kw}$(red), $bad_{pi} \rightarrow bad_{kw}$(gray), $good_{pi} \rightarrow good_{kw}$(white), $bad_{pi} \rightarrow good_{kw}$(green).
5. We show the colored proteins with PyMol and present them in Fig.8.

Our finding is that pretraining knowledge corrects for both disordered and ordered regions. However, it seems that the correction is more pronounced for regions with well-defined secondary structures (α-helix and β-sheet).

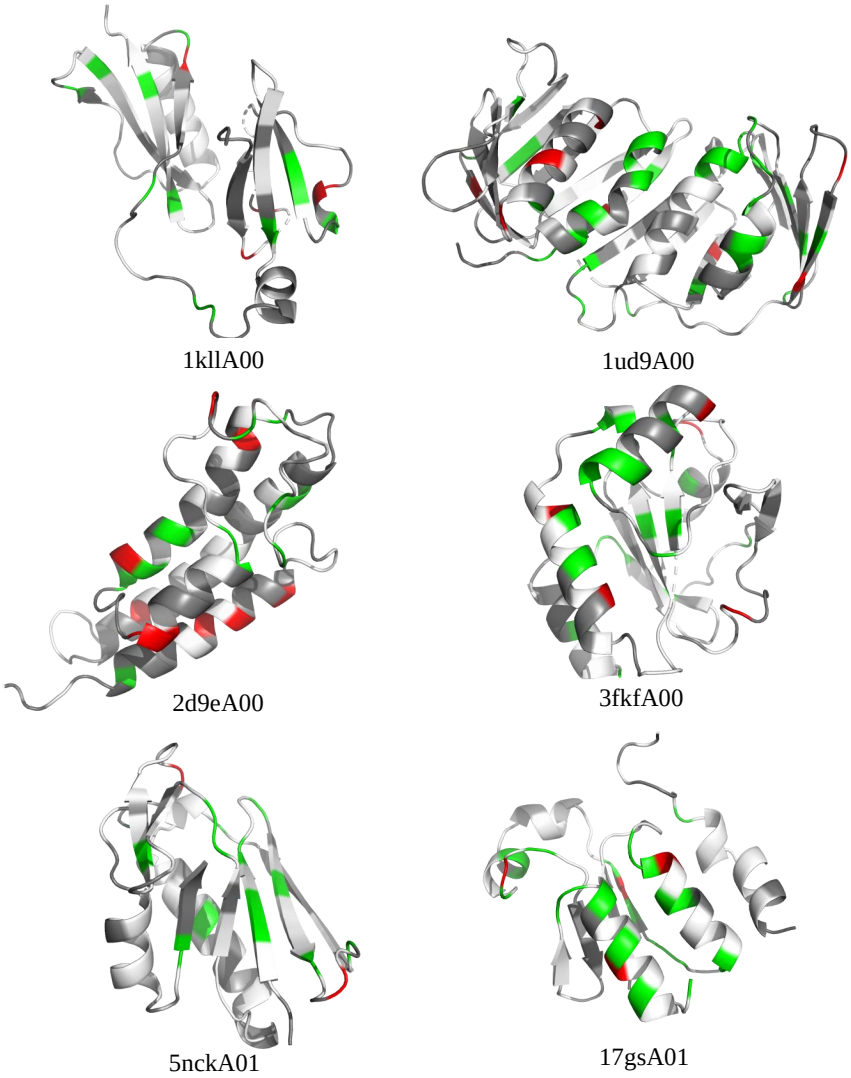


Figure 8: The colored proteins.