pLM-Guided Inverse Folding for Antibody Sequence Design

Anonymous Authors[1]

Abstract

Inverse folding, predicting amino acid sequences from three-dimensional structures, is a foundational task in computational protein design, yet it is hindered by the scarcity of structural data, which limits model training and risks overfitting. The standard approach fine-tunes general inverse folding models on domain-specific structural datasets like antibodies, but such data remain expensive. To enable inverse folders to benefit more from abundant sequence data, we propose combining ProteinMPNN, a general protein inverse folding model, with IgLM, an antibody-specific language model, via a training-free weighted ensemble of their predictions at inference time. Evaluated on antibody and nanobody structures, our results show that this approach substantially improves amino acid recovery over ProteinMPNN alone, approaching the performance of antibody-specific models like AntiFold while generating more diverse sequences. Even models already fine-tuned on antibody structures (AbMPNN) benefit from language model guidance, demonstrating that it complements structural fine-tuning and leads to more natural-looking sequences that still satisfy structural constraints.

Introduction

Inverse folding, the task of predicting an amino acid sequence conditioned on a three-dimensional backbone, remains a foundational component of modern *in silico* protein design pipelines (Watson et al., 2023; Frank et al., 2024; Pacesa et al., 2025). In these workflows, a backbone is generated first, followed by inverse folding to derive a compatible sequence, and subsequent validation steps, such as refolding to assess structural self-consistency. Despite recent progress toward end-to-end all-atom protein generation (Stark et al., 2025; Butcher et al., 2025), inverse folding continues to play a critical role as a complementary model for refining and validating sequence designs.

A central limitation of inverse folding, and protein design more broadly, is the scarcity of experimentally determined protein structures, which remain costly and time-consuming to obtain (Slabinski et al., 2007; Ding et al., 2022). To mitigate this, several methods augment training data with synthetic structures generated *in silico* by folding models (Hsu et al., 2022; Dreyer et al., 2023; Høie et al., 2024). Because sequence data are far more abundant than structural data, this enables large-scale dataset expansion, typically filtered by folding confidence scores. However, reliance on synthetic structures may bias inverse folding models toward algorithmically favorable and less natural sequence distributions, limiting diversity and deviating from true biological variability.

An alternative way to exploit abundant sequence data is to incorporate language models trained on large protein sequence corpora (Lin et al., 2023; Nijkamp et al., 2023; Ferruz et al., 2022). In the antibody and nanobody domain, such data are particularly abundant due to extensive repertoire sequencing efforts (Olsen et al., 2022). Coupling antibody language models with inverse folding has the potential to improve prediction accuracy while preserving features characteristic of natural antibodies, including favorable developability properties such as solubility, low aggregation propensity, and reduced immunogenicity (Raybould et al., 2019).

Here, we propose combining an antibody-specific language

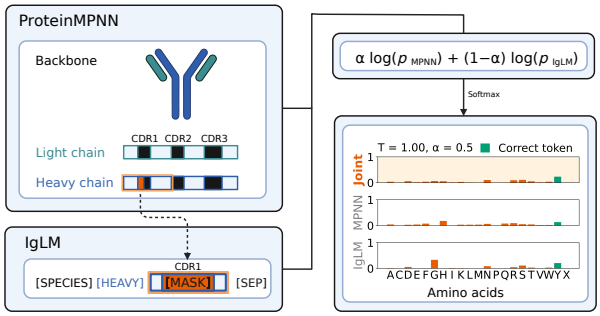

Figure 1. Overview of model ensembling. ProteinMPNN (or AbMPNN) uses the full antibody/nanobody structure, including the target chain and binder framework regions. IgLM predicts one chain at a time, conditioning on the surrounding framework sequence and previously sampled residues for the current CDR. The residue being sampled is highlighted in orange. Log-probabilities from both models (see Methods) are combined and passed through a softmax with temperature T to produce a categorical distribution for sampling the next residue.

model, which captures natural sequence distributions, with an inverse folder that enforces structural constraints, as illustrated in Figure 1. Instead of fine-tuning inverse folding models on antibody structures, we integrate both components at inference time via ensemble sampling, enabling an effective combination of their complementary signals. We evaluate this approach on highly variable CDR regions, assessing its impact on sequence recovery, structural consistency, and sequence naturalness. By leveraging abundant sequence data instead of limited structural information, this strategy can readily benefit from ongoing advances in language modeling.

Related Work

Protein Inverse Folding leverages diverse machine learning architectures, including graph neural networks (e.g., ProteinMPNN (Dauparas et al., 2022)), sequence-based Transformers (e.g., ESM-IF1 (Hsu et al., 2022)), structure-aware graph Transformer models (e.g., PiFold (Gao et al., 2022)), and discrete diffusion approaches, to predict sequences from structural templates (Ektefaie et al., 2024; Bai et al., 2025). Specialized variants, fine-tuned for properties like solubility (Goverde et al., 2024) or specific protein families, further enhance performance. Antibody-focused inverse folders, including AbMPNN (Dreyer et al., 2023) and AntiFold (Høie et al., 2024), are derived from general protein inverse folders, by continued fine-tuning on antibody datasets. These approaches supplement training with synthetic structures, e.g., from ABodyBuilder2 (Abanades et al., 2023), whereas others, like IgDesign (Shanehsazzadeh et al., 2023), rely exclusively on experimentally determined structures. Discrete diffusion models like AntiDIF (Branson & Deane, 2025) and RADAb (Wang et al., 2024) increase prediction diversity while maintaining high sequence recovery rates.

Protein Language Models draw heavily from advances in natural language processing (NLP), with two main paradigms: masked language models (e.g., the ESM family (Lin et al., 2023)) and autoregressive models (e.g., ProtGPT2, ProGen2 (Ferruz et al., 2022; Nijkamp et al., 2023)). These models have been applied to a range of tasks, including sequence annotation, (partial) sequence generation and evaluation, and classification of biological properties. Additionally, they serve as efficient alternatives to time-consuming multiple sequence alignments in protein structure prediction pipelines (Fang et al., 2023). A current trend is the integration of structural and sequence information into multimodal language models, such as ESM3 (Hayes et al., 2025). In the antibody domain, masked language models (e.g., AntiBERTa, AbBERT (Leem et al., 2022; Vashchenko et al., 2022)) are more prevalent and have demonstrated effectiveness in downstream tasks such as paratope prediction (Kalemati et al., 2024). Autoregressive models, such as IgLM (Shuai et al., 2023), generate sequences residue by residue and are designed to produce antibody sequences while assigning likelihood-based scores that reflect their naturalness, making them well-suited for both sampling and sequence scoring.

Integration of Language Models with Inverse Folding has already been explored in the past. One prominent approach is LM-Design (Zheng et al., 2023), a framework that combines ProteinMPNN's structural encoding with a language model (specifically ESM2). This integration requires retraining the language model. While LM-Design is presented as a general framework rather than a specific model, it has been fine-tuned on antibody structure data, as shown by IgDesign (Shanehsazzadeh et al., 2023). Our approach differs from this method in that it requires no retraining, operates purely at inference time, and leverages an antibody-specific rather than a general protein language model. Modern multimodal models such as ESM3 can perform inverse folding without additional training, as they jointly model sequence and structure and are trained on large-scale protein sequence datasets, including antibodies.

Method

To study ensembling with an antibody language model during inference, we rely on ProteinMPNNN as the inverse folding component. ProteinMPNN is a compact and computationally efficient inverse folding model with strong performance across diverse protein design tasks and broad adoption in design pipelines (Watson et al., 2023; Frank et al., 2024; Pacesa et al., 2025). We also include AbMPNN, a fine-tuned version of ProteinMPNN trained specifically on antibody structures, to assess whether ensembling remains effective when the structure-based model is already antibody-specific.

For the language model component, we choose IgLM, which is specifically trained on antibody sequences. IgLM supports autoregressive and bidirectional sequence generation and aligns well with ProteinMPNN's sampling strategy. In addition, IgLM's log-likelihood scores correlate with ProteinMPNN's amino acid recovery rates (see Appendix A.3), indicating that it can distinguish between high- and low-quality ProteinMPNN predictions and supports its use in ensembling. By modeling natural antibody sequence distributions, IgLM encourages antibody-like predictions.

ProteinMPNN consists of a backbone encoder and a sequence decoder, trained jointly to maximize the likelihood of sequences given their corresponding structures. During inference, the sequence decoder generates residues autoregressively without a fixed positional order, conditioning each prediction on the backbone structure and previously sampled residues. Residues are sampled from a multinomial

distribution, with stochasticity controlled by a temperature parameter.

IgLM, which is built on the GPT-2 architecture (Radford et al., 2019), likewise generates amino acid sequences through autoregressive sampling. IgLM captures bidirectional sequence relationships by masking specific regions of the input sequence and predicting the missing residues in the context of the entire sequence, with the predicted residues appended to the input sequence. Nevertheless, it requires a left-to-right sampling process within one masked-out region. We therefore restrict sampling of ProteinMPNN to a left-to-right order within each chain. This does not affect ProteinMPNN's sampling quality (see Appendix A.3).

To merge the predictions of ProteinMPNN and IgLM, we denote their log-probability outputs at decoding step t as

$$\ell_{\mathrm{IgLM}}(s_t) = \log P_{\mathrm{IgLM}}(s_t \mid s_{<t}),$$
$$\ell_{\mathrm{MPNN}}(s_t) = \log P_{\mathrm{MPNN}}(s_t \mid s_{<t}, X)$$

where s_t denotes the amino acid type at step t, and $s_{<t}$ the amino acids sampled at all previous steps, given the left-to-right decoding order. X denotes the input structure.

For IgLM, we mask only the CDR under design and provide the flanking framework residues as context.

The ensemble is formed in log-probability space via a convex combination (with coefficient $0 \leq \alpha \leq 1$):

$$\ell_{\mathrm{ens}}(s_t) = \alpha\, \ell_{\mathrm{MPNN}}(s_t) + (1-\alpha)\, \ell_{\mathrm{IgLM}}(s_t).$$

The ensemble distribution is obtained by applying a softmax with temperature $\tau > 0$ over all amino acids,

$$P_{\mathrm{ens}}(s_t) = \frac{\exp(\ell_{\mathrm{ens}}(s_t)/\tau)}{\sum_a \exp(\ell_{\mathrm{ens}}(a)/\tau)}.$$

The ensemble can be interpreted as a weighted product-of-experts combination of the model predictions. While the models are not strictly independent due to shared sequence conditioning, this formulation empirically improves sampling performance, as shown in Figure 2 for antibodies and Figure 14 for nanobodies.

The parameter α controls the weighting between the two models, with $\alpha = 0.5$ representing a complete balance between them. We selected α empirically based on performance on the validation set.

Results

We evaluate whether ensembling a general protein inverse folding model (ProteinMPNN) with IgLM can match the performance of antibody-specific models (AbMPNN), and whether AbMPNN also further benefits from ensembling.

ProteinMPNN+IgLM and AbMPNN+IgLM are compared against their base models (ProteinMPNN and AbMPNN)

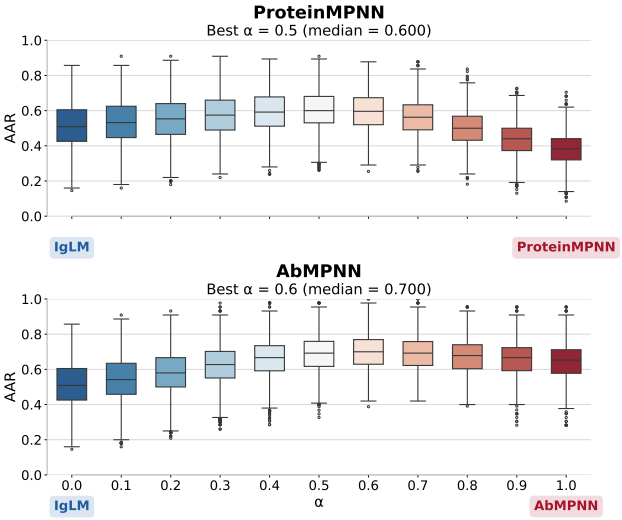

Figure 2. Amino acid recovery rates of CDR loops across different α values for ProteinMPNN or AbMPNN ensembles with IgLM, evaluated on antibody validation structures at sampling temperature 0.25 with 10 predictions per structure.

and other inverse folding methods on CDR loop inpainting, targeting the complementarity-determining regions (CDR loops), the most variable regions of antibodies and nanobodies that form the antigen-binding site. Evaluation is performed on a test set of 196 antibody and 93 nanobody structures from the PDB released after ESM3. To reduce redundancy and prevent data leakage, we filter the set such that no pair of structures exceeds 90% sequence similarity with each other, and no structure exceeds 90% sequence similarity to any previously released antibody or nanobody structures available before the cutoff date.

The ensemble weight α is optimized using amino acid recovery rate (AAR), a standard sequence-recovery metric in protein design, on a validation set of 1,126 antibody and 374 nanobody structures (Figures 2 and 14). Tuning α for AbMPNN is less straightforward, as it was trained on a substantial fraction of these structures. Nonetheless, ensembling yields consistent but modest improvements in both cases. ProteinMPNN benefits from a smaller α than AbMPNN, suggesting that stronger IgLM guidance is advantageous in this setting.

Figure 3a shows that combining ProteinMPNN with IgLM substantially improves CDR recovery, closing the gap to antibody-specific methods such as AntiFold, though still lagging behind AbMPNN, particularly on heavy-chain CDR3. Gains from ensembling AbMPNN with IgLM are smaller but consistent. Nanobody inverse folding remains challenging for all models, reflecting the lower conservation of heavy chains. LM-Design, itself a pLM-based approach, achieves the strongest performance on nanobodies. Overall, ensemble approaches consistently outperform their base models.

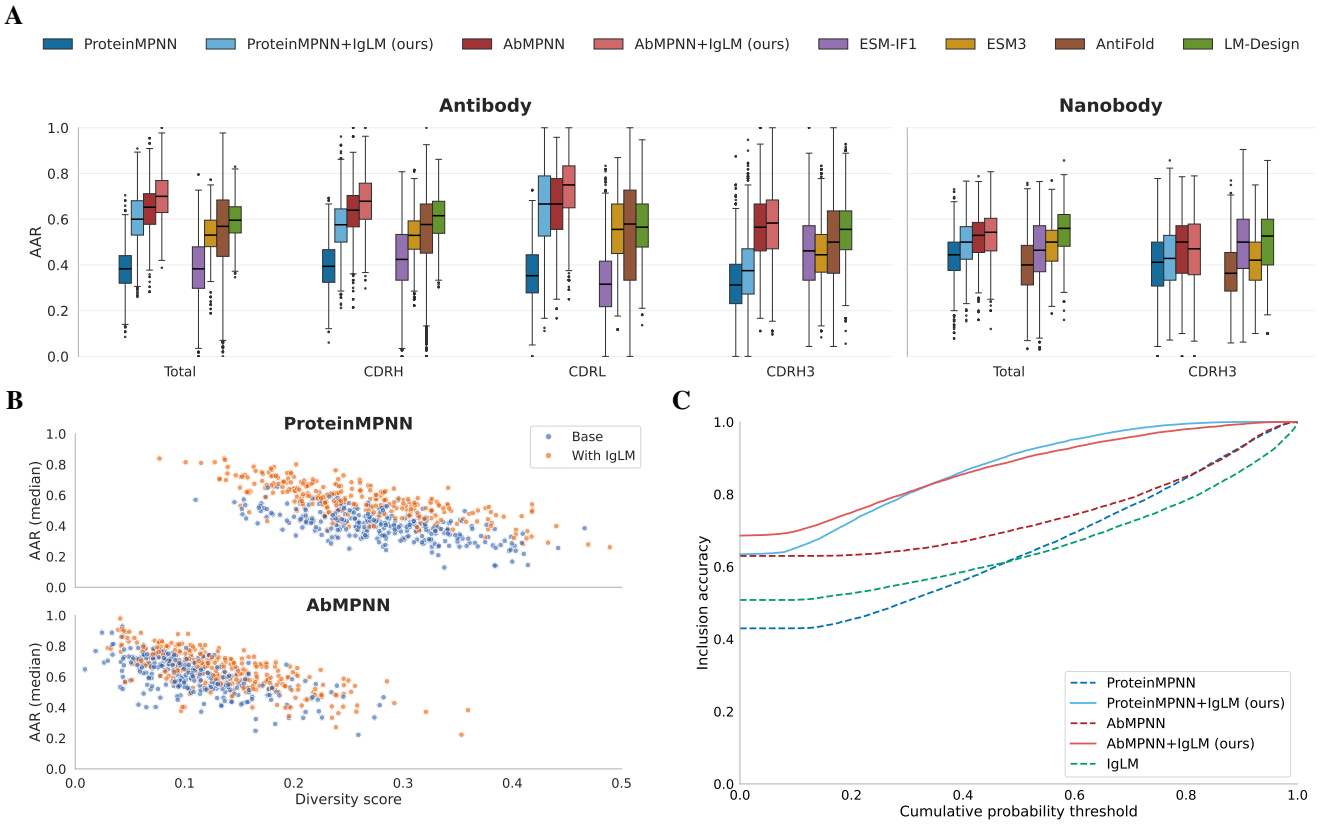

Figure 3. (A) Amino acid recovery rates (AAR) for CDR loops of different models, computed from 10 predictions per sequence on the antibody and nanobody test set at a fixed sampling temperature of 0.25. (B) Diversity score and median AAR per structure, illustrating differences between base and ensemble models for ProteinMPNN and AbMPNN. (C) Fraction of CDR positions where the ground-truth amino acid is included within the set of top-probability tokens whose cumulative probability mass is below threshold p. Curves closer to the upper-left indicate higher confidence in the correct token, as it is recovered at lower cumulative probability thresholds across positions.

Figure 3b shows that for ProteinMPNN, ensembling improves AAR while preserving diversity, whereas AbMPNN sees small AAR improvement and increased diversity. Figure 3c reveals a key property. We evaluate predictions from a cumulative-probability perspective, where correctness is defined as whether the ground-truth token appears within the smallest set of most probable tokens whose total probability mass exceeds a threshold p. While AbMPNN achieves high top-1 accuracy, its probability mass is highly concentrated around the top prediction, with negligible mass beyond it. In contrast, the ensembles distribute probability more broadly and more often include the correct token within lower-probability regions, making them more robust for sampling-based design.

Additional analyses in Appendix A.2 show that the ensembles produce highly natural sequences, as measured by ProGen2 and IgLM log-likelihoods, closely matching native sequences, a property not achieved by ProteinMPNN or AbMPNN. We further find that the generated sequences maintain structural consistency, as reflected in Boltz2 (Passaro et al., 2025) refolding experiments, where ensembles consistently rank among the top-performing models in RMSD to native structures, despite variability. Beyond these metrics, we analyze amino acid profiles and ensembling behavior, focusing on when ensembling is most effective. The ensemble typically follows the more confident model, but reflects agreement when both models are uncertain.

Discussion

Guiding general protein inverse folders with domain-specific language models improves sequence recovery via inference-time ensembling without modifying model parameters. This approach also maintains sequence diversity and yields more balanced predictions across positions. Structural fine-tuning remains important, as antibody-specific inverse folding models substantially improve challenging regions such as heavy-chain CDR3. Notably, language-model guidance further improves even these specialized models, suggesting the best performance comes from combining structural fine-tuning with complementary sequence-level knowledge from large-scale language models.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author

Impact Statement

This work introduces a simple, training-free approach to integrate antibody language models with inverse folding, improving sequence recovery and diversity without requiring additional structural data. By enabling more rapid use of abundant sequence information, this approach may accelerate protein and antibody design workflows, particularly in settings where structural data are limited. Potential applications include therapeutic antibody engineering and biomolecular design. However, as with other generative models in biology, there is a risk that such methods could be misused to design biologically active sequences without sufficient experimental and biosafety evaluation. In addition, model biases inherited from training data may influence generated sequences. Careful wet-lab validation and responsible use remain essential when applying these methods in real-world settings.

References

Abanades, B., Wong, W. K., Boyles, F., Georges, G., Bujotzek, A., and Deane, C. M. Immunebuilder: Deep-learning models for predicting the structures of immune proteins. *Communications Biology*, 6(1):575, 2023.

Bai, P., Miljković, F., Liu, X., De Maria, L., Croasdale-Wood, R., Rackham, O., and Lu, H. Mask-prior-guided denoising diffusion improves inverse protein folding. *Nature Machine Intelligence*, pp. 1–13, 2025.

Branson, N. and Deane, C. Antidif: Accurate and diverse antibody specific inverse folding with discrete diffusion. *bioRxiv*, pp. 2025–07, 2025.

Butcher, J., Krishna, R., Mitra, R., Brent, R. I., Li, Y., Corley, N., Kim, P. T., Funk, J., Mathis, S., Salike, S., et al. De novo design of all-atom biomolecular interactions with rfdiffusion3. *bioRxiv*, 2025.

Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R. J., Milles, L. F., Wicky, B. I., Courbet, A., de Haas, R. J., Bethel, N., et al. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378 (6615):49–56, 2022.

Ding, W., Nakai, K., and Gong, H. Protein design via deep learning. *Briefings in Bioinformatics*, 23(3):bbac102, 03 2022. ISSN 1477-4054. doi: 10.1093/bib/bbac102. URL https://doi.org/10.1093/bib/bbac102.

Dreyer, F. A., Cutting, D., Schneider, C., Kenlay, H., and Deane, C. M. Inverse folding for antibody sequence design using deep learning. *arXiv preprint arXiv:2310.19513*, 2023.

Ektefaie, Y., Viessmann, O., Narayanan, S., Dresser, D., Kim, J. M., and Mkrtchyan, A. Reinforcement learning on structure-conditioned categorical diffusion for protein inverse folding. *arXiv preprint arXiv:2410.17173*, 2024.

Fang, X., Wang, F., Liu, L., He, J., Lin, D., Xiang, Y., Zhu, K., Zhang, X., Wu, H., Li, H., et al. A method for multiple-sequence-alignment-free protein structure prediction using a protein language model. *Nature Machine Intelligence*, 5(10):1087–1096, 2023.

Ferruz, N., Schmidt, S., and Höcker, B. Protgpt2 is a deep unsupervised language model for protein design. *Nature communications*, 13(1):4348, 2022.

Frank, C., Khoshouei, A., Fuβ, L., Schiwietz, D., Putz, D., Weber, L., Zhao, Z., Hattori, M., Feng, S., de Stigter, Y., et al. Scalable protein design using optimization in a relaxed sequence space. *Science*, 386(6720):439–445, 2024.

Gao, Z., Tan, C., Chacón, P., and Li, S. Z. Pifold: Toward effective and efficient protein inverse folding. *arXiv preprint arXiv:2209.12643*, 2022.

Gordon, C., Lu, A. X., and Abbeel, P. Protein language model fitness is a matter of preference. *bioRxiv*, pp. 2024–10, 2024.

Goverde, C. A., Pacesa, M., Goldbach, N., Dornfeld, L. J., Balbi, P. E., Georgeon, S., Rosset, S., Kapoor, S., Choudhury, J., Dauparas, J., et al. Computational design of soluble and functional membrane protein analogues. *Nature*, 631(8020):449–458, 2024.

Hayes, T., Rao, R., Akin, H., Sofroniew, N. J., Oktay, D., Lin, Z., Verkuil, R., Tran, V. Q., Deaton, J., Wiggert, M., et al. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025.

Høie, M. H., Hummer, A., Olsen, T. H., Aguilar-Sanjuan, B., Nielsen, M., and Deane, C. M. Antifold: Improved antibody structure-based design using inverse folding. *arXiv preprint arXiv:2405.03370*, 2024.

Hsu, C., Verkuil, R., Liu, J., Lin, Z., Hie, B., Sercu, T., Lerer, A., and Rives, A. Learning inverse folding from millions of predicted structures. In *International conference on machine learning*, pp. 8946–8970. PMLR, 2022.

Kalemati, M., Noroozi, A., Shahbakhsh, A., and Koohi, S. Paraantiprot provides paratope prediction using antibody and protein language models. *Scientific Reports*, 14(1): 29141, 2024.

<anon.email@domain.com>.

Submitted to the 2026 Workshop on Generative and Agentic AI for Biology (ICML 2026). Do not distribute.

Leem, J., Mitchell, L. S., Farmery, J. H., Barton, J., and Galson, J. D. Deciphering the language of antibodies using self-supervised learning. *Patterns*, 3(7), 2022.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637): 1123–1130, 2023.

Nijkamp, E., Ruffolo, J. A., Weinstein, E. N., Naik, N., and Madani, A. Progen2: exploring the boundaries of protein language models. *Cell systems*, 14(11):968–978, 2023.

Olsen, T. H., Boyles, F., and Deane, C. M. Observed antibody space: A diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. *Protein Science*, 31(1):141–146, 2022.

Pacesa, M., Nickel, L., Schellhaas, C., Schmidt, J., Pyatova, E., Kissling, L., Barendse, P., Choudhury, J., Kapoor, S., Alcaraz-Serna, A., et al. One-shot design of functional protein binders with bindcraft. *Nature*, 646(8084):483–492, 2025.

Passaro, S., Corso, G., Wohlwend, J., Reveiz, M., Thaler, S., Somnath, V. R., Getz, N., Portnoi, T., Roy, J., Stark, H., et al. Boltz-2: Towards accurate and efficient binding affinity prediction. *BioRxiv*, 2025.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Raybould, M. I. J., Marks, C., Krawczyk, K., Taddese, B., Nowak, J., Lewis, A. P., Bujotzek, A., Shi, J., and Deane, C. M. Five computational developability guidelines for therapeutic antibody profiling. *Proceedings of the National Academy of Sciences*, 116(10):4025–4030, 2019. doi: 10.1073/pnas.1810576116. URL https://www.pnas.org/doi/abs/10.1073/pnas.1810576116.

Shanehsazzadeh, A., Alverio, J., Kasun, G., Levine, S., Calman, I., Khan, J. A., Chung, C., Diaz, N., Luton, B. K., Tarter, Y., et al. Igdesign: In vitro validated antibody design against multiple therapeutic antigens using inverse folding. *bioRxiv*, pp. 2023–12, 2023.

Shuai, R. W., Ruffolo, J. A., and Gray, J. J. Iglm: Infilling language modeling for antibody sequence design. *Cell Systems*, 14(11):979–989, 2023.

Slabinski, L., Jaroszewski, L., Rodrigues, A. P., Rychlewski, L., Wilson, I. A., Lesley, S. A., and Godzik, A. The challenge of protein structure determination—lessons from structural genomics. *Protein Science*, 16(11):2472–2482, 2007.

Stark, H., Faltings, F., Choi, M., Xie, Y., Hur, E., O'Donnell, T., Bushuiev, A., Uçar, T., Passaro, S., Mao, W., et al. Boltzgen: Toward universal binder design. *bioRxiv*, pp. 2025–11, 2025.

Vashchenko, D., Nguyen, S., Goncalves, A., da Silva, F. L., Petersen, B., Desautels, T., and Faissol, D. Abbert: learning antibody humanness via masked language modeling. *bioRxiv*, pp. 2022–08, 2022.

Wang, Z., Ji, Y., Tian, J., and Zheng, S. Retrieval augmented diffusion model for structure-informed antibody design and optimization. *arXiv preprint arXiv:2410.15040*, 2024.

Watson, J. L., Juergens, D., Bennett, N. R., Trippe, B. L., Yim, J., Eisenach, H. E., Ahern, W., Borst, A. J., Ragotte, R. J., Milles, L. F., et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976): 1089–1100, 2023.

Zheng, Z., Deng, Y., Xue, D., Zhou, Y., Ye, F., and Gu, Q. Structure-informed language models are protein designers. In *International conference on machine learning*, pp. 42317–42338. PMLR, 2023.

A. Appendix

A.1. Extended methods

Diversity score

We quantified the diversity of our sequence predictions using the metric introduced by (Branson & Deane, 2025) for their discrete diffusion–based inverse folding model.

Sequence diversity D across a set of M predicted sequences $\{\hat{S}_m\}_{m=1}^{M}$ is defined as the average pairwise dissimilarity:

$$D = \frac{1}{M(M-1)} \sum_{j=1}^{M} \sum_{\substack{k=1 \\ k \neq j}}^{M} \Big(1 - \mathrm{SR}(\hat{S}_j, \hat{S}_k)\Big),$$

$$\mathrm{SR}(\hat{S}_j, \hat{S}_k) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{\hat{S}_j[i] = \hat{S}_k[i]\},$$

where N is the sequence length, and $\mathbf{1}\{\cdot\}$ denotes the indicator function. Here, SR represents the fraction of matching residues between two sequences. This metric captures the average fraction of differing amino acids between predicted sequences and can be applied directly to sequences generated for a given structure.

A.2. Analysis of Model Behavior

Naturalness

We assess the naturalness of the designed sequences using log-likelihood scores from IgLM and ProGen2 to evaluate whether ensembling with a language model improves sequence quality. IgLM is included as a reference model, as its inclusion in the ensemble is expected to increase log-likelihood values by construction. ProGen2 is included as an independent measure of naturalness, as it has previously been used for this purpose in general protein settings (Gordon et al., 2024).

The analysis shown in Figure 4 demonstrates that, regardless of the underlying base model, ensembling results in designed sequences with log-likelihood scores closer to those of the native sequences. In contrast, sequences generated without ensembling tend to exhibit lower log-likelihoods, indicating reduced naturalness.

Structural validation

To evaluate whether sequences designed by the inverse folding models satisfy structural constraints beyond sequence-level agreement (amino acid recovery), we used Boltz2 (Passaro et al., 2025) to refold the designed sequences. Owing to the large number of sequences considered (10 designs per inverse folding model for each PDB structure in the test set, totaling approximately 24,000 sequences), the MSA component of Boltz2 was disabled, and each sequence was refolded only 10 times.

As a reference, we also refolded the native PDB sequences ten times using Boltz2 (Figure 5). The resulting target-aligned binder RMSD values were often relatively high, indicating that Boltz2 does not always reproduce the experimentally observed binding pose. Similar behavior was observed when refolding the native sequences with MSA enabled. While Boltz2 confidence scores were generally consistent with expected ranges and correlated with both target-aligned binder RMSD and framework-aligned CDR RMSD, higher variability was observed for certain structures. Some variability was also observed across repeated Boltz2 runs on the same native sequence, though these differences were limited.

For each structure, we compute two RMSD-based metrics. First, we measure the RMSD between the native binder structure and the designed binder structure after aligning both to the target structure, providing an overall measure of structural agreement in the binding context. Second, we compute a CDR-loop RMSD, where we align the framework regions of the binder and then measure the RMSD over the CDR loops only. This second metric isolates variability in the most functionally relevant regions of the binder while controlling for differences in the conserved scaffold.

While the results, shown in Figure 5, show a small difference favoring the ensemble approach and the base inverse folding models (AbMPNN and ProteinMPNN), the high variance prevents strong conclusions. Differences between nanobody and antibody RMSD values are largely attributable to their differing residue counts and molecular sizes.

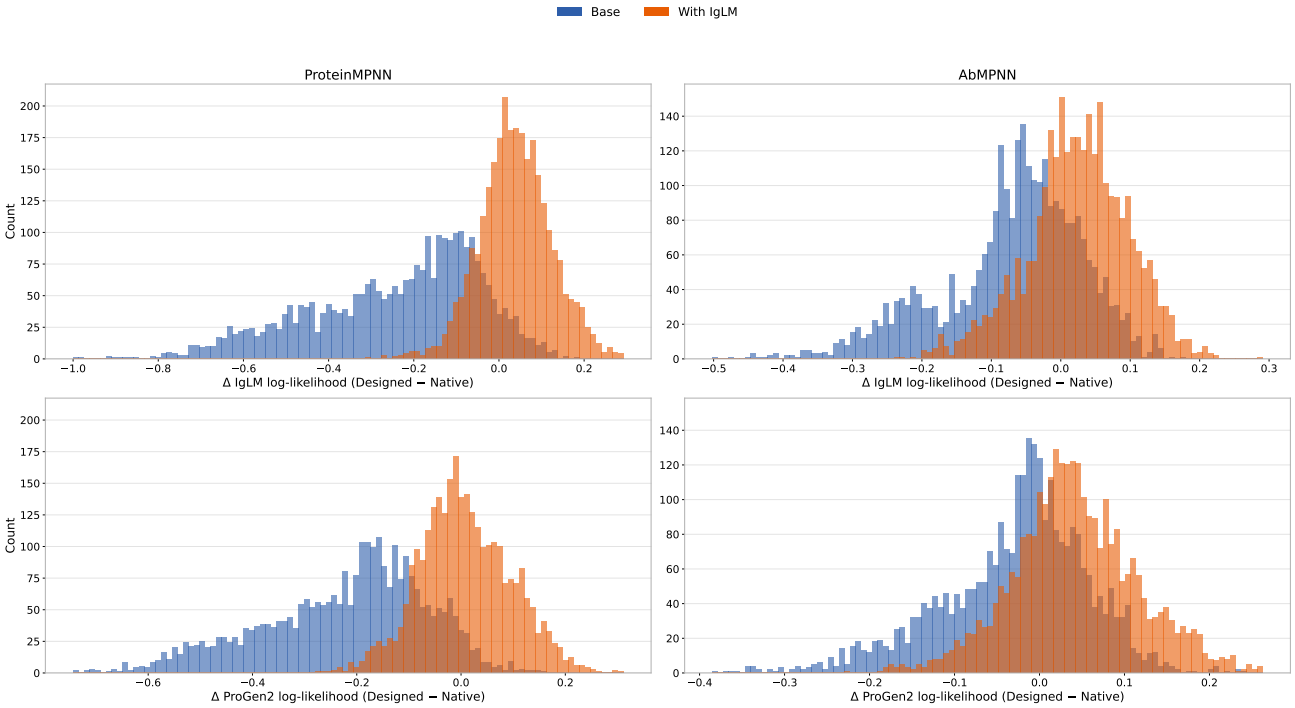

Figure 4. Shown are histograms of log-likelihood differences between designed and native sequences for the base models AbMPNN and ProteinMPNN, both with and without IgLM ensembling, based on the antibody and nanobody test set. Log-likelihoods are obtained using IgLM and ProGen2.

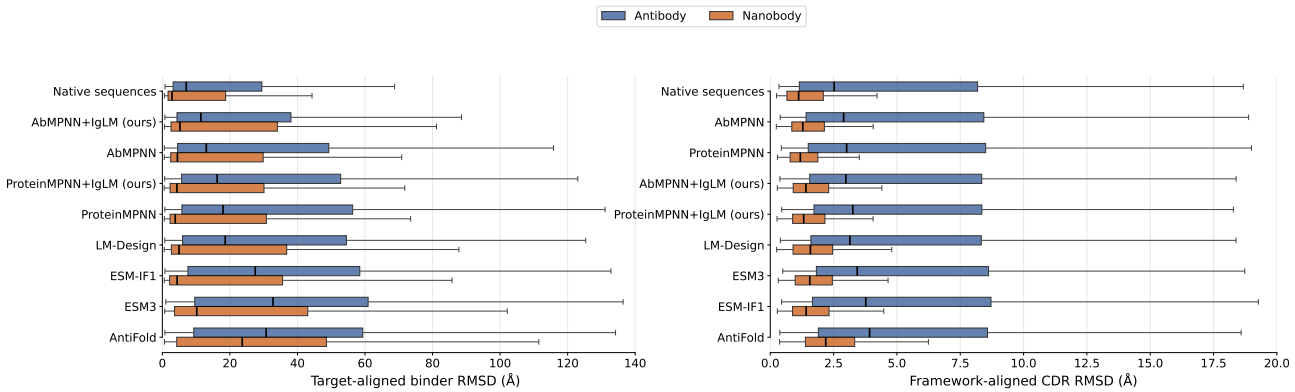

Figure 5. RMSD values between 10 Boltz2 predictions for each native and designed sequence (labeled by model name in the plot) in the test set and their corresponding native structures. Target-aligned RMSD is shown for the full binder, while framework-aligned RMSD is shown for the CDR regions.

Ensemble mechanism analysis

To better understand when and how ensembling is beneficial, we analyze in more detail the ProteinMPNN+IgLM ensemble, where the effects of ensembling are most pronounced. To do so, we evaluate predictions from ProteinMPNN, IgLM, ProteinMPNN+IgLM, AbMPNN, and AbMPNN+IgLM on the test set under a conditional setting where the subsequence up to the position being predicted is fixed to the correct (native) sequence. This differs from standard autoregressive sampling, where the true (ground-truth) tokens are not available during generation, and sequences are generated solely from previously predicted tokens. This evaluation protocol ensures that all models are evaluated under identical ground-truth context conditioning.

Figure 6 shows the difference in cross-entropy loss between the combined model (ProteinMPNN+IgLM) and ProteinMPNN alone. At temperature T=1 (standard sampling temperature), improvements of ensembling are mainly observed in cases where ProteinMPNN and IgLM strongly disagree, suggesting that the IgLM signal is more informative in these regions. For the lower temperature setting (T=0.25), a different pattern emerges. When the two models strongly disagree (high divergence), ensembling typically improves performance. In contrast, when ProteinMPNN is highly confident (i.e., exhibits low Shannon entropy), its predictions tend to dominate the ensemble, leading to behavior similar to ProteinMPNN alone. Finally, in cases where ProteinMPNN shows high uncertainty but the KL divergence between the two models is low, indicating that IgLM is also uncertain, the ensemble can occasionally degrade performance, as both models provide weak or uninformative signals.

Looking at region-specific signals in Figure 8, the overall trends remain consistent. At T=1, the ensemble cross-entropy loss is often higher than that of the best single model. ProteinMPNN performs better in the CDR3 loops, whereas IgLM achieves lower loss in the remaining loop regions. At T=0.25, however, the ensemble on average outperforms both ProteinMPNN and IgLM across all regions, and is often better than AbMPNN, and comparable to AbMPNN+IgLM.

Overall, these results indicate that the ensemble predominantly combines the strengths of both models by deferring to the more confident prediction. However, there are also cases where ensembling actively improves prediction quality over both models, as shown in Figure 7. In particular, when the correct token is a plausible candidate for both models, but they assign probability mass to different peaks, ensembling can create a new peak in the combined distribution, leading to improved predictions.

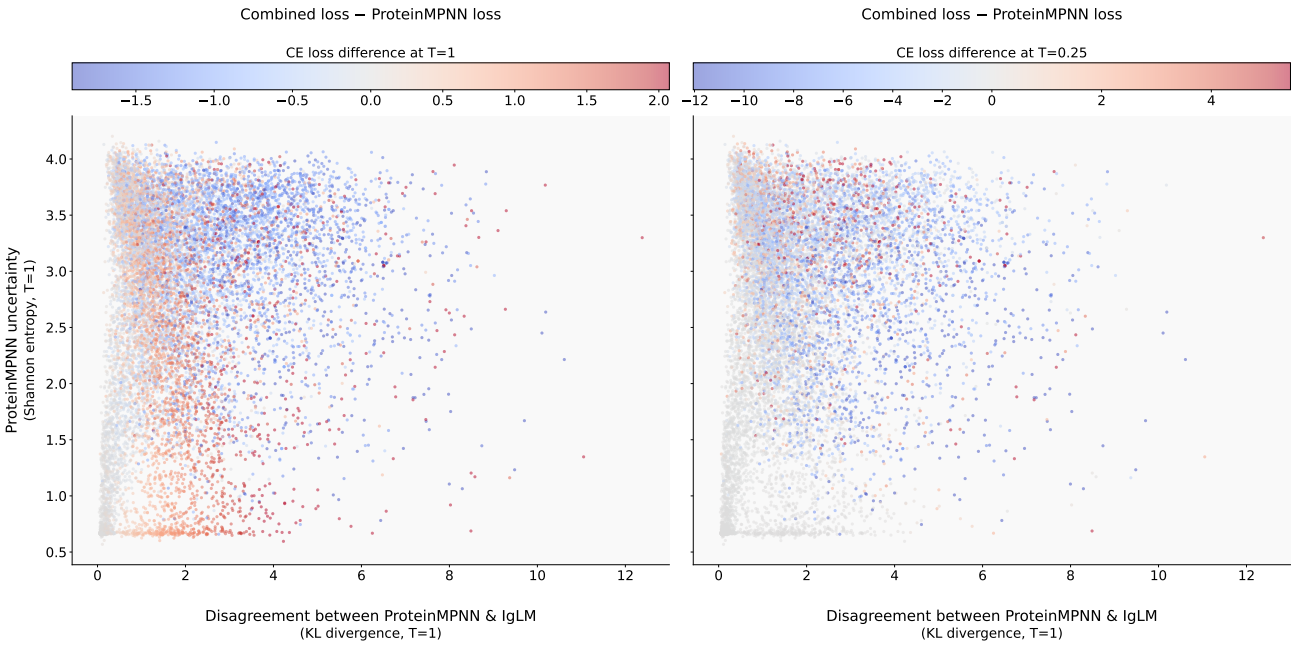

Figure 6. Each point in the plot represents the difference in cross-entropy loss between the combined model (ProteinMPNN+IgLM) and ProteinMPNN alone, with blue indicating that the combined model performs better and red indicating that ProteinMPNN performs better. The x-axis shows the KL divergence between ProteinMPNN and IgLM predictions, used as a measure of model divergence, while the y-axis shows the Shannon entropy of ProteinMPNN predictions, used as a measure of model uncertainty. The plot is shown for two different temperature settings (T=1, T=0.25).

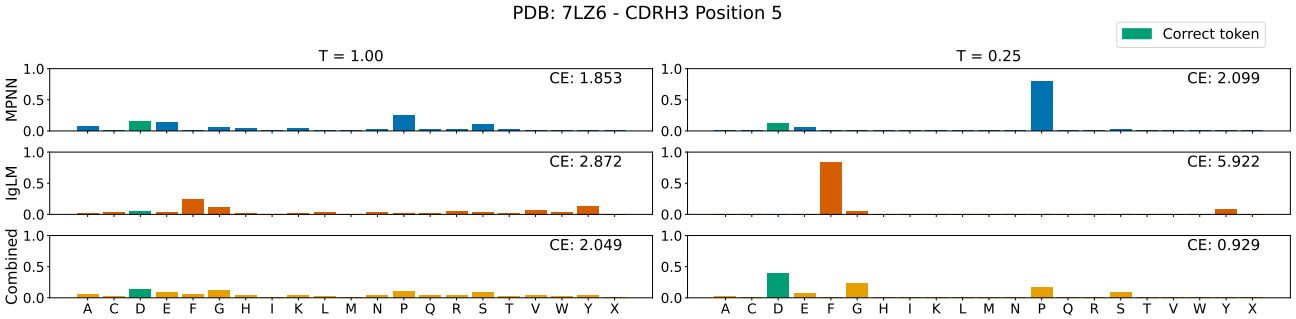

Figure 7. Shown is an example of ensembling ProteinMPNN and IgLM, where predictions are first combined at temperature T=1 and then rescaled to T=0.25 by temperature adjustment of the resulting distribution for evaluation. For each model, the predicted categorical distribution over amino acids at a given position is shown, along with the cross-entropy (CE) loss with respect to the ground-truth amino acid. The correct token is highlighted in green. The ensemble redistributes probability mass between the individual models, which can shift the highest-probability token and change the resulting CE compared to the individual predictions. At T=1, the distributions are more diffuse, and the ensemble does not improve CE over ProteinMPNN alone. In contrast, at T=0.25, the sharper distributions lead to more decisive probability shifts, allowing the ensemble to assign the highest probability to the correct token and thereby reduce CE.

Amino Acid Frequencies of Models Compared to Native Sequences

We computed amino acid frequencies in the CDR loops of test set structures and compared them to the frequencies in sequences designed by each model, shown in Figure 9. This comparison provides a coarse measure of how well generated sequences reproduce native amino acid composition in antibody binding regions. Ensembling improves ProteinMPNN predictions in terms of matching the native amino acid frequency distribution in CDR loops, but has no positive effect on AbMPNN. The LM-based models LM-Design and ESM3 show high overall similarity to native sequences, although not particularly trained on antibody sequences only. General protein inverse folding models show the largest deviation from native frequencies. Across the better-performing models, there is a tendency to oversample the most frequent amino acids, potentially reflecting training biases toward amino acid recovery objectives.

Comparison of amino acid recovery rates across the test set

Figure 10 shows model performance for all CDR loops in the test set.

Diversity comparison of model predictions

Figure 11 compares the sequence diversity across models. In general, general protein inverse folding models, which exhibit lower recovery rates, produce more diverse sequences on average. ESM3 shows comparatively low diversity despite its weaker recovery performance.

A.3. Ablations and Design Decisions

Relationship between amino acid recovery rate and IgLM log-likelihood

Although the correlation in Figure 12 is weak and sometimes negative, amino acid recovery generally correlates with IgLM log-likelihood for both ProteinMPNN and AbMPNN, more strongly in the light chain. Aggregating both models increases the correlation. This suggests IgLM log-likelihood can serve as an empirical proxy for similarity to the native sequence.

Comparison of decoding strategies: Random vs. Left-to-Right

In ProteinMPNN, the decoding order is not fixed but is randomly sampled during inference, consistent with the training procedure. Using a left-to-right decoding order, which is required by the ensemble setup, does not significantly affect sequence recovery or sequence diversity in our experiments, see Figure 13.

Determination of α for nanobody structures

Tuning of α on nanobody structures, following the same strategy used for antibodies (Figure 14).

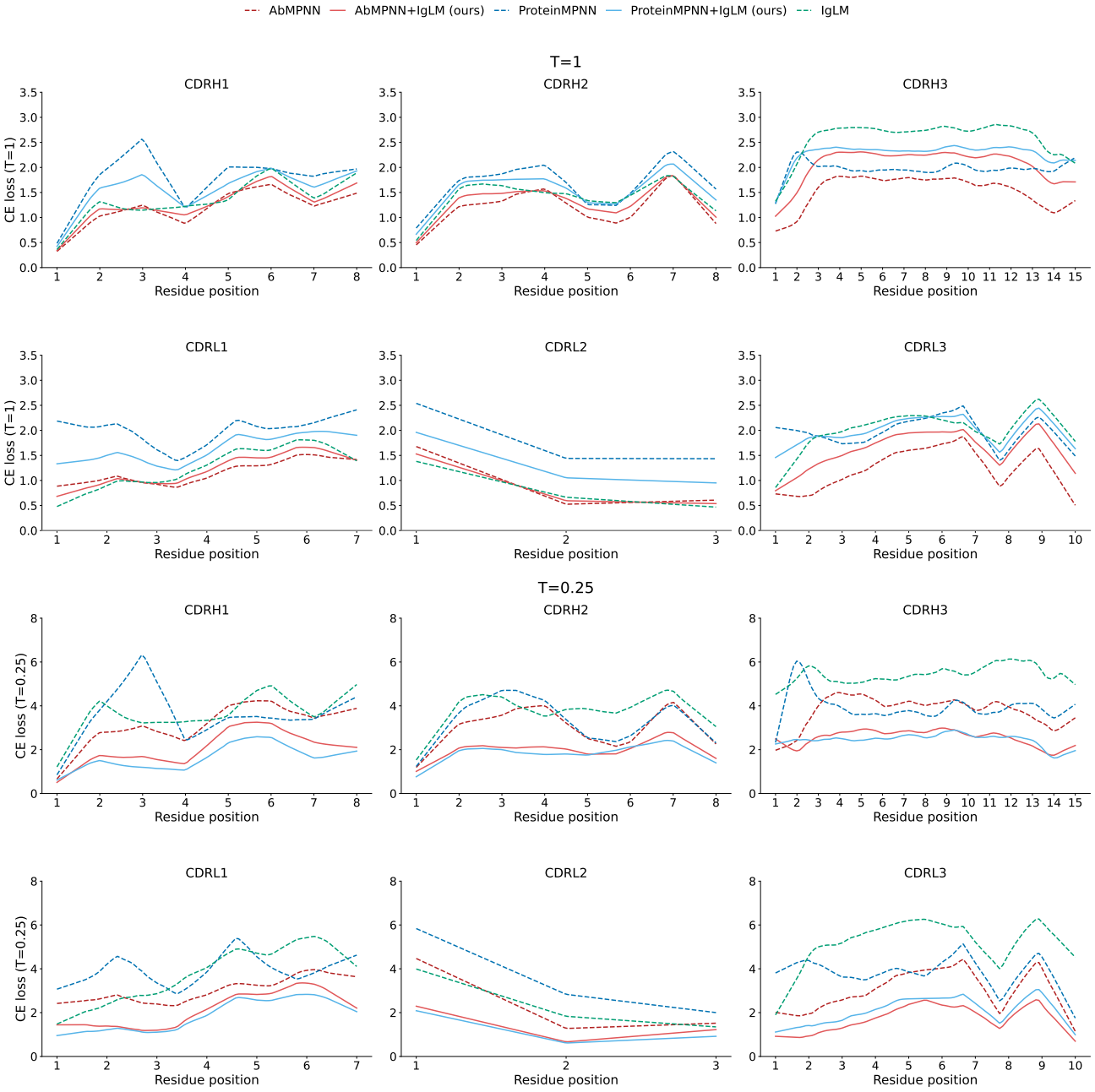

Figure 8. The mean cross-entropy loss for different CDR loops is shown for each model at temperatures T=1 and T=0.25. Since CDR loops vary in length, direct position-wise comparison is not possible. To address this, each CDR loop is first normalized to a fixed length by linearly interpolating the cross-entropy loss values onto a grid of 100 evenly spaced positions along the loop. The mean cross-entropy loss is then computed over these 100 normalized positions. Finally, results are rescaled to account for differences in the original average CDR length.

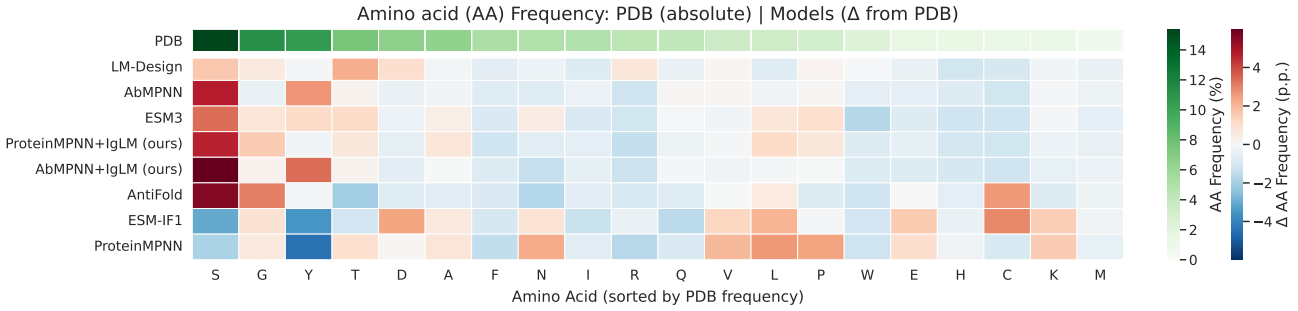

Figure 9. Amino acid frequencies in the CDR loops of all test set structures are compared with those predicted by each model. The results are shown as differences relative to the empirical amino acid frequencies observed in the native PDB structures.

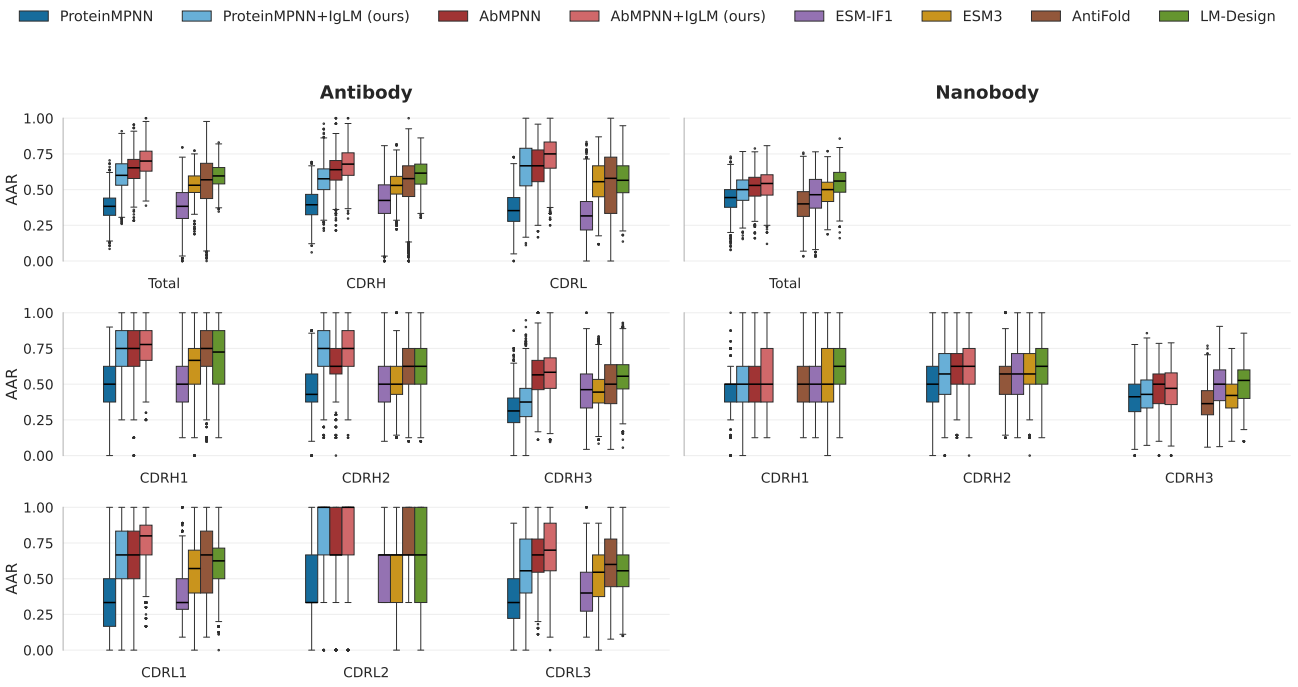

Figure 10. Amino acid recovery rates for CDR loops of different models, computed from 10 predictions per sequence on the antibody and nanobody test set at a fixed sampling temperature of 0.25. The plot also reports results separately for each CDR region.

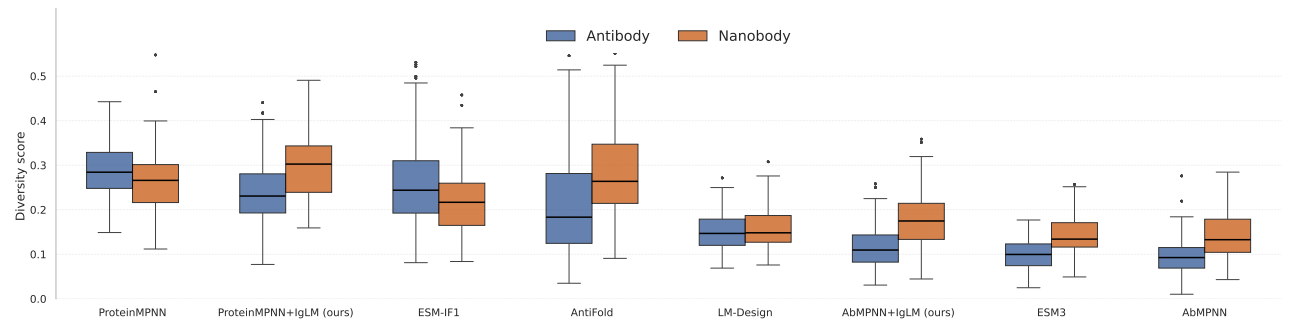

Figure 11. Model-wise diversity scores, computed according to Branson et al., for the antibody and nanobody test sets. For each structure, all 10 designed sequences generated by a given model were considered, and a single diversity score was calculated based on the variability among those sequences.

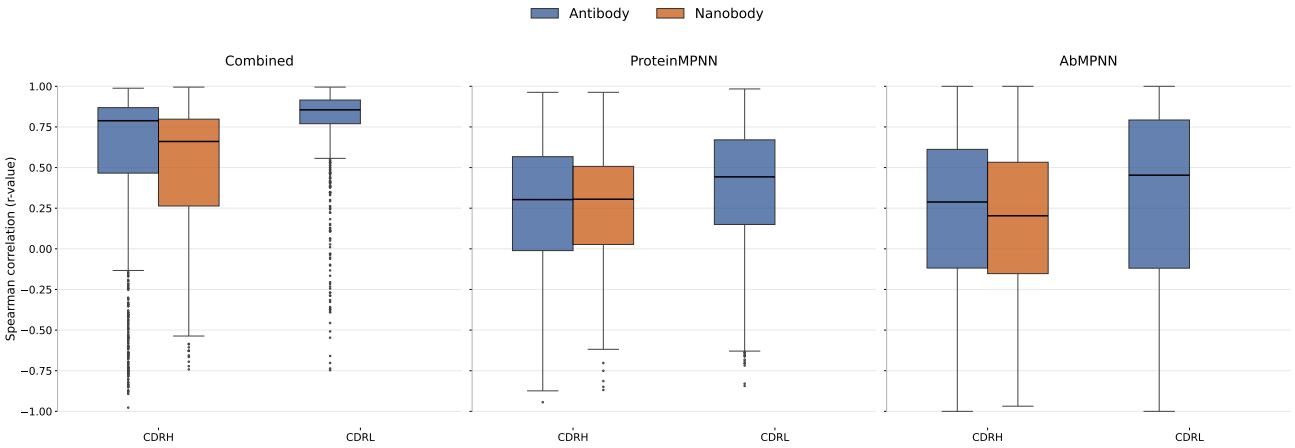

Figure 12. Correlation analysis of amino acid recovery rates for AbMPNN and ProteinMPNN predictions on heavy- and light-chain CDR loops, and IgLM log-likelihoods. Results are based on 10 predictions per structure for the nanobody and antibody validation set. Spearman correlations were computed separately for each PDB structure. Boxplots show the distribution for the r-value of correlation analyses.

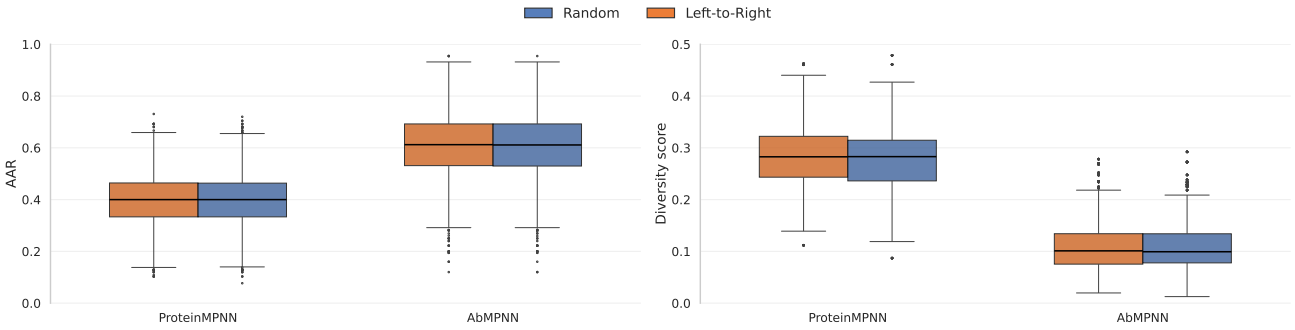

Figure 13. Amino acid recovery rates and diversity score for CDR loops under two decoding strategies, random and left-to-right, computed from 10 predictions of ProteinMPNN and AbMPNN on antibody and nanobody test set at a fixed temperature of 0.25.

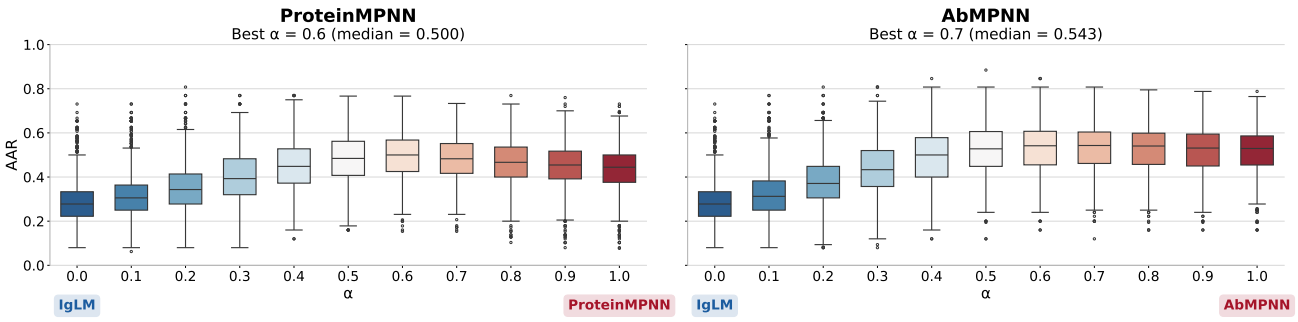

Figure 14. Amino acid recovery rates of CDR loops at different α values for ensembles of ProteinMPNN or AbMPNN combined with IgLM, evaluated on nanobody structures from the validation set, with the sampling temperature fixed at 0.25 and 10 predictions per structure.