Antibody CDR design by ensembling inverse folding with protein language models

Diego del Alamo[1], Rahel Frick[2], Daphné Truan[3], Joel D. Karpiak[4]
Protein Design & Informatics, GSK Research & Development
[1] Baar, ZG, Switzerland, [2] Heidelberg, DE, [3] Stevenage, UK, [4] Collegeville, PA, USA

Inference using both inverse folding and language models

Modern machine learning (ML) tools surpass traditional physics-based methods in designing proteins that express in vitro by learning nuanced patterns from massive databases of protein sequences and structures [1, 2, 3]. Protein language models (PLMs) such as ESM-2 and inverse folding models such as ProteinMPNN in particularly have seen widespread use [1, 4]. These two methods rely on distinct architectures and to learn from sequence and structural data, respectively. Attempts to unify them into a single ML tool capable of leveraging both sources of information have encountered mixed success, with reports of marginal improvements in sequence recovery [5], no improvements in downstream tasks [6], or "catastrophic forgetting" of information learned during pretraining [7].

A workaround involves supplementing structural datasets, which typically comprise tens of thousands of training examples, with hundreds of thousands to millions of computational models generated from the same sequence sets used to train PLMs [8]. This has proven useful for inverse folding of antibodies, heterooligomeric proteins that are widely adapted and reengineered due to their ability to bind a seemingly limitless range of targets with high affinity [9-12]. Models trained on the PDB and computational models generated using paired heavy/light chain sequences from OAS [13], such as AbMPNN [10], a fine-tuned version of ProteinMPNN, show improved antibody sequence recovery relative to those trained on just experimental structures or predicted models. Yet this predictive power comes at the cost of more complex training regimes, a trade-off with unclear benefit given the rapid pace at which new generic inverse folding models are released.

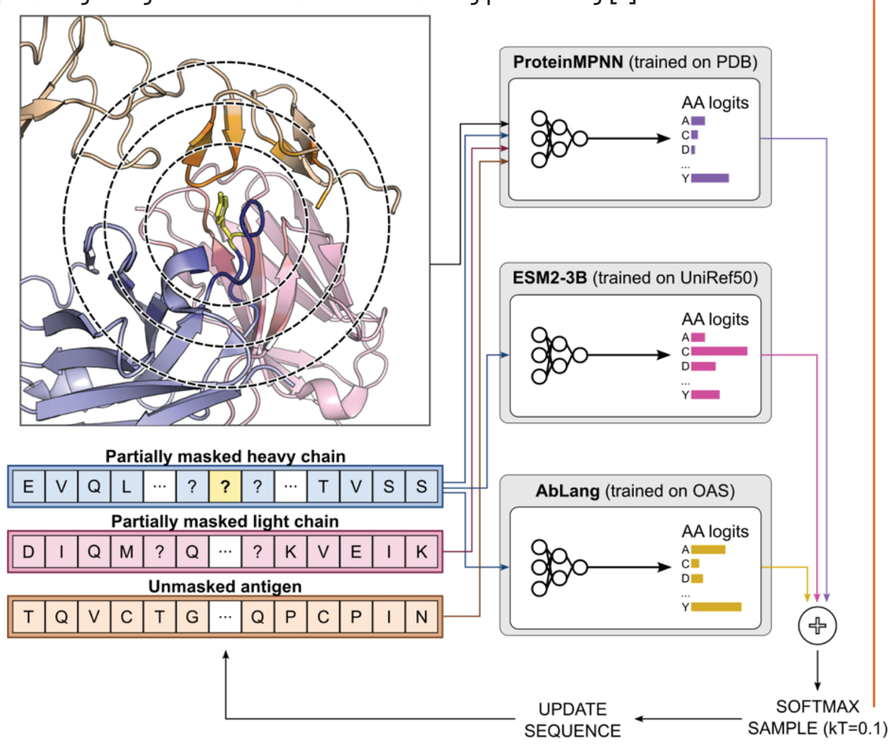

Figure 1. Schematic for ensembling inverse folding models, which design residues using local structural characteristics, with protein language models, which global sequence information.

Here we show that the performance of such antibody-specific inverse folding models can be matched or exceeded by ensembling off-the-shelf generic inverse folding models and antibody-specific language models [14], but not generic protein language models, without any retraining or fine-tuning. Both types of methods compute probability distributions of all 20 canonical amino acids across a predetermined set of masked residues (Figure 1), and ensembling such outputs has previously been shown to yield state-of-the-art performance in various zero-shot prediction tasks [15]. This approach establishes a baseline against which ML-based design tools jointly running inference on sequence and structure could be evaluated.

Structure-based methods generate unrealistic sequences

This study focuses on designing the six complementarity-determining regions (CDRs) of antibodies, which mediate antigen binding. Simultaneous unrestrained inverse folding using ProteinMPNN of all six CDRs of a previously published benchmark set [16] led to a highly diverse sequences, as judged by residue-level Shannon entropy values, but with poor sequence recovery (Figures 2 and 3). Sequence identity to the closest V-gene, calculated using ANARCI [17, 18], was also found to be poor (not shown). These results were broadly consistent with previous reports [10-12].

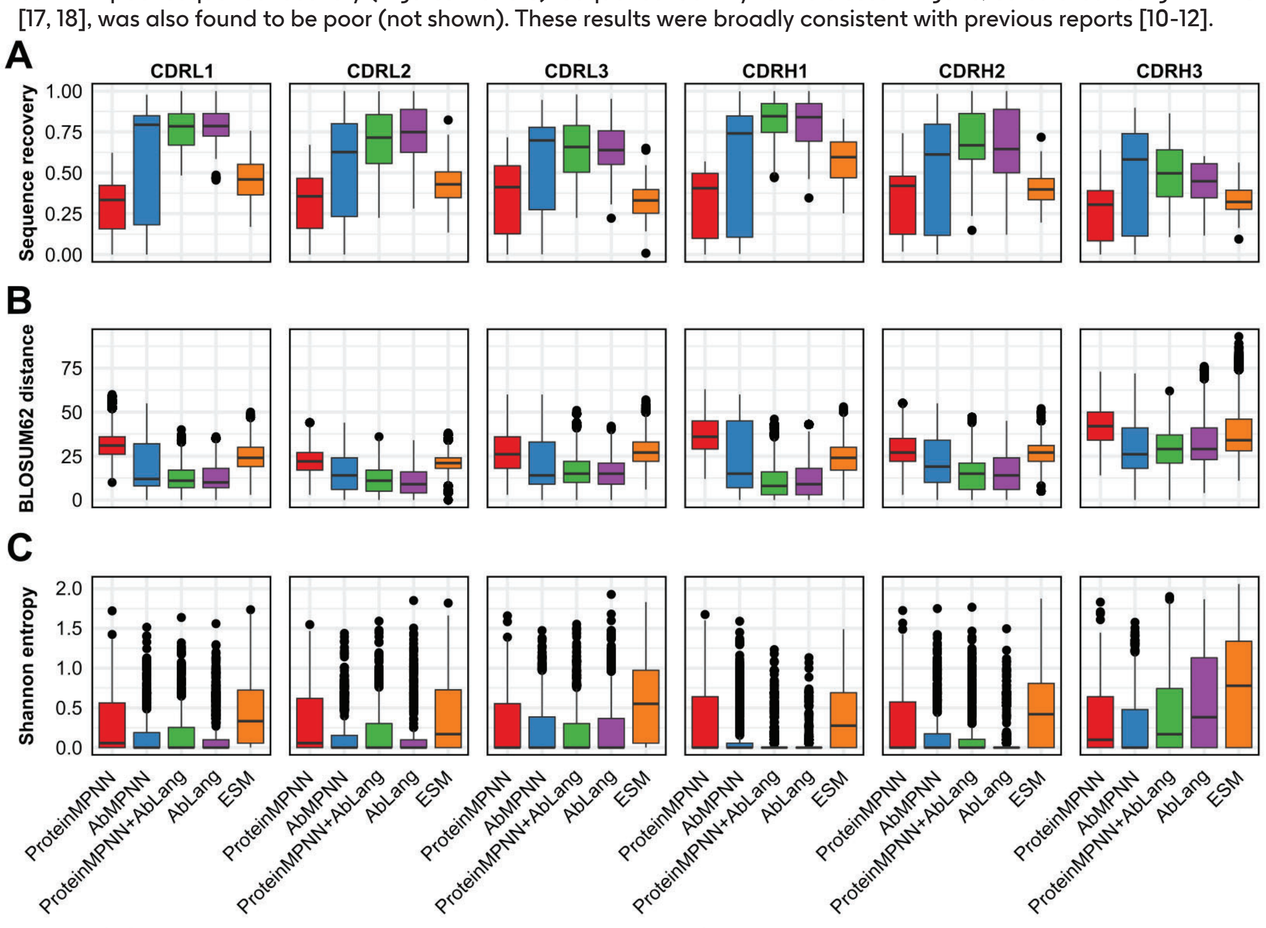

Figure 2. Sequence recovery and similarity to native antibodies. A) Sequence recovery was found to be poor across all CDRs using the inverse folding model ProteinMPNN or ESM alone. The combination of ProteinMPNN and AbLang outperformed either method in isolation, particularly on CDRH3 designs. B) Sequence similarity calculated using the BLOSUM62 distance matrix. Values were flipped so that zero indicates a perfect match and greater values indicate greater dissimilarity. C) Average per-residue diversity of CDR designs.

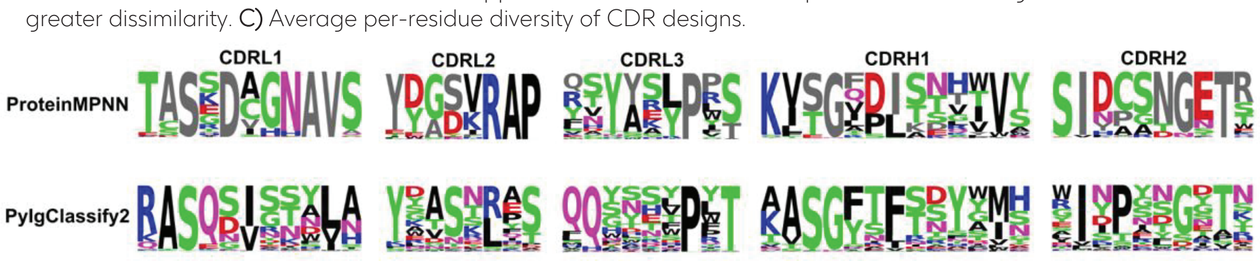

Figure 3. Sequence logos of ProteinMPNN designs of Trastuzumab (PDB: 1N8Z) alongside those for equivalent conformational clusters. CDRH3 is omitted because it does not form canonical clusters.

AbMPNN, a version of ProteinMPNN trained on antibody structural data, showed broad improvements in all respects, including improved sequence recovery and V-gene sequence identity. Yet, relative to other approaches discussed below, many sequences retained high negative log-likelihoods to PSSMs for both PyIgClassify2 conformational clusters and the broader OAS, indicating persistent yet reduced unrealistic design (Figures 3 and 4).

Antibody language models do not yield sufficient diversity

Results with the PLM AbLang [15] differed between CDRs H1/L1 and H2/L2, which can be inferred by the V-gene evident from the unmasked framework sequence, and CDR H3/L3, which cannot. Among the former, we observed sharp drops in negative log-likelihood to PSSMs (Figure 4), along with higher V-gene sequence identity and sequence recovery relative to ProteinMPNN designs (Figure 2). In contrast, designs of the CDR H3 showed high Shannon entropy and poor recovery using AbLang.

Predictions from ESM sampled a far more diverse pool of sequences than any other method, consistent with previous results [19]. While this appeared to come at the cost of sequence recovery, a surprisingly strong and unexpected correspondence to PyIgClassify2 clusters was observed [20]. As the ESM model used here also provides the basis for the structure prediction model ESMFold [4], we surmise that these results are driven by a deeper, more general understanding of protein folding, whereas the performance of AbLang likely results from rote memorization. Such results contrast with, but do not contradict, previous reports showing poorer zero-shot capabilities among antibody-specific PLMs compared to generalist PLMs [21, 22].

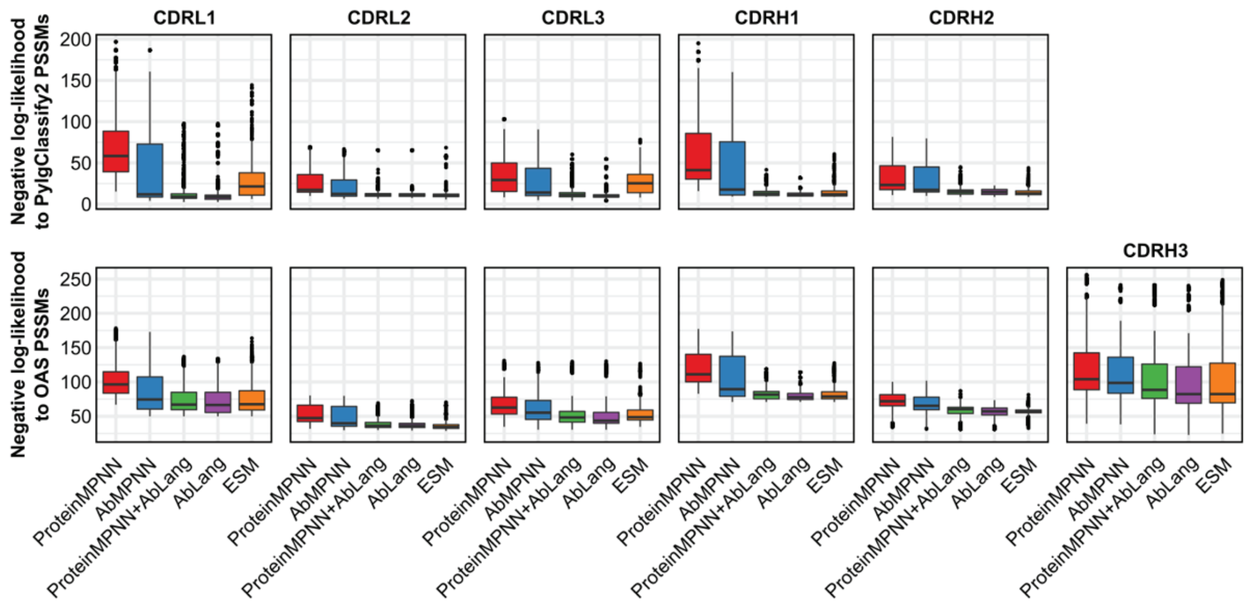

Figure 4. Antibody-specific methods generally outperform general methods in designing sequences that are more consistent with sequences sharing the same conformational clusters (top) and CDR lengths in general (bottom). CDR H3 does not form canonical clusters and was therefore omitted from the former analysis.

Combining PLMs and inverse folding

On CDRs 1 and 2, ProteinMPNN + AbLang retained the high sequence recovery of PLMs in isolation (Figure 2). On CDR H3, it matched or outperformed both individual methods on 63% of predictions, whereas predictions from either AbLang or ESM were inconclusive and diverse (Figure 5A). Design using ProteinMPNN + AbLang + ESM provided no benefit over ProteinMPNN + AbLang. AbMPNN outperformed ProteinMPNN + AbLang on some loops, but the latter provided consistent performance, whereas the former's recovery distribution was highly bimodal (Figure 5B).

A further example motivates how ProteinMPNN + AbLang arrive at different conclusions from AbMPNN. A study investigating binding of Trastuzumab variants against its target HER2 found Y105 on the CDR H3 to be indispensable to binding [23], and overrepresented among *de novo* high-affinity designs [24]. Yet a tyrosine was introduced at position 105 in only 12% of AbMPNN designs and 0% of ProteinMPNN Trastuzumab designs in our benchmark, compared to 68% of AbLang designs and 91% of ProteinMPNN+AbLang designs. But when evaluating a library of ~35,000 variants, all of which had Y105 [23], AbMPNN predictions had greater precision and recall on HER2 binding prediction (AUC: 0.793) than ProteinMPNN+AbLang predictions (AUC: 0.766; Figure 6).

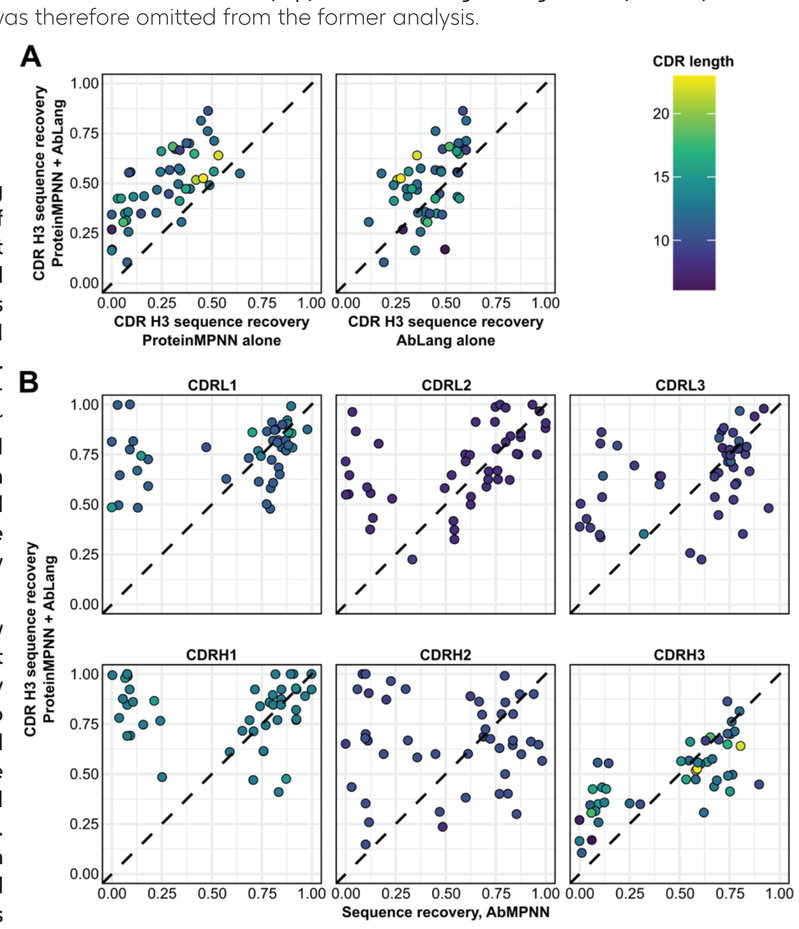

Figure 5. Comparison of ProteinMPNN+AbLang to various methods. A) Comparison in CDR H3 loop designs to ProteinMPNN alone and AbLang alone. B) Comparison of sequence recovery in all six CDRs by ProteinMPNN+AbLang, compared to AbMPNN, a bespoke version of ProteinMPNN fine-tuned exclusively on antibody structures and structural models. Results in both plots show the average values of 100 designs for each PDB structure using each method.

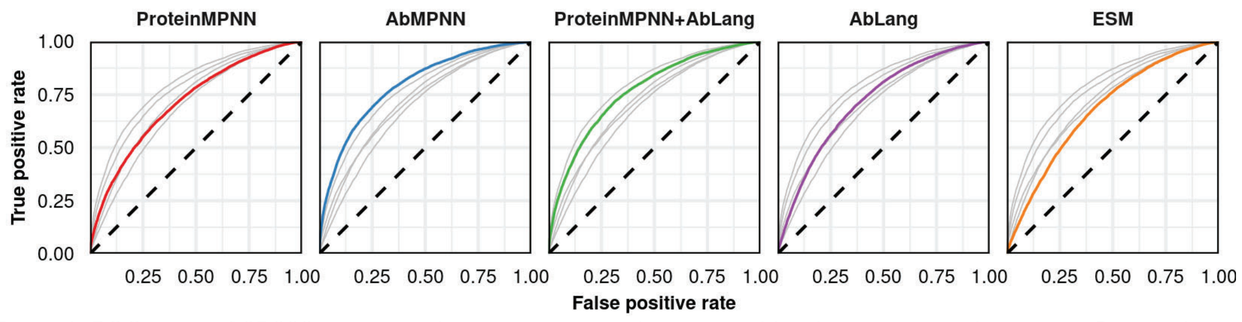

Figure 6. ROC curve of 35,000 trastuzumab mutants previously measured by deep mutational scanning. Per-residue probabilities were calculated using either the crystal structure with a fully masked CDR H3, or the complete heavy chain Fab (for ESM) or Fv (for AbLang) with fully masked CDR H3.

Concluding remarks

Recently several manuscripts have designed inverse folding models that integrate PLMs, with demonstrated improvements relative to exclusively structural information for protein design. Yet, to our knowledge, no comparisons have been carried out against naïvely ensembled sequence- and structure-based ML methods for sequence design. Such an approach could serve as a baseline against which future attempts to combine both sources of information should be compared. Here we show how it proved surprisingly effective at designing CDRs H1/L1 and H2/L2 of antibodies, with diminished effectiveness at CDR H3. This shows how inverse folding can be broadly improved using PLMs, while exposing failure points to be addressed by future ML methods trained on both sources of information.

References

[1] "Robust deep learning-based protein sequence design using ProteinMPNN." doi.org/10.1126/science.add2187
[2] "Large language models generate functional protein sequences across diverse families" doi.org/10.1038/s41587-022-01618-2
[3] "Language models generalize beyond natural proteins" doi.org/10.1101/2022.12.21.521521
[4] "Evolutionary-scale prediction of atomic-level protein structure with a language model" doi.org/10.1126/science.ade2574
[5] "Masked inverse folding with sequence transfer for protein representation learning" doi.org/10.1093/protein/gzad015
[6] "Enhancing Protein Language Models with Structure-based Encoder and Pre-training" doi.org/10.48550/arXiv.2303.06275
[7] "ProstT5: Bilingual Language Model for Protein Sequence and Structure" doi.org/10.1101/2023.07.23.550085
[8] "Learning inverse folding from millions of predicted structures" doi.org/10.1101/2022.04.10.487779
[9] "AntiFold: Improved antibody structure design using inverse folding" openreview.net/forum?id=bxZMKHtlL6
[10] "Inverse folding for antibody sequence design using deep learning" doi.org/10.48550/arXiv.2310.19513
[11] "In vitro validated antibody design against multiple therapeutic antigens using generative inverse folding" doi.org/10.1101/2023.12.08.570889
[12] "Contextual protein and antibody encodings from equivariant graph transformers" doi.org/10.1101/2023.07.15.549154
[13] "Observed Antibody Space: A Resource for Data Mining Next-Generation Sequencing of Antibody Repertoires" doi.org/10.4049/jimmunol.1800708
[14] "AbLang: an antibody language model for completing antibody sequences" doi.org/10.1093/bioadv/vbac046
[15] "Combining Structure and Sequence for Superior Fitness Prediction" openreview.net/forum?id=8PbTU4exnV
[16] "RosettaAntibodyDesign (RAbD): A general framework for computational antibody design" doi.org/10.1371/journal.pcbi.1006112
[17] "ANARCI: antigen receptor numbering and receptor classification" doi.org/10.1093/bioinformatics/btv552
[18] "Accelerated Profile HMM Searches" doi.org/10.1371/journal.pcbi.1002195
[19] "Addressing the antibody germline bias and its effect on language models for improved antibody design" doi.org/10.1101/2024.02.02.578678
[20] "A penultimate classification of canonical antibody CDR conformations" doi.org/10.1101/2022.10.12.511988
[21] "Efficient evolution of human antibodies from general protein language models" doi.org/10.1038/s41587-023-01763-2
[22] "ProGen2: Exploring the boundaries of protein language models" doi.org/10.1016/j.cels.2023.10.002
[23] "Optimization of therapeutic antibodies by predicting antigen specificity from antibody sequence via deep learning" doi.org/10.1038/s41551-021-00699-9
[24] "Antibody optimization enabled by artificial intelligence predictions of binding affinity and naturalness" doi.org/10.1101/2022.08.16.504181