IgBlend: Unifying 3D Structure and Sequence for Antibody LLMs

Cedric Malherbe **Talip Uçar**

Centre for AI, DS&AI, BioPharmaceuticals R&D, AstraZeneca
{cedric.malherbe, talip.ucar}@astrazeneca.com

Abstract

Large language models (LLMs) trained on antibody sequences have shown significant potential in the rapidly advancing field of machine learning-assisted antibody engineering and drug discovery. However, current state-of-the-art antibody LLMs often overlook structural information, which could enable the model to more effectively learn the functional properties of antibodies by providing richer, more informative data. In response to this limitation, we introduce IgBlend, which integrates both the 3D coordinates of backbone atoms (C-alpha, N, and C) and antibody sequences. Our model is trained on a diverse dataset containing over 4 million unique structures and more than 200 million unique sequences, including heavy and light chains as well as nanobodies. We rigorously evaluate IgBlend using established benchmarks such as complementarity-determining region (CDR) editing and affinity scoring and demonstrate that IgBlend consistently outperforms current state-of-the-art models across all benchmarks. Furthermore, experimental validation shows that the model's log probabilities correlate well with measured binding affinities.

1 Introduction

Antibodies are key components of the adaptive immune system, capable of recognizing and neutralizing a wide range of pathogens, including viruses, bacteria, and other foreign invaders. Their ability to bind specific targets with high affinity makes them essential tools in therapeutic development. Recent advancements in natural language processing (NLP) have led to the creation of foundational language models that can learn from and modify antibody sequences [Olsen et al., 2022b, 2024, Prihoda et al., 2022]. Moreover, the three-dimensional (3D) structure of an antibody is closely linked to its specificity, affinity, and interaction with antigens. Therefore, capturing the relationship between sequence and structure is crucial for tasks such as affinity maturation, de novo antibody design, and optimizing antibody-antigen interactions for therapeutic applications. While current language models excel at either sequence-to-sequence or structure-to-sequence (inverse folding) tasks, relying on only one of these modalities at the input limits their capability and flexibility in more complex antibody engineering tasks [Olsen et al., 2022b, 2024, Prihoda et al., 2022, Høie et al., 2023]. In this paper, we introduce IgBlend, a multi-modal model designed to incorporate both sequence and structural information for antibody engineering. Our approach can utilize either sequence, structure, or both, enabling the model to not only sample sequences that can fold to the same parental backbone but also generate more diverse sequences, providing greater flexibility in designing antibody sequences. Moreover, by utilizing both experimentally resolved structures [Dunbar et al., 2014] and synthetic data generated through structure prediction models [Abanades et al., 2023b, Ruffolo et al., 2023], we aim to improve model performance on key antibody engineering tasks. Our contributions can be summarized as follows:

Machine Learning for Structural Biology Workshop, NeurIPS 2024.

- We introduce IgBlend, a model that learns antibody representations from either sequence, structure or sequence-structure pairs when structural data is available.
- We present a pre-training strategy with multiple sub-objectives as well as a procedure for training and dataset processing, all of which can broadly be applied to other multi-modal training settings.
- We empirically demonstrate that integrating structural information, even when synthetically generated, significantly improves the performance of large models across a wide range of benchmarks.
- We show that IgBlend 's log probabilities correlate well with measured binding affinities.

To save space, we differ the background, notations and related works to the Appendix.

2 Methods

2.1 Model architecture

The proposed architecture, IgBlend, is illustrated in Fig 1 and consists of three primary components: a structure encoder that handles the backbone coordinates of the antibody, a sequence encoder that processes the amino acid sequence, and a multi-modal trunk that processes the combined structural and sequential representations. Below, we provide a detailed description of these key components.

Structure encoder. The structure encoder generates an abstract representation vector for each set of coordinates $\mathbf{x}_i \in \mathbb{R}^{3\times3}$ from the full sequence of coordinates $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \mathbb{R}^{3\times3\times n}$. This representation (a 512-dimensional embedding) encapsulates the geometry of the global backbone structure. The architecture comprises four GVP-GNN (Graph Neural Network Geometric Vector Perceptron) layers [Jing et al., 2020], followed by two generic Transformer encoder layers [Vaswani et al., 2017]. This design is invariant to rotation and translation of the input coordinates and has been demonstrated to effectively capture protein geometries in various learning tasks [Jing et al., 2020], including structure-to-sequence models such as ESM-inverse folding [Hsu et al., 2022] and AntiFold [Høie et al., 2023]. The input to the encoder is the series of residue coordinates $\mathbf{x}$, and a local reference frame is established for each amino acid, following the approach used in AlphaFold2 [Jumper et al., 2021]. A change of basis is then performed according to this local reference frame, rotating the vector features from the GVP-GNN outputs into the local reference frames of each amino acid. Finally, the output of the GVP is processed through two Transformer blocks, producing a 512-dimensional embedding for each residue. Notably, each or all sets of coordinates can be masked using the $*$ token.

Sequence encoder. In parallel to the structure encoder, the sequence encoder generates a vector representation (i.e., embedding of size 512) for each amino acid $\mathbf{s}_i \in \mathbb{A}$ in the full sequence $\mathbf{s} = (\mathbf{s_1}, \ldots, \mathbf{s_n})$. The architecture consists of a one-hot encoding embedding followed by two blocks

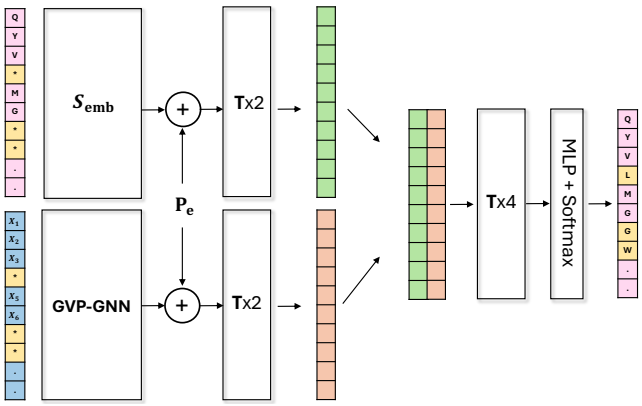

Figure 1: Architecture of the IgBlendmodel. It takes as input both: a series of amino acids (top) and a series of 3D coordinates (bottom). The symbol * denotes either a masked amino acid or a masked set of coordinates. Note that the model can process each modality independently by setting all the tokens of one modality to mask. S_e denotes the sequence embedding (i.e. look-up table), T denotes a transformer block, P_e denotes the sinusoidal position embedding. The sequence encoder is displayed of the bottom left, the structure encoder on the bottom left and on the right the multi-modality processor.

of a standard Transformer model [Vaswani et al., 2017]. This architecture has already been shown to learn relevant information from antibody sequences in [Olsen et al., 2022b] and [Olsen et al., 2024]. Specifically, the module utilizes sinusoidal positional embeddings, a SwiGLU activation function [Shazeer, 2020], and has an embedding dimension of 512. Additionally, any amino acid within the sequence can be masked using the masked token $*$.

Multi-modality encoder. The fusion layer processes both modalities in two steps. First, it combines the abstract representations from the sequence and structure encoders by concatenating them along the embedding dimension, forming a single vector of size 1024 for each residue. It then processes the concatenated modalities through a series of four Transformer blocks using a SwiGLU activation function.

Classification head. Finally, the classification head consists of a multi-layer perceptron (MLP) followed by a softmax function and processes the multi-modal representation to generate a probability distribution over amino acid types at each position. Further details on the architecture can be found in Appendix B.

2.2 Datasets and pre-training objectives

The model was trained on more than 4M structures and and more than 200M sequences as detailed in Appendix C. To train the multi-modal IgBlend, we use a balanced representation of heavy and light chains. We employed a specialized masked language modeling objective capable of handling both sequential and structural data by minimizing the sum of three losses based on cross-entropy:

$$\mathcal{L}_{\text{multi-modal}} := \mathcal{L}_{\text{seq2seq}} + \mathcal{L}_{\text{seq+struct2seq}} + \mathcal{L}_{\text{struct2seq}}$$

where $\mathcal{L}_{\text{seq2seq}}$ denotes a sequence to sequence objective, $\mathcal{L}_{\text{seq+struct2seq}}$ denotes a sequence plus structure to sequence objective and $\mathcal{L}_{\text{struct2seq}}$ denotes the structure to sequence task. Hence, the model learns to perform all these tasks in parrallel. For completeness, all the individual losses are fully described in Appendix D.

3 Empirical results

In this section, we evaluate the impact of incorporating structural information into the pre-training of antibody LLMs. Our evaluation focuses on two tasks: (i) editing CDRs, and (ii) scoring sequences for HER2 binding. We compare the performance of IgBlend with five existing open-source antibody and nanobody language models, including AbLang [Olsen et al., 2022b], AbLang2 [Olsen et al., 2024], AntiBERTy [Ruffolo et al., 2021], Sapiens [Prihoda et al., 2022], Nanobert [Hadsund et al., 2024], which is a nanobody specific LLM, and two inverse folding models, including AntiFold [Høie et al., 2023] and ESM-IF [Hsu et al., 2022].

3.1 Editing Complementarity-determining region (CDR)

First, we focused on the task of editing/recovering the CDR regions of a single chain, which is of particular importance in the process of optimizing antibodies for affinity. In this task, one of the CDR region is randomly fully masked, i.e., we select a mask $\mathcal{M}_s \in \{\text{CDR1, CDR2, CDR3}\}$, and the models are asked to predict the masked residues within that fully masked CDR. To investigate the impact of different modalitites set as input, each model recovers the sequence as follows: $\widehat{\mathbf{s}} = \text{Model}(\mathbf{s}_{/\mathcal{M}_s})$ for seq-only models, $\widehat{s} = \text{Model}(\mathbf{s}_{/\mathcal{M}_s}, \mathbf{x})$ for structure guided sequential models and $\widehat{s} = \text{Model}(\mathbf{x})$ for inverse folding models. We evaluated the models using the same inputs for seq-only, structure guided and inverse folding models over the 1.000 sequences sampled from the test using Equation (1) for each chain type, unseen during the training of IgBlend, and we recorded the percentage of successfully recovered residues. The results can be found in Figure 2 with all models being evaluated on the same masked sequences. To further investigate the models' ability to design sequences that can fold into the same backbone conformation, we conducted a consistency check between generated sequences and their structure $\mathbf{x}$. To do so, we kept the best models in each category for comparison: AbLang2 for heavy/light chains, Nanobert for nanobodies and AntiFold for inverse folding. We then sampled 500 sequences per chain type from the test distribution and asked the model to recover a masked CDR region. Then, for each of recovered sequence $\widehat{s}$, we computed its

structural approximation $\widehat{\mathbf{x}} = \text{IgFold}(\widehat{\mathbf{s}})$ using IgFold with PyRosetta refinement. Finally, given a recovered sequence and structure pair $(\widehat{\mathbf{s}}, \widehat{\mathbf{x}})$, the ground truth $(\mathbf{s}, \mathbf{x})$ and the masked CDR region $\mathcal{M}_s$, we computed the Levenshtein$(\{\mathbf{s}_i, i \in \mathcal{M}_s\}, \{\widehat{\mathbf{s}}_i, i \in \mathcal{M}_s\})$ distance between sequences to measure diversity as well as the RMSD$(\{\mathbf{x}_i, i \in \mathcal{M}_s\}, \{\widehat{\mathbf{x}}_i, i \in \mathcal{M}_s\})$ between the original and predicted structures as a proxy for the structural similarity for the masked region. Results are also collected in Figure 2 and extended results can be found in Appendix. A few key observations emerged:

- First, similar to the previous experiments, the top-performing sequence-only models (AbLang, AbLang2, AntiBERTy, IgBlend) exhibited comparable performance across the different CDR regions. It is noteworthy, however, that IgBlend displays an accuracy of more than 9% above the best performing model on nanobodies. We also consistently observe that incorporating more information as input improves the performance of IgBlend for all chain types (i.e. IgBlend(Seq+Struct Guided)>IgBlend(Seq+Masked Struct)>IgBlend(Seq-only)). Moreover, incorporating structural information alongside the masked sequence (IgBlend(seq+struct guidance)) considerably improves the performance over the best seq-only models (i.e. 11.8% on CDR3-H, 6.74% on CDR3-L and 15.43 on CDR3-N)
- Second, we observe that IgBlend (Struct Guided)—which uses both sequential information $(\mathbf{s}_{\mathcal{M}_s})$ and structural information $(\mathbf{x}_{\mathcal{M}_x})$—performs more similarly to IgBlend (inverse folding), which uses only structure $\mathbf{x}$, than to IgBlend (seq-only), which uses only sequence $\mathbf{s}_{\mathcal{M}_s}$. This suggests that when re-editing entire CDR regions, IgBlend relies more on structural information than on sequential information.
- In terms of structural similarity, we note that IgBlend (structure guided) achieves the highest percentage of structures within the lowest RMSD bin for each chain type, even surpassing the best inverse folding models. Specifically, IgBlend records 37%, 46%, and 51% for H-CDR3, L-CDR3, and N-CDR3, respectively, compared to AntiFold's 32%, 5%, and 26%. This indicates that, in addition to outperforming sequential models, IgBlend (seq+Struct guided) generates sequences with greater structural accuracy than AntiFold.

Mode	Model	Heavy			Light			Nanobody		
		CDR1	CDR2	CDR3	CDR1	CDR2	CDR3	CDR1	CDR2	CDR3
Sequence Only	**AbLang**	82.97	**80.53**	41.68	72.21	69.27	67.47	43.73	45.09	20.90
	AbLang2	82.85	80.31	41.62	72.94	69.66	68.03	43.05	41.43	20.16
	Antiberty	82.90	80.37	41.23	72.64	69.20	68.61	40.48	47.76	23.12
	Sapiens	81.44	77.13	38.45	71.18	67.22	63.03	44.25	39.99	19.79
	Nanobert	57.33	40.00	24.02	10.16	08.53	07.22	60.49	61.09	29.08
	IgBlend	**83.15**	80.33	**41.84**	**73.14**	**69.79**	**68.70**	**62.58**	**63.81**	**29.53**
Inverse Folding	**AntiFold**	75.41	70.99	36.97	57.05	58.98	49.12	44.70	44.92	22.02
	ESM-IF	49.90	44.19	19.65	33.68	43.70	31.46	30.74	39.98	15.34
	IgBlend	**86.18**	**84.44**	**52.69**	**76.69**	**82.03**	**73.9**	**69.72**	**72.58**	**43.77**
Seq + Masked Struct	**IgBlend**	84.00	80.61	43.37	74.00	73.10	70.61	65.93	64.75	32.28
Seq + Struct Guided	**IgBlend**	**87.27**	**85.04**	**53.65**	**77.08**	**83.59**	**75.44**	**71.40**	**73.52**	**44.96**

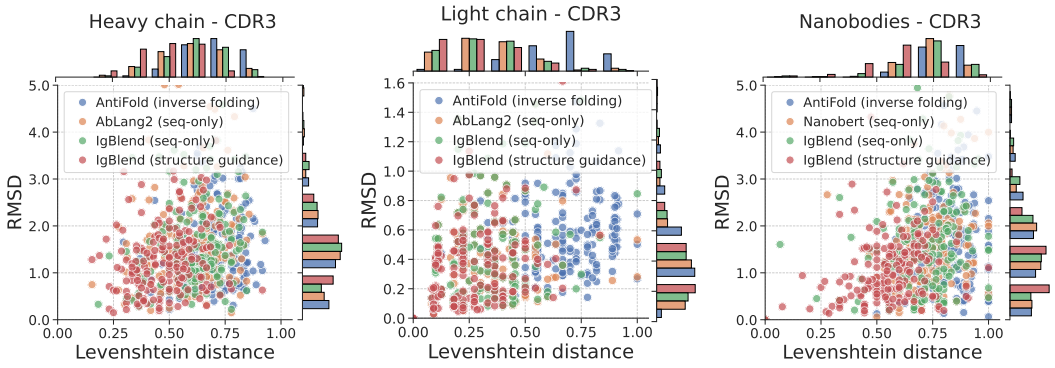

Figure 2: **CDR in-filling results:** One CDR region (CDR1, CDR2, or CDR3) is fully masked, and the model attempts to recover it. **Top:** The table shows the average percentage of correctly recovered residues for heavy chain (H), light chain (L) and nanobodies (N). **Bottom:** The graphs display the Levenshtein distances and RMSE in the masked CDR3 regions for each chain type.

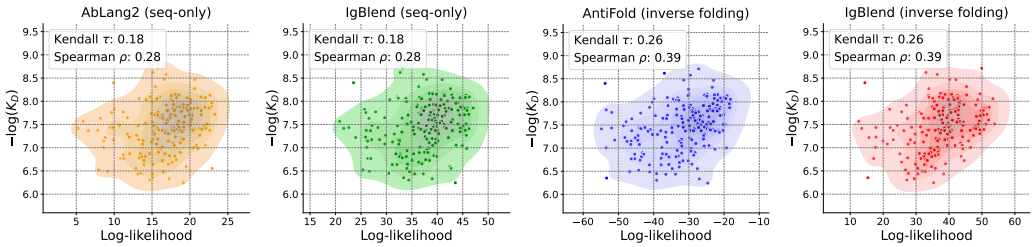

Figure 3: **Trastuzumab-HER2 H-CDR3 editing on zero-shot dataset.** Each model scores sequences using their log-likelihood on HCDR3. The scatter plot displays the log-probability (score) of each sequence on the x-axis and the $-\log(K_D)$ values on the y-axis. Additionally, the density of the point cloud is displayed.

3.2 Human epidermal growth factor receptor (HER2): H-CDR3 design

Finally, we assessed IgBlend's ability to score sequences for H-CDR3 design using experimental data. We utilized the dataset published by Shanehsazzadeh et al. [2023], where a machine learning model is used to re-design H-CDR3 sequences targeting HER2. Specifically, they chose the therapeutic antibody trastuzumab, which targets HER2, as a template and re-designed the heavy chain's CDR3, conditioned on the modeled HCDRs were conditioned on the HER2 antigen backbone structure derived from PDB:1N8Z (Chain C), trastuzumab framework sequences, and the trastuzumab-HER2 epitope. The K_D values for the generated sequences were then measured using a Fluorescence-activated Cell Sorting (FACS)-based ACE assay. To evaluate the ability of antibody language models to pre-screen sequences likely to show binding affinity, we scored the generated sequences in the zero-shot dataset (maintaining the same H-CDR3 length as trastuzumab) by calculating their log-likelihood in the HCDR3 region with both AbLang2 and IgBlend (seq-only). To determine if incorporating structural information improves scoring, we also evaluated the sequences using IgBlend (inverse folding), where trastuzumab's backbone structure $\mathbf{x}$ approximated using IgFold was provided as input, guiding the model to favor sequences with structures similar to trastuzumab and AntiFold. The $-\log(K_D)$ values against the likelihood scores from each model are displayed in Figure 3 as well as the Spearman correlation and Kendall tau. Note also that additional results computed on the associated control dataset can be found in Appendix E.2.

- First, it is evident that sequence-only models, such as AbLang2 and IgBlend (seq-only), offer limited utility in predicting which sequences are likely to have strong binding affinities. This can be attributed to the fact that these models are trained on sequence data alone, with no *explicit* information about conformational states that would favor binding to HER2. As a result, they tend to generate biologically plausible sequences without prioritizing binding affinity (since much of the training set is derived from the OAS dataset, which lacks binding information).
- In contrast, structure-guided models such as IgBlend (inverse folding) and AntiFold demonstrate a stronger positive correlation between binding affinity and scoring, indicating their greater accuracy in identifying high-affinity sequences. IgBlend (inverse folding), which relies solely on backbone structure, further emphasizes the critical role of structural context in guiding models toward more favorable binding configurations.

This contrast highlights that while sequence-only models are effective at generating biologically viable sequences, structure-based models are superior for evaluating binding affinity, emphasizing the need to integrate structural data for more accurate predictions in H-CDR3 design.

4 Conclusion and future work

In this study, we explored the impact of incorporating structural information into antibody LLMs to improve their performance. We outlined the pre-training objectives and compared our model against existing sequence-based and inverse folding models, providing empirical evidence that structural guidance consistently improves performance across all benchmarks. However, we note that these performance gains come at the expense of reduced sequence diversity. Additionally, we showed that log-likelihood can effectively be used to rank sequences for binding affinity, with structure-based models showing a higher correlation with experimentally validated data. In future work, we aim to incorporate side-chain information and expand the structural datasets to enhance model accuracy.

References

Brennan Abanades, Tobias H Olsen, Matthew I J Raybould, Broncio Aguilar-Sanjuan, Wing Ki Wong, Guy Georges, Alexander Bujotzek, and Charlotte M Deane. The Patent and Literature Antibody Database (PLAbDab): an evolving reference set of functionally diverse, literature-annotated antibody sequences and structures. *Nucleic Acids Research*, 52(D1):D545–D551, 11 2023a. ISSN 0305-1048. doi: 10.1093/nar/gkad1056. URL `https://doi.org/10.1093/nar/gkad1056`.

Brennan Abanades, Wing Ki Wong, Fergus Boyles, Guy Georges, Alexander Bujotzek, and Charlotte M Deane. ImmuneBuilder: Deep-learning models for predicting the structures of immune proteins. *Communications Biology*, 6(1):575, 2023b.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. 2018.

James Dunbar and Charlotte M Deane. ANARCI: antigen receptor numbering and receptor classification. *Bioinformatics*, 32(2):298–300, 2016.

James Dunbar, Konrad Krawczyk, Jinwoo Leem, Terry Baker, Angelika Fuchs, Guy Georges, Jiye Shi, and Charlotte M Deane. SAbDab: the structural antibody database. *Nucleic acids research*, 42(D1):D1140–D1146, 2014.

François Ehrenmann, Patrice Duroux, Véronique Giudicelli, and Marie-Paule Lefranc. Standardized sequence and structure analysis of antibody using IMGT®. *Antibody engineering*, pages 11–31, 2010.

Daria Frolova, Marina Pak, Anna Litvin, Ilya Sharov, Dmitry Ivankov, and Ivan Oseledets. Mulan: Multimodal protein language model for sequence and structure encoding. *bioRxiv*, pages 2024–05, 2024.

Johannes Thorling Hadsund, Tadeusz Satława, Bartosz Janusz, Lu Shan, Li Zhou, Richard Röttger, and Konrad Krawczyk. nanobert: a deep learning model for gene agnostic navigation of the nanobody mutational space. *Bioinformatics Advances*, 4(1):vbae033, 2024.

Tomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, et al. Simulating 500 million years of evolution with a language model. *bioRxiv*, pages 2024–07, 2024.

Michael Heinzinger, Konstantin Weissenow, Joaquin Gomez Sanchez, Adrian Henkel, Milot Mirdita, Martin Steinegger, and Burkhard Rost. Bilingual language model for protein sequence and structure. *bioRxiv*, pages 2023–07, 2023.

Magnus Høie, Alissa Hummer, Tobias Olsen, Morten Nielsen, and Charlotte Deane. Antifold: Improved antibody structure design using inverse folding. In *NeurIPS 2023 Generative AI and Biology (GenBio) Workshop*, 2023.

Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, and Alexander Rives. Learning inverse folding from millions of predicted structures. In *International conference on machine learning*, pages 8946–8970. PMLR, 2022.

Bowen Jing, Stephan Eismann, Patricia Suriana, Raphael John Lamarre Townshend, and Ron Dror. Learning from protein structure with geometric vector perceptrons. In *International Conference on Learning Representations*, 2020.

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.

Wolfgang Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, 32(5):922–923, 1976.

Tobias H Olsen, Fergus Boyles, and Charlotte M Deane. Observed Antibody Space: A diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. *Protein Science*, 31(1):141–146, 2022a.

Tobias H Olsen, Iain H Moal, and Charlotte M Deane. AbLang: an antibody language model for completing antibody sequences. *Bioinformatics Advances*, 2(1):vbac046, 2022b.

Tobias Hegelund Olsen, Iain H Moal, and Charlotte Deane. Addressing the antibody germline bias and its effect on language models for improved antibody design. *bioRxiv*, pages 2024–02, 2024.

David Prihoda, Jad Maamary, Andrew Waight, Veronica Juan, Laurence Fayadat-Dilman, Daniel Svozil, and Danny A Bitton. BioPhi: A platform for antibody design, humanization, and humanness evaluation based on natural antibody repertoires and deep learning. In *MAbs*, volume 14, page 2020203. Taylor & Francis, 2022.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE, 2020.

Jeffrey A Ruffolo, Jeffrey J Gray, and Jeremias Sulam. Deciphering antibody affinity maturation with language models and weakly supervised learning. *arXiv preprint arXiv:2112.07782*, 2021.

Jeffrey A Ruffolo, Lee-Shin Chu, Sai Pooja Mahajan, and Jeffrey J Gray. Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies. *Nature communications*, 14(1):2389, 2023.

Amir Shanehsazzadeh, Sharrol Bachas, Matt McPartlon, George Kasun, John M Sutton, Andrea K Steiger, Richard Shuai, Christa Kohnert, Goran Rakocevic, Jahir M Gutierrez, et al. Unlocking de novo antibody design with generative artificial intelligence. *bioRxiv*, pages 2023–01, 2023.

Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

Martin Steinegger and Johannes Söding. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology*, 35(11):1026–1028, 2017.

Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. Saprot: Protein language modeling with structure-aware vocabulary. *bioRxiv*, pages 2023–10, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

A Background, related work and notations

Background on antibodies. In humans, antibodies are classified into five isotypes: IgA, IgD, IgE, IgG, and IgM. This work primarily focuses on IgG antibodies, which are Y-shaped glycoproteins produced by B-cells (see Figure 4), as well as nanobodies, which are antibody fragments consisting of a single monomeric variable domain. Henceforth, "antibody" will specifically refer to IgG antibodies. Antibodies consist of distinct regions that play specific roles in the immune response. The Fab (fragment antigen-binding) region, composed of both variable (V) and constant (C) domains from the heavy and light chains, is primarily responsible for antigen binding. Within this region, the antigen-binding site is formed by the variable domains — VH for the heavy chain and VL for the light chain — which determine the specificity of the antibody and enable it to recognize and bind to specific antigens. The Fv (fragment variable) region is the smallest functional unit of an antibody that can still bind to an antigen. It consists solely of the variable domains (VH and VL) of the heavy and light chains, without the constant domains. Within the variable domains, there are two key distinct regions: the framework regions and the complementarity-determining regions (CDRs). The framework regions provide structural support, maintaining the overall shape of the variable domains, while the CDRs, comprising three loops on both the VH and VL chains, are directly involved in binding to the antigen. These CDRs are crucial for the precise recognition and interaction with specific antigens. While the Fv region is essential for the initial recognition and binding of antigens, it lacks the effector functions present in the full antibody. The Fab region, being larger and more complex due to the inclusion of both variable and constant domains, is generally more stable and has a higher affinity for antigens. The Fv region, on the other hand, is simpler and more easily engineered for various applications, such as in the development of single-chain variable fragment (scFv) antibodies. The base of the Y-shaped antibody, known as the Fc (fragment crystallizable) region, is involved in regulating immune responses. It interacts with proteins and cell receptors, ensuring that the antibody generates an appropriate immune response. Moreover, nanobodies, which are small, single-domain antibodies derived from heavy-chain-only antibodies found in certain animals such as camels and llamas, are even more compact than traditional Fv regions. They retain full antigen-binding capacity while offering advantages such as increased stability and easier production, making them valuable tools in both therapeutic and diagnostic applications.

Related work. In recent years, significant efforts have been made to develop antibody foundation models by adapting approaches from natural language processing (NLP). For instance, AbLang [Olsen et al., 2022b], a BERT-like model [Devlin et al., 2018], is specifically trained on sequences from the immunoglobulin protein superfamily. It has proven useful in various tasks, such as restoring missing residues, analyzing affinity maturation trajectories, and identifying paratope residues—those involved in antigen recognition. Similarly, AntiBERTy [Ruffolo et al., 2021] is another BERT-based model designed for antibody-related applications, while Sapiens [Prihoda et al., 2022] is a specialized language model tailored for immunoglobulins. Recent research has also addressed challenges like germline bias and optimized predictions for non-germline residues [Olsen et al., 2024]. Collectively, these models have advanced our understanding of antibody diversity, maturation, and binding properties, significantly impacting the field of drug discovery and demonstrating the potential of language models in antibody research. In parallel and concurrently following the advancements of AlphaFold [Jumper et al., 2021], recent work has focused on inferring protein structure from sequences only, leading to models like ImmuneBuilder [Abanades et al., 2023b] and IgFold [Ruffolo et al., 2023]. These models have made it possible to generate high quality inferred structures on a large scale, a feat previously limited by the scarcity of structural data. Conversely, structural systems that can generate a sequences from structure only such as ESM-IF [Hsu et al., 2022] and [Høie et al., 2023] have also been developed. Due to the success of structural systems, there has been a growing interest in the recent years in directly incorporating the knowledge about the protein structure into protein LLMs to increase their capacities. For instance, ProtT5 [Heinzinger et al., 2023], SaProt [Su et al., 2023], MULAN [Frolova et al., 2024] and ESM3 [Hayes et al., 2024] are early protein models that incorporate the structural information through the use of structural tokens learned independently of the overall system or by plugging a structural adapted. However, to the best of our knowledge, the questions on how to design such systems for antibodies and on the potential benefits are still open and there is currently no antibody LLM that integrates both structural and sequential information. In this paper, we follow this route and show that we can directly learn a joint representation of both the structure and sequence in the pre-training phase of antibodies LLMs. As a result, we show that such systems improve on both structural and sequential only antibodies models.

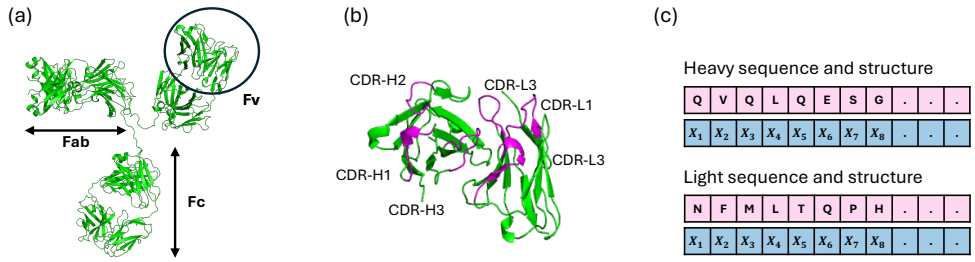

Figure 4: (a) Antibody structure with antigen binding (Fab), crystallizable (Fc), and variable (Fv) regions, (b) Zoom over the variable region which contains an heavy and a light chain, CDRs regions are displayed in magenta, (C) Modalities that we exploit in this paper for antibody modeling.

Notations. For any single unpaired chain (heavy, light or nanobody), we denote the backbone structure and sequence of the chain with n residues as follows:

$$\text{structure: } \mathbf{x} := (\mathbf{x}_1, \dots, \mathbf{x}_n) \in \mathbb{R}^{3\times 3\times n} \text{ and sequence: } \mathbf{s} := (\mathbf{s}_1, \dots, \mathbf{s}_n) \in \mathbb{A}^n$$

where $\mathbf{x}_i \in \mathbb{R}^{3\times 3}$ represents the 3D coordinates of the C-alpha, N, and C atoms of the i^{th} residue, while $\mathbf{s}_i \in \mathbb{A} :=$ [A, R, N, D, C, E, Q, G, H, I, L, K, M, F, P, S, T, W, Y, V, *] specifies the amino acid type corresponding to the i^{th} residue, where $i \in \{1, \dots, n\}$. For consistency in notation, we will also use $*$ to denote the unknown token for both structure and sequence tokens, acknowledging a slight abuse of notation. Moreover, we stress that this work solely focuses on unpaired sequences and leaves the fine-tuning on purely paired sequences for future work. Throughout the rest of this paper, we will also use $\mathbb{P}$, $\mathbb{E}$ and $\mathbb{I}$ to represent the standard probability, expectation and indication function taking values in $\{0, 1\}$, respectively. To compute the differences between two sequences $(\mathbf{s}, \widehat{\mathbf{s}}) \in \mathbb{A}^{|\mathbf{s}|}$ of the same length, we will use the normalized Levenshtein distance: $\text{Levenshtein}(\mathbf{s}, \widehat{\mathbf{s}}) = (1/|\mathbf{s}|) \cdot \sum_{i=1}^{|\mathbf{s}|} \mathbb{I}\{\mathbf{s}_i \neq \widehat{\mathbf{s}}_i\}$. To compute differences between two backbone structures $(\mathbf{x}, \widehat{\mathbf{x}}) \in \mathbb{R}^{3\times 3\times |\mathbf{x}|}$, we will also use the Root Mean Square Deviation (RMSD) defined as $\text{RMSD}(\mathbf{x}, \widehat{\mathbf{x}}) = \arg\min_{R\in\Omega_3, t\in\mathbb{R}^3} (1/3|\mathbf{x}|) \cdot \sum_{i\leq|\mathbf{x}|, j\leq 3} \|\mathbf{x}_i^j - R^*\widehat{\mathbf{x}}_i^j - t^*\|_2^2)^{1/2}$ where $R^* \in \mathbb{R}^{3\times 3}$ and $t^* \in \mathbb{R}^3$ respectively denote the optimal rotation matrix and translation after finding the optimal rigid alignment with the Kabsch algorithm [Kabsch, 1976] between the backbone structures where $\Omega_3 \subset \mathbb{R}^{3\times 3}$ denotes the set of 3D rotations and $\|\cdot\|_2$ denotes the standard Euclidean distance.

B Architectural details

We collect here the full details of the IgBlendarchitecture used in the paper, discribed in Table 1.

C Data preparation

Data source. To create a model capable of processing both sequential and structural information, we needed to address the significant asymmetry in the availability of data across these modalities (204M sequences and 3M structures as shown in Table 2). Therefore, we compiled two datasets: (1) a structural dataset $\mathcal{D}_{\text{struct}}$, which includes structures paired with their corresponding sequences, and (2) a sequential dataset $\mathcal{D}_{\text{seq}}$, which consists solely of sequence data. These datasets were derived from four primary sources: SAbDab [Dunbar et al., 2014], which contains experimentally determined structures using techniques like electron crystallography and X-ray diffraction; PLAbDab [Abanades et al., 2023a], which provides sequences derived from patents; OAS datasets [Olsen et al., 2022a], which compile and annotate immune repertoires; and INDI, which contains sequences of nanobodies. Given the relatively small number of experimentally determined structures (e.g., only about 2,000 samples from SAbDab as shown in Table 2), we expanded our structural dataset by incorporating inferred structures. In addition to the inferred structures already present in the PLAbDab dataset (folded with ImmuneBuilder), we generated additional structures from the OAS paired, unpaired and INDI. The paired sequences from OAS were folded with ImmuneBuilder [Abanades et al., 2023b] and a clustered version of the unpaired OAS and INDI dataset were folded using IgFold [Ruffolo et al.,

Structure module	value
gvp_eps	0.0001
gvp_node_hidden_dim_scalar	512
gvp_node_hidden_dim_vector	256
gvp_num_encoder_layers	4
gvp_dropout	0.1
gvp_encoder_embed_dim	512
transformer_encoder_layers	2
encoder_embed_dim	512
transformer_dropout	0.1
encoder_attention_heads	8
encoder_ffn_embed_dim	1024
Sequence Module	
d_model	512
dropout	0.1
layer_norm_eps	0.0001
nhead	8
activation	SwiGLU
dim_feedforward	512
layer_norm_eps	0.0001
Multi-modal encoder	
d_model	1024
num_layers	4
n_head	16
dim_feedforward	1024
activation	SwiGLU
prediction_head	
d_model	1024
activation	GELU

Table 1: Hyper-parameters of IgBlend.

2023]. This process resulted in approximately 4 million unique structures. For the sequential dataset, we extracted data from four repertoires: OAS paired, OAS unpaired, PLAbDab paired, PLAbDab unpaired and INDI.

Modality	Heavy sequences	Light sequences	Heavy structures	Light structures
OAS paired	1 804 122	443 129	1 418 312	535 130
OAS unpaired	156 314 998	34 464 420	1 057 850	643 647
PLAbDab paired	51 740	45 620	47 554	42 021
PLAbDab unpaired	139 706	89 743	-	-
INDI (nanobodies)	11 231 660	-	895 008	-
SAbDab	-	-	2 056	2 024
Total	169 542 226	35 042 912	3 420 780	1 222 822

Table 2: Number of unique samples per modalities and chain types after the first pre-processing step.

Data processing. For each of the datasets $\mathcal{D}_{\text{struct}}$ and $\mathcal{D}_{\text{seq}}$, we begin by removing all duplicates, defined as pairs of data with identical sequences. Next, only the data that meet the following criteria are retained: (1) no unknown residues, (2) no missing residues, and (3) no shorter than expected IMGT regions [Ehrenmann et al., 2010], as determined by running ANARCI [Dunbar and Deane, 2016]. After these cleaning steps, we are left with two datasets: $\mathcal{D}_{\text{struct}} = \{(\mathbf{s}, \mathbf{x})_1, \ldots, (\mathbf{s}, \mathbf{x})_{|\mathcal{D}_{\text{struct}}|}\}$, which contains pairs of sequences and structures, and $\mathcal{D}_{\text{seq}} = \{(s, *)_1, \ldots, (s, *)_{|\mathcal{D}_{\text{seq}}|}\}$, which contains only sequential information. These datasets are then further divided into heavy, light and nanobodies chain samples, resulting in $\mathcal{D}_{\text{struct}} = \mathcal{D}_{\text{struct,H}} \cup \mathcal{D}_{\text{struct,L}} \cup \mathcal{D}_{\text{struct,N}}$ and $\mathcal{D}_{\text{seq}} = \mathcal{D}_{\text{seq,H}} \cup \mathcal{D}_{\text{seq,L}} \cup \mathcal{D}_{\text{seq,N}}$. The number of unique samples remaining in each dataset is summarized in Table 2. Due to the significant imbalance in the number of samples across modalities, as noted in Table 2, we implemented a new sampling scheme to rebalance the data. For each modality M $\in \{\text{seq}, \text{struct}\}$ and each chain type C $\in \{\text{L}, \text{H}, \text{N}\}$, we clustered the datasets $\mathcal{D}_{\text{M,C}}$ using MMseqs2 [Steinegger

and Söding, 2017], clustering over the full sequences with the parameters "−cov-mode 1", "−c 0.8", and "−min_seq_id 0.8" for the sequential datasets and over the concatenated CDR regions with the parameter "−min_seq_id 0.9" for the structure datasets. This process resulted in a union of n_{cluster} clustered samples $\mathcal{D}_{\text{M,C}} = \bigcup_{i=1}^{n_{\text{cluster}}} \mathcal{C}_{\text{M,C}}(i)$ for each modality and chain type. Based on these clusters, we defined the distributions $\mathcal{P}(\mathcal{D}_{\text{struct}})$ and $\mathcal{P}(\mathcal{D}_{\text{seq}})$ over each dataset modality as follows: first, we sample a chain type C with equal probability: $\mathbb{P}(\text{C = H}) = \mathbb{P}(\text{C = L}) = \mathbb{P}(\text{C = N}) = 1/3$, then we select a sample within the corresponding dataset $\mathcal{D}_{\text{C,M}}$ according to the size of its corresponding cluster:

$$\mathbb{P}(\mathbf{s}, \mathbf{x})_{|\text{M,C}} = \begin{cases} 1/|\mathcal{C}_{\text{M,C}}(i_s)| & \text{if } (\mathbf{s}, \mathbf{x}) \in \mathcal{D}_{\text{M,C}} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where i_s denotes the index of the cluster containing $\mathbf{s}$, and $|\mathcal{C}_{\text{M,C}}(i_s)|$ indicates the size of its corresponding cluster. This clustering-based distribution approach allows us to preserve the entire dataset while re-weighting each cluster to enhance diversity in the training set. Additionally, during the clustering process, 10 of the clusters are reserved for validation and another 10 for testing. The reserved clusters are entirely excluded from the training set and have less than 0.8 sequence identity with the training data, forcing the validation and test sets to be too dissimilar from the training set.

D Pre-training objectives

To train the multi-modal IgBlend, we use the data distribution defined by Equation (1), ensuring a balanced representation of heavy and light chains across the two datasets, $\mathcal{D}_{\text{seq}}$ and $\mathcal{D}_{\text{struct}}$. We employ a specialized masked language modeling objective capable of handling both sequential and structural data. The model parameters, θ, are optimized by minimizing the sum of three losses based on cross-entropy:

$$\mathcal{L}_{\text{multi-modal}} := \mathcal{L}_{\text{seq2seq}} + \mathcal{L}_{\text{seq+struct2seq}} + \mathcal{L}_{\text{struct2seq}} \tag{2}$$

where:

$$\begin{cases} \mathcal{L}_{\text{seq2seq}} & = & \mathbb{E}_{(\mathbf{s},*)\sim\mathcal{P}(\mathcal{D}_{\text{seq}})} \left[\sum_{i \in \mathcal{T}_s} -\log(p_\theta(s_i | \mathbf{s}_{/\mathcal{M}_s}, *) \right] \\ \mathcal{L}_{\text{seq+struct2seq}} & = & \mathbb{E}_{(\mathbf{s},\mathbf{x})\sim\mathcal{P}(\mathcal{D}_{\text{struct}})} \left[\sum_{i \in \mathcal{T}_s} -\log(p_\theta(s_i | \mathbf{s}_{/\mathcal{M}_s}, \mathbf{x}_{/\mathcal{M}_x}) \right] \\ \mathcal{L}_{\text{struct2seq}} & = & \mathbb{E}_{(\mathbf{s},\mathbf{x})\sim\mathcal{P}(\mathcal{D}_{\text{struct}})} \left[\sum_{i \in \mathcal{T}_s} -\log(p_\theta(s_i | *, \mathbf{x}) \right] \end{cases} \tag{3}$$

with $p_\theta(\mathbf{s}_i | \mathbf{s}, \mathbf{x})$ denoting the output of the softmax layer shown in Figure 1 at position $i \in \{1, \ldots, n\}$, given $(\mathbf{s}, \mathbf{x})$ as input. The masking strategy for each pre-training objective is outlined below, defining the positions of the amino acids to predict $\mathcal{T}_s$, the masked residues in the sequence $\mathcal{M}_s$, and the masked structures $\mathcal{M}_x$:

- **seq2seq.** This task, used in training sequence-only antibody models [Devlin et al., 2018, Olsen et al., 2024], is applied to the sequential dataset $\mathcal{D}_{\text{seq}}$, which lacks structural information (i.e., $\mathbf{x} = *$). For each sequence, between 10% and 40% of the amino acids are selected for masking using one of two methods: (i) randomly sampling individual residues throughout the sequence or (ii) masking continuous spans of residues, with the starting position chosen at random. The positions of the residues to be predicted are the same as those masked, $\mathcal{M}_s = \mathcal{T}_s$. The masked residues in $\mathcal{M}_s$ are then processed using one of three strategies: (a) replaced by the unknown token $*$ with 80% probability, (b) substituted with a different amino acid with 10% probability, or (c) left unchanged with 10% probability. The masking distribution is also slightly adjusted to ensure balanced coverage of both CDR and framework regions.
- **seq+struct2seq.** Both sequential and structural information are used to predict masked amino acids, with masking applied to both the structure and sequence simultaneously. The

same residues are used for both prediction and masking, with $\mathcal{T}_s = \mathcal{M}_x$. Following the seq2seq approach, 10% to 40% of the amino acids are masked, using a mix of continuous spans and random positions. With equal probability, we either (i) mask the corresponding coordinates $\mathcal{M}_x = \mathcal{M}_s$ or (ii) retain the full structural information $\mathcal{M}_x = \emptyset$ to use it as guidance.

- **struct2seq.** Only the structural information from the structural dataset $\mathcal{D}_{\text{struct}}$ is used to predict amino acids $\mathbf{s}_i$ at specific target positions $\mathcal{T}_s$. The input sequence data is completely disregarded, replaced by a series of unknown tokens $*$, leaving only the structural information $\mathbf{x}$. The target positions for amino acid prediction, $\mathcal{T}_s$, are chosen using the same distribution as in the seq2seq task, alternating between continuous spans and random positions.

By using this combination of pre-training objectives, the model dedicates equal time to each task individually.

D.1 Training details

The model was trained on 8 A10G GPUs using a distributed DDP strategy and the PyTorch Zero Redundancy Optimizer [Rajbhandari et al., 2020]. The total number of training steps was predetermined at 125,000. The learning rate was warmed up over the first 200 steps to a peak of 0.001, after which it was gradually reduced to zero using a cosine scheduler. Training was conducted in 16-bit precision. To conserve memory and enable a larger batch size, gradient activation checkpointing was implemented immediately after the structural module. The effective batch size was set to 90 per GPU, resulting in a total batch size of 720 samples per step. The AdamW optimizer was used with a weight decay parameter of 0.1, epsilon of 0.00001, and betas of [0.9, 0.95] for regularization. More details can be found in the Appendix B.

E Experimental results

E.1 CDR editing

Figure 5 collects the result of the CDR recovery experiment in all CDR regions.

E.2 HER2 experiments

Figure 6 displays the results of the HER2 HCDR3 design for the control dataset provided within the same paper.

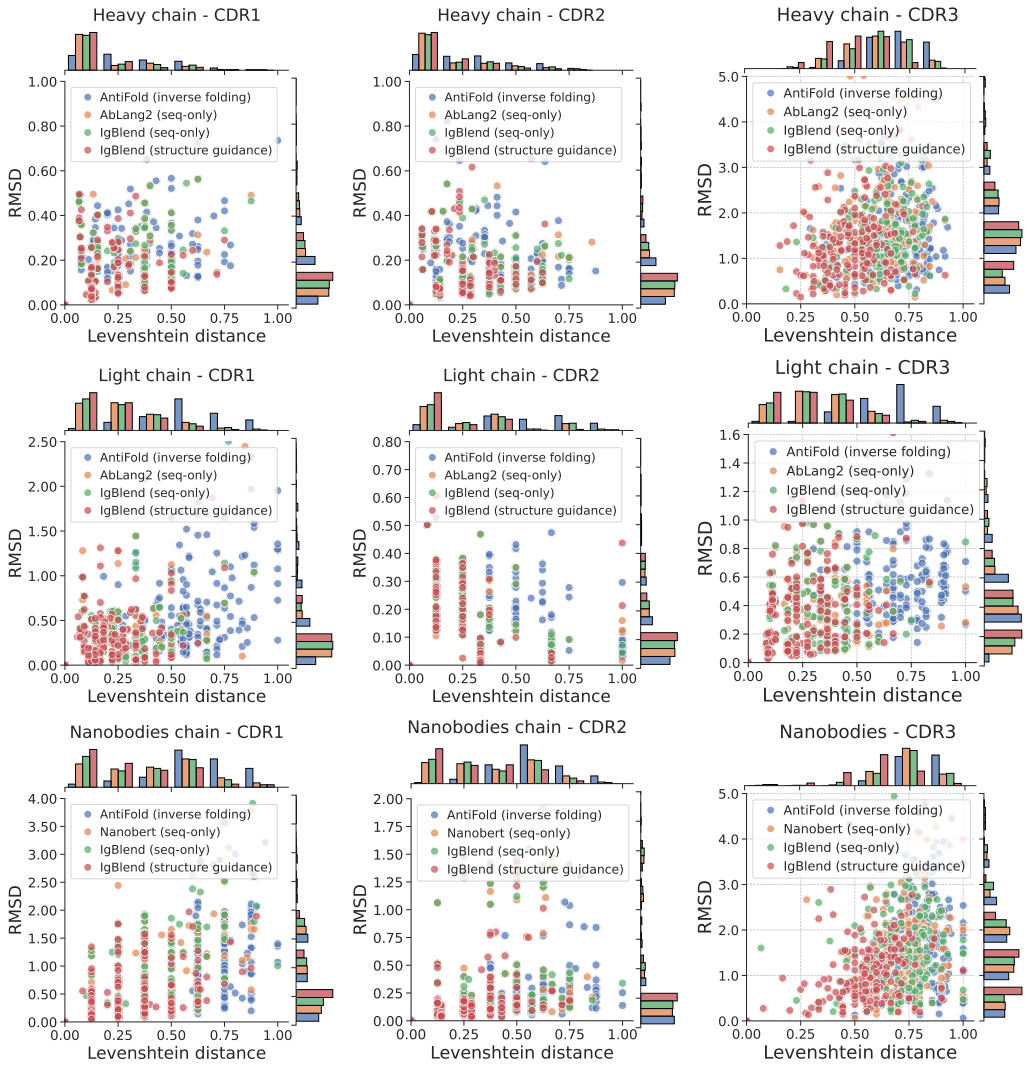

Figure 5: **CDR recovery results:** One series of aminco acid of the sequence is fully masked (one CDR), and the model attemps to recover it. AntiFold only uses the structural information. IgBlend (structure guidance) uses the masked sequence and the structure information. The distances (both Levenshtein and RSME) are only computed in the masked CDR regions. The x-axis displays the Levensthein distance of the generated sequences to the original one and the y-axis reports the RMSE of the generated sequence with regards to the original structure.

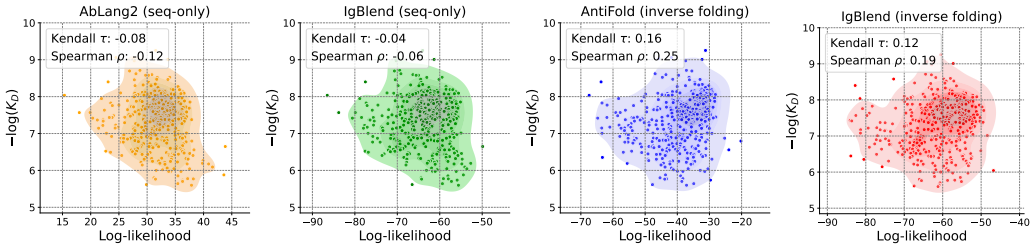

Figure 6: **Trastuzumab-HER2 H-CDR3 editing on control dataset.** Each model scores sequences using their log-likelihood on all CDRS. The scatter plot displays the log-probability (score) of each sequence on the x-axis and the $-\log(K_D)$ values on the y-axis. Additionally, the density of the point cloud is displayed.