Benchmarking Generative Models for Antibody Design

Talip Uçar **Cedric Malherbe** **Ferran Gonzalez**
Centre for AI, BioPharmaceuticals R&D, AstraZeneca
{talip.ucar, cedric.malherbe, ferran.gonzalez}@astrazeneca.com

Abstract

Generative models trained on antibody sequences and structures have shown great potential in advancing machine learning-assisted antibody engineering and drug discovery. Current state-of-the-art models are primarily evaluated using two categories of in silico metrics: sequence-based metrics, such as amino acid recovery (AAR), and structure-based metrics, including root-mean-square deviation (RMSD), predicted alignment error (pAE), and interface predicted template modeling (ipTM). While metrics such as pAE and ipTM have been shown to be useful filters for experimental success, there is no evidence that they are suitable for ranking, particularly for antibody sequence designs. Furthermore, no reliable sequence-based metric for ranking has been established. In this work, using real-world experimental data from seven diverse datasets, we extensively benchmark a range of generative models, including LLM-style, diffusion-based, and graph-based models. We show that log-likelihood scores from these generative models correlate well with experimentally measured binding affinities, suggesting that log-likelihood can serve as a reliable metric for ranking antibody sequence designs. Additionally, we scale up one of the diffusion-based models by training it on a large and diverse synthetic dataset, significantly enhancing its ability to predict and score binding affinities. Our implementation is available at: https://github.com/AstraZeneca/DiffAbXL

1 Introduction

Antibodies are crucial components of the immune system and have become indispensable tools in therapeutics and diagnostics due to their ability to specifically recognize and bind to a wide range of antigens. Engineering antibodies to improve their affinity, specificity, and stability is a rapidly advancing field, increasingly driven by machine learning and computational approaches. Generative models trained on antibody sequences and structures hold great promise in accelerating antibody design and drug discovery. However, current state-of-the-art models typically rely on in silico evaluation metrics, divided into two primary categories: sequence-based metrics, such as amino acid recovery (AAR), and structure-based metrics, such as root-mean-square deviation (RMSD) between predicted and actual structures. Recent advances in structural prediction, notably AlphaFold, have significantly improved our ability to predict protein structures. These tools provide structure-based confidence metrics, such as predicted Local Distance Difference Test (pLDDT), predicted alignment error (pAE), predicted template modeling (pTM), interface predicted template modeling (ipTM), and DockQ scores. Some of these metrics, such as pAE and ipTM, have been demonstrated to be effective filters for distinguishing between high-quality and low-quality structural models, thereby enhancing the chances of experimental success [Abramson et al., 2024, Watson et al., 2023]. While these structure-based metrics are valuable for filtering and assessing model performance, they are not suitable for ranking antibody sequence designs, particularly when it comes to predicting binding affinity and functional efficacy. Moreover, existing sequence-based metrics, such as AAR, provide limited insights into functional performance, as there is no established proxy derived from sequence information alone that accurately predicts binding affinity. This poses a substantial challenge in prioritizing antibody candidates for experimental validation.

Preprint. Under review.

Physics-based approaches provide energy-based metrics by modeling biological systems and accounting for factors such as protein flexibility, explicit solvents, co-factors, and entropic effects. However, the correlation between these metrics and experimentally measured binding affinities is often low [Bennett et al., 2023], and there is no strong evidence that they are effective for ranking antibody designs based on affinity. These methods also face significant challenges, including high computational costs and difficulties in automation [Alford et al., 2017], which limits their utility for large-scale affinity predictions.

In this work, we address these limitations by conducting a rigorous evaluation of state-of-the-art generative models for antibody design, using seven diverse real-world datasets and a range of generative models, including Large Language Model (LLM)-style, diffusion-based, and graph-based models. We demonstrate that log-likelihood scores from these models correlate strongly with experimentally measured binding affinities, positioning log-likelihood as a reliable metric for ranking antibody sequence designs. Furthermore, we scale up one of the existing diffusion-based generative models by training it on a large and diverse synthetic dataset, significantly enhancing its ability to predict and score binding affinities. Our scaled model outperforms existing models in terms of its correlation with experimentally measured affinities. By leveraging experimental validation and addressing the shortcomings of current in silico metrics, our work introduces log-likelihood as a reliable and practical metric for ranking antibody sequence designs. This approach provides a direct link between computational model outputs and experimentally measured binding affinities, offering a clear path for prioritizing high-affinity antibody candidates. Our findings suggest that log-likelihood-based ranking can streamline experimental efforts, ultimately accelerating the discovery and development of next-generation therapeutic antibodies.

Background on Antibodies Human antibodies are classified into five isotypes: IgA, IgD, IgE, IgG, and IgM. This work focuses on IgG antibodies—Y-shaped glycoproteins produced by B-cells and nanobodies, which are single-domain antibody fragments (see Figure 1a for reference). Hereafter, "antibody" refers specifically to IgG antibodies. Antibodies have regions with distinct immune functions. The Fab (fragment antigen-binding) region, comprising variable (V) and constant (C) domains from both heavy and light chains, binds antigens. Within this region, the variable domains (VH and VL) form the antigen-binding site and determine specificity. The Fv (fragment variable) region is the smallest unit capable of antigen binding, consisting only of VH and VL without constant domains. Within variable domains are framework regions and complementarity-determining regions (CDRs). Framework regions maintain structural integrity, while CDRs—three loops on both VH and VL—directly bind antigens and are crucial for specific recognition. The Fv region, essential for antigen recognition, lacks the effector functions of the full antibody. The Fab region, including both variable and constant domains, is more stable and has higher antigen affinity. The Fv region is simpler and easier to engineer for applications such as single-chain variable fragment (scFv) antibodies. The Fc (fragment crystallizable) region at the antibody's base regulates immune responses by interacting with proteins and cell receptors. Nanobodies are compact, single-domain antibodies derived from heavy-chain-only antibodies found in animals such as camels and llamas. Smaller than traditional Fv regions, they retain full antigen-binding capacity and offer increased stability and easier production, making them valuable in therapeutic and diagnostic applications.

2 Related Work

The application of deep learning to antibody and protein design has garnered significant attention in recent years, driven by advancements in natural language processing (NLP) and geometric deep learning. These generative models can be categorized into three broad approaches: LLMs, graph-based methods, and diffusion-based methods. Additionally, they can be distinguished by their input-output modalities: sequence-to-sequence, structure-to-sequence (inverse folding), sequence-structure co-design, and sequence-structure-to-sequence frameworks. Below, we review related work across these categories, focusing on both protein and antibody design.

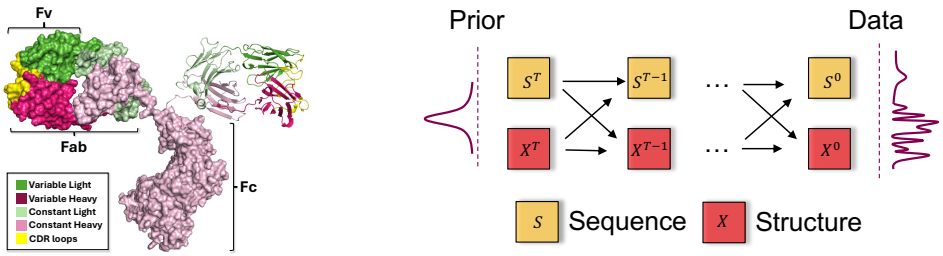
LLM-based approaches LLMs, drawing from advancements in natural language processing (NLP), have been applied extensively to both protein and antibody design. These models can be categorized based on their input-output modalities. In the broader domain of protein design, sequence-to-sequence models such as ESM [Rives et al., 2021] have demonstrated success in tasks such as sequence recovery and mutation effect prediction. These models focus on identifying patterns within protein sequences and have improved our ability to generate functional proteins from sequence data. In many cases, these

models are benchmarked using experimental data from Deep Mutational Scans (DMS), predicting the likelihood of amino acid substitutions and their correlation with measured protein fitness. However, comprehensive benchmarks that assess model predictions of antibody affinity beyond single-amino acid mutations or indels, particularly those incorporating antigen information, are still lacking [Notin et al., 2024]. On the other hand, structure-to-sequence models such as ESM-IF [Hsu et al., 2022] predict amino acid sequences that fold into the same fixed backbone structure, providing a solution to the inverse folding problem. Recent work in protein design has also introduced sequence-structure co-design models, which use both sequence and structural information as input and output. One such model is ESM-3 [Hayes et al., 2024], which incorporates not only sequence and structure but also functional information to improve the design of proteins. This co-design approach allows for the generation of sequences that not only match a desired structure but also fulfill specific functional requirements. Such advancements represent a key shift towards integrating multiple modalities in a single framework for more accurate protein design. In the context of antibodies, several LLM-based models have been developed for specific immunoglobulin-related tasks. The sequence-to-sequence models such as AbLang [Olsen et al., 2022b], AbLang-2 [Olsen et al., 2024], AntiBERTy [Ruffolo et al., 2021], and Sapiens [Prihoda et al., 2022] leverage architectures such as BERT [Devlin et al., 2018] to model antibody sequences and are particularly effective in tasks such as residue restoration and paratope identification. However, these models focus mainly on sequence information and do not incorporate structural data, limiting their ability to design sequences with associated structural properties. Similarly, structure-to-sequence models such as AntiFold [Høie et al., 2023] focus on the inverse folding problem for antibodies, generating sequences that fit a given structural backbone. While these approaches offer valuable insights, they still treat sequence and structure separately. To bridge this gap, recent efforts have introduced models that incorporate both sequence and structure. For example, LM-Design [Zheng et al., 2023] and IgBlend [Anonymous, 2025] represent a new class of sequence-structure-to-sequence models that leverage both modalities at the input to design proteins and antibodies respectively. By learning joint representations of sequence and structure, these models provide a more holistic approach to protein and antibody design, improving the design of sequences that are structurally and functionally coherent.

Graph-based approaches Graph-based methods have become prominent in antibody design due to their ability to represent the geometric structure of antibody regions. These models treat antibody structures as graphs, where nodes correspond to residues or atoms, and edges capture the spatial relationships between them. This allows for the co-design of sequences and structures in a way that respects the underlying geometry of antibodies. For instance, Jin et al. [2021] proposed an iterative method to simultaneously design sequences and structures of CDRs in an autoregressive manner, continuously refining the designed structures. Building on this, Jin et al. [2022] introduced a hierarchical message-passing network that focuses specifically on HCDR3 design, leveraging epitope information to guide the design process. Another approach by Kong et al. [2022] uses SE(3)-equivariant graph networks to incorporate antibody and antigen information, enabling a more comprehensive design of CDRs. These models emphasize sequence-structure co-design, ensuring that generated sequences conform to structural constraints while also optimizing for antigen binding.

Diffusion-based approaches Diffusion-based models have recently emerged as a powerful approach for antibody design. These models generate new sequences and/or structures by simulating a process that progressively refines noisy input into coherent output, holding promise for capturing intricate dependencies in complex biological systems, such as protein folding dynamics and molecular interactions, over multiple iterations [Abramson et al., 2024, Jing et al., 2024]. Moreover, they have proven effective in antibody design due to their ability to handle geometric and structural constraints. Luo et al. [2022] introduced a diffusion model, DiffAb, that integrates residue types, atom coordinates and orientations to generate antigen-specific CDRs, incorporating both sequence and structural information. More recently, Martinkus et al. [2024] proposed AbDiffuser, a diffusion-based model that incorporates domain-specific knowledge and physics-based constraints to generate full-atom antibody structures, including side chains. Another recent approach, AbX [Zhu et al.], is a score-based diffusion model with continuous timesteps, which jointly models the discrete sequence space and the SE(3) structure space for antibody design.

Our contribution The correlation between likelihood and binding affinity is previously observed in works such as [Shanehsazzadeh et al., 2023], where the authors used their inverse folding model IgMPNN to design antibodies targeting HER2. However, the observation was solely based on computing the percentage of binders in sampled antibody library, where higher percentage of binders



(a) Regions of IgG antibody

(b) Reverse diffusion process

Figure 1: **a)** Regions of IgG antibody (PDB ID: 1igt) shown as surface (left and bottom) and cartoon (right). **b)** Iterative reverse diffusion process for DiffAbXL shown only for the position—Gaussian distribution $\mathcal{N}(\cdot) \in \mathbb{R}^3$.

in a smaller library is used as indication to draw the conclusion. In this work, i) We show the direct correlation between log-likelihood and binding affinity and we do so by conducting the same experiment across seven datasets, proving its generaliazibility; ii) We conduct the experiments across different types of generative models, and show their applicability in ranking antibody sequences.; iii) Building on these diffusion-based approaches, we extend one of the existing models, DiffAb [Luo et al., 2022], by training it on a large and diverse synthetic dataset, as well as a small dataset of experimentally determined antibody structures. This scaling significantly enhances the model's ability to predict and rank antibody designs based on binding affinities, addressing one of the key challenges in antibody design: ranking. By incorporating experimental validation, we demonstrate that log-likelihood scores from this scaled-up DiffAb model correlate well with experimentally measured binding affinities, positioning it as a robust tool for antibody sequence design and ranking. Our work contributes to the growing body of research to move beyond simple filtering and towards effective ranking of designs based on experimental success.

3 Method

In this work, to include a scaled version of a diffusion-based generative model, we adapted the diffusion modelling approach, DiffAb, proposed by Luo et al. [2022]. By scaling, we mean increasing the total input sequence length up to 450 residues, as well as expanding the dataset by orders of magnitude, while keeping the model architecture parameters mostly unchanged. Similar to other works in the literature, we trained it on designing CDR3 of the heavy chain (HCDR3) of the antibody as it contributes the most to the diversity and specificity of antibodies [Jin et al., 2022, Xu and Davis, 2000, Zhou et al., 2024]. We refer to this model as DiffAbXL-H3. We also trained another version for designing all six CDRs (DiffAbXL-A).

3.1 DiffAbXL

Data representation We represent the ith amino acid in a given input $\mathcal{V}$ by its type $s_i \in \{\text{ACDEFGHIKLMNPQRSTVWY}\}$, C_α coordinate $\boldsymbol{x}_i \in \mathbb{R}^3$, and orientation $\boldsymbol{O}_i \in \text{SO}(3)$, where $i = 1, 2, ..., N$ and N is the total number of amino acids in $\mathcal{V}$. An input $\mathcal{V}$ consists of one or more masked regions $\mathcal{M}$, which undergo a diffusion process, and the remaining unmasked regions $\mathcal{U}$, which serve as context. Here, $\mathcal{U}$ is the union of the context regions of the antibody and the antigen, where the antigen is optional, such that $\mathcal{V} = \mathcal{M} \cup \mathcal{U}$. If multiple regions are masked, $\mathcal{M}$ refers to the set of masked regions (e.g., all six CDR regions on the heavy and light chains), and $\mathcal{U}$ refers to the remaining unmasked regions. Each masked region $\mathcal{M}_k$ has m_k amino acids at indexes $j_k = l_k + 1, ..., l_k + m_k$, where k indexes the masked regions. The generation task is defined as modeling the conditional distribution $P(\mathcal{M}|\mathcal{U})$, where $\mathcal{M} = \bigcup_k \{(s_{j_k}, \boldsymbol{x}_{j_k}, \boldsymbol{O}_{j_k}) | j_k = l_k + 1, ..., l_k + m_k\}$ is the set of regions to be generated, conditioned on the context $\mathcal{U} = \{(s_i, \boldsymbol{x}_i, \boldsymbol{O}_i) | i \in \{1, ..., N\} \setminus \bigcup_k \{l_k + 1, ..., l_k + m_k\}\}$.

Diffusion Process Training a diffusion probabilistic model consists of two interconnected Markov chains, referred as forward and reversed diffusion, each governing a distinct diffusion process. The forward diffusion process incrementally introduces noise into the data, ultimately approximating the

prior distribution. Conversely, the generative diffusion process initiates from the prior distribution and iteratively refines it to produce the desired data distribution.

Forward diffusion Starting from time $\tau = 0$, noise is incrementally introduced into the data, ultimately approximating the prior distribution at time step $\tau = T$. We use the multinomial $\mathcal{C}(\cdot)$, Gaussian $\mathcal{N}(\cdot) \in \mathbb{R}^3$, and isotropic Gaussian distribution $\mathcal{IG} \in \text{SO}(3)$ to add noise to the type, position, and orientation of amino acids, respectively:

$$q(s_j^t|s_j^0) = \mathcal{C}(\mathbb{1}(s_j^t)|\bar{\alpha}_a^t \cdot \mathbb{1}(s_j^0) + (1-\bar{\alpha}_a^t) \cdot \mathbf{1}/K) \tag{1}$$

$$q(\boldsymbol{x}_j^t|\boldsymbol{x}_j^0) = \mathcal{N}(\boldsymbol{x}_j^t|\sqrt{\bar{\alpha}_p^t} \cdot \boldsymbol{x}_j^0, (1-\bar{\alpha}_p^t) \cdot \boldsymbol{I}) \tag{2}$$

$$q(\boldsymbol{O}_j^t|\boldsymbol{O}_j^0) = \mathcal{IG}_{SO_{(3)}}(\boldsymbol{O}_j^t|\text{ScaleRot}(\sqrt{\bar{\alpha}_o^t} \cdot \boldsymbol{O}_j^0), (1-\bar{\alpha}_o^t)) \tag{3}$$

where $(s_j^0, \boldsymbol{x}_j^0, \boldsymbol{O}_j^0)$ denotes the type, initial position, and orientation of the jth amino acid in one of the masked regions $\mathcal{M}$, while $(s_j^t, \boldsymbol{x}_j^t, \boldsymbol{O}_j^t)$ refers to their values with added noise at time step $\tau = t$. Moreover, $\mathbb{1}$ refers to one-hot encoding of amino acids, $\mathbf{1}$ is a twenty-dimensional vector filled with ones, $\boldsymbol{I}$ is the identity matrix and K is the total number of amino acid types (i.e., 20 in our case). In $\{\bar{\alpha}_a^t, \bar{\alpha}_p^t, \bar{\alpha}_o^t\}$, $\bar{\alpha}^t$ is defined as $\bar{\alpha}^t = \prod_{\tau=1}^{t}(1-\bar{\beta}^\tau)$, where $\bar{\beta}^\tau$ is the noise schedule for type ($\bar{\beta}_a^\tau$), position ($\bar{\beta}_p^\tau$), and orientation ($\bar{\beta}_o^\tau$) of amino acids in each masked region of $\mathcal{M}$ at a given time τ.

Reverse diffusion For the forward diffusion processes above, we define the corresponding reverse diffusion process as follows:

$$p(s_j^{t-1}|\mathcal{M}^t, \mathcal{U}) = \mathcal{C}(\text{F}(\mathcal{M}^t, \mathcal{U})[j]) \tag{4}$$

$$p(\boldsymbol{x}_j^{t-1}|\mathcal{M}^t, \mathcal{U}) = \mathcal{N}(\boldsymbol{x}_j^{t-1}|\boldsymbol{\mu}_p(\mathcal{M}^t, \mathcal{U}), \beta_p^t \cdot \boldsymbol{I}) \tag{5}$$

$$p(\boldsymbol{O}_j^{t-1}|\mathcal{M}^t, \mathcal{U}) = \mathcal{IG}_{SO_{(3)}}(\boldsymbol{O}_j^{t-1}|\text{H}(\mathcal{M}^t, \mathcal{U})[j], \beta_o^t) \tag{6}$$

where $\boldsymbol{\mu}_p(\mathcal{M}^t, \mathcal{U}) = \frac{1}{\sqrt{\alpha_p^t}}(\boldsymbol{x}_j^t - \frac{\beta_p^t}{\sqrt{1-\bar{\alpha}_p^t}}\epsilon_p(\mathcal{M}^t, \mathcal{U})[j])$, and we use $\text{F}(\cdot)[j]$, $\epsilon_p(\cdot)[j]$, and $\text{H}(\cdot)[j]$ to predict the type, the standard Gaussian noise ϵ_j for the position, and the denoised orientation matrix of amino acid j in each masked region $\mathcal{M}$.

Objective function The training objective is defined as the sum of three losses:

$$L_{total} = \mathbb{E}_{t\sim\text{Uniform}(1\ldots T)}[L_a^t + L_p^t + L_o^t], \tag{7}$$

$$L_a^t = \mathbb{E}_{\mathcal{M}^t\sim p}\left[\frac{1}{|\mathcal{M}|}\sum_k \sum_{j=l_k+1}^{l_k+m_k} D_{KL}(q(s_j^{t-1}|s_j^t, s_j^0)||p(s_j^{t-1}|\mathcal{M}^t, \mathcal{U}))\right], \tag{8}$$

$$L_p^t = \mathbb{E}_{\mathcal{M}^t\sim p}\left[\frac{1}{|\mathcal{M}|}\sum_k \sum_{j=l_k+1}^{l_k+m_k} \|\epsilon_j - \epsilon_p(\mathcal{M}^t, \mathcal{U})[j]\|^2\right], \tag{9}$$

$$L_o^t = \mathbb{E}_{\mathcal{M}^t\sim p}\left[\frac{1}{|\mathcal{M}|}\sum_k \sum_{j=l_k+1}^{l_k+m_k} \|(\boldsymbol{O}_j^0)^T\tilde{\boldsymbol{O}}_j^{t-1} - \boldsymbol{I}\|^2\right] \quad \text{and} \quad \tilde{\boldsymbol{O}}_j^{t-1} = \text{H}(\mathcal{M}^t, \mathcal{U})[j], \tag{10}$$

where ϵ_j is a standard Gaussian noise applied to the position $\boldsymbol{x}_j$ of the jth amino acid, and the summations over k account for each masked region $\mathcal{M}_k$ that contains m_k amino acids indexed by $j = l_k+1, ..., l_k+m_k$. The objective functions help the model accurately reconstruct amino acid types, positions, and orientations from noisy data. L_a^t ensures correct amino acid type predictions by comparing true and predicted distributions using KL divergence. L_p^t minimizes the difference between predicted and actual noise in positions, restoring spatial coordinates. L_o^t aligns the predicted and actual orientation matrices by comparing their product with the identity matrix. Together, these

losses train the model to recover the masked regions consistently and accurately. Finally, for an exhaustive explanation of diffusion processes, we refer the reader to the seminal works of [Sohl-Dickstein et al., 2015, Ho et al., 2020], where the main aspects of diffusion models, including their theoretical foundations and practical applications, are covered.

3.2 Training

DiffAbXLs are trained on a combined dataset sourced from SAbDab [Dunbar et al., 2014] and approximately 1.5 million structures generated using ImmuneBuilder2 [Abanades et al., 2023] with paired sequences from the Observed Antibody Space (OAS) [Olsen et al., 2022a]. To ensure high-quality training data, we filtered the structures from the SAbDab dataset following the same procedure as [Luo et al., 2022], removing structures with a resolution worse than 4Å and discarding antibodies that target non-protein antigens. Next, we clustered antibodies from the combined dataset of OAS-paired sequences and SAbDab structures based on their HCDR3 sequences (or LCDR3 if HCDR3 does not exist in the sample), using a 50% sequence identity threshold for clustering. The training and test splits were determined based on cluster-based splitting. The test set included clusters containing the 19 antibody-antigen complexes from the test set used in [Luo et al., 2022] as well as 60 complexes from the RAbD dataset introduced in [Adolf-Bryfogle et al., 2018]. For validation, 20 additional clusters were selected, with the remainder used for training to maximize the training data. Both DiffAbXL-H3 and DiffAbXL-A share the same architecture and hyper-parameters, and are trained for 10 epochs using the AdamW optimizer with an initial learning rate of 1e-4, and a ReduceLROnPlateau scheduler (details provided in Section A.2 of the Appendix). For further details on the model architecture and parameters, please refer to Section A.1 of the Appendix.

3.3 Evaluation

For each sequence in a batch, we compute the log-likelihood of the masked region given the context. In this work, we mask out either all CDRs in antibodies and nanobodies, or only the CDRs where the mutations are applied. When determining the CDR region, we either use the union of several numbering schemes (AHo, IMGT, Chothia, Kabat) to account for variations in CDR definitions, or, if the designed CDRs extend beyond these regions, graft the designed CDRs into the parental sequence and use the grafted region for masking. Let $P_j(s_j \mid \mathcal{U})$ denote the posterior probability of amino acid s_j at position j conditioned on $\mathcal{U}$. To ensure numerical stability, we compute the log probabilities as $\log P'_j(s_j \mid \mathcal{U}) = \log\left(P_j(s_j \mid \mathcal{U}) + \varepsilon\right)$, where ε is a small constant (e.g., 1×10^{-9}). The log-likelihood for the sequence is then calculated by summing over the masked positions:

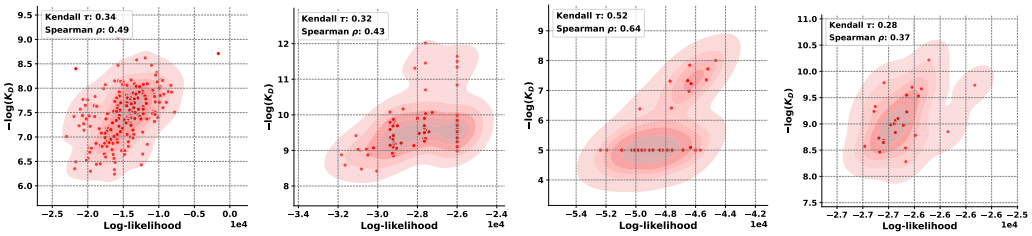
$$\mathrm{LL} = \sum_{j=l+1}^{l+m} \log P'_j\left(s_j \mid \mathcal{U}\right). \tag{11}$$

For consistency, we apply the same log-likelihood computation to BERT-style LLMs. This approach has been previously used to evaluate sequence recovery rates in works such as AntiBERTy [Ruffolo et al., 2021], AbLang [Olsen et al., 2022b], AbLang2 [Olsen et al., 2024], and IgBlend. Specifically, in this case, $P'_j(s_j \mid \mathcal{U})$ corresponds to the output of the final softmax layer at position j, where the prediction is conditioned on the rest of the context $\mathcal{U}$, which is provided as input to the model.

Optionally, if a parent sequence is provided, with amino acids s_j^{parent}, we can adjust the log-likelihood score by subtracting the parent's log-likelihood:

$$\mathrm{LL}_{\text{adjusted}} = \sum_{j=l+1}^{l+m} \left[\log P'_j\left(s_j \mid \mathcal{U}\right) - \log P'_j\left(s_j^{\text{parent}} \mid \mathcal{U}\right)\right]. \tag{12}$$

Unless otherwise specified, we use Equation 11 in our results (see Table 1). After computing the log-likelihoods for all sequences, we assess their relationship with experimental labels y_i (e.g., binding affinities measured in the form of either $-log(K_D)$, $-log(IC50)$, or $-log(qAC50)$) by computing Spearman's rank correlation coefficient ρ and Kendall's tau τ. The pseudocode used for computing correlations and their variance for diffusion models can be found in Section A.3 of the Appendix. Finally, for models such as DiffAbXL that use both sequence and structure at their input,



(a) Absci HER2 zero-shot (b) AZ Target-2 (c) Nature HEL (d) Nature HER2

Figure 2: **Results for DiffAbXL: a)** DiffAbXL-H3-DN for Absci zero-shot HER2 data **b)** DiffAbXL-A-SG for AZ Target-2, **c)** DiffAbXL-A-SG for Nature HEL, **d)** DiffAbXL-A-DN for Nature HER2.

we compute their scores in two modes: i) **De Novo (DN)**[1]**:** We mask both sequence and structure of the region and compute the log-likelihood of the sequence in the masked region at the output; ii) **Structure Guidance (SG):** We mask only the sequence, and use the structure to guide the sampling.

4 Experiments

4.1 Datasets

In this study, we employ seven datasets from three sources—Absci HER2 [Shanehsazzadeh et al., 2023], Nature [Porebski et al., 2024], and AstraZeneca (AZ)—each providing diverse experimental antibody and nanobody data for evaluating the performance of our models.

The Absci HER2 datasets [Shanehsazzadeh et al., 2023] focus on re-designed heavy chain complementarity-determining regions (HCDRs) of the therapeutic antibody Trastuzumab, targeting HER2. The HCDRs were generated using a two-step procedure: i) CDR loop prediction using a machine learning model conditioned on the HER2 antigen backbone structure from PDB:1N8Z (Chain C), Trastuzumab's framework sequences, and the Trastuzumab-HER2 epitope. Then, antibody sequences are sampled using an inverse folding model on predicted structures. HCDR3 lengths ranging from 9 to 17 residues were sampled based on their distribution in the OAS database, while HCDR1 and HCDR2 sequences were fixed at 8 residues—common lengths for these regions. The affinity (K_D) values of these generated sequences were measured using a Fluorescence-activated Cell Sorting (FACS)-based ACE assay. Two datasets are published: (1) the "zero-shot binders" dataset, comprising 422 HCDR3 sequences, from which we utilize those with a HCDR3 length of 13 residues (matching Trastuzumab), and (2) the SPR control dataset, which contains binders and non-binders with varying HCDR regions, where we use only the binders.

The Nature datasets, published by Porebski et al. [2024], provide experimental results for three targets: HER2, HEL, and IL7. For HER2, mutations are present solely in the HCDR3 region, while for IL7, mutations occur in both LCDR1 and LCDR3 regions. In contrast, the HEL dataset consists of nanobodies with mutations across all three CDR regions. These datasets contain 25, 19, and 38 data points for HER2, IL7, and HEL respectively. We use $IC50$ measurement for IL7 and K_D for HER2 and HEL. Additionally, for models that require structural inputs, we predicted the structures using the parental sequences for HER2, IL7 and HEL by using ImmuneBuilder2, IgFold, and NanoBodyBuilder2 respectively [Abanades et al., 2023, Ruffolo et al., 2023] — estimated errors are shown in Table 4 of the Appendix.

The AZ datasets[2] include two distinct antibody libraries designed for two targets. The Target-1 dataset, which is based on rational design, features mutations across four regions (HCDR1-3, LCDR3), and comprises 24 data points. The Target-2 dataset consists of 85 data points and is a combination of three libraries: two rationally designed libraries (one with mutations in three heavy chain CDRs and the other with mutations in three light chain CDRs), and a third library designed using a machine learning model with mutations across all six CDRs. Finally, we use $qAC50$ measurements for

[1]We define the true "De Novo" design as the process of designing an entire antibody from scratch for a specific target sequence. In this work, we use the term mainly to clarify the distinction from the structure-guidance mode.

[2]The AZ datasets and the parental sequence of Nature IL7 are proprietary and will not be disclosed.

Table 1: Summary of the results for Spearman correlation. Abbreviations: DN: De Novo mode, SG: Structure Guidance mode, NA: Epitope or complex structure required, but not available. *, **, *** indicate p-values under 0.05, 0.01 and 1e-4 respectively.

Approach	Model	Dataset						
		Absci HER2		Nature			AZ	
		Zero Shot (K_D)	**Control** (K_D)	**HEL** (K_D)	**IL7** ($IC50$)	**HER2** (K_D)	**Target-1** ($qAC50$)	**Target-2** (K_D)
Graph	**MEAN**	$0.36 \pm 0.00^{***}$	-0.04 ± 0.00	0.25 ± 0.00	$-0.46 \pm 0.00^{*}$	0.02 ± 0.00	-0.37 ± 0.00	0.03 ± 0.00
	dyMEAN	$0.37 \pm 0.00^{***}$	$0.15 \pm 0.00^{**}$	NA	NA	NA	-0.15 ± 0.00	0.03 ± 0.01
LLM	**IgBlend**	$0.27 \pm 0.04^{***}$	0.04 ± 0.02	-0.09 ± 0.08	$-0.84 \pm 0.04^{***}$	-0.10 ± 0.11	0.07 ± 0.09	$0.36 \pm 0.05^{***}$
	AbLang	$0.30 \pm 0.03^{***}$	0.03 ± 0.02	0.17 ± 0.11	$-0.84 \pm 0.04^{***}$	-0.13 ± 0.08	0.09 ± 0.12	$0.35 \pm 0.04^{***}$
	AbLang2	$0.30 \pm 0.02^{***}$	0.02 ± 0.02	0.29 ± 0.04	$-0.83 \pm 0.04^{***}$	-0.07 ± 0.08	0.09 ± 0.09	$0.36 \pm 0.06^{***}$
	AntiBERTy	$0.26 \pm 0.03^{***}$	0.00 ± 0.02	0.07 ± 0.07	$-0.84 \pm 0.03^{***}$	-0.17 ± 0.09	0.09 ± 0.08	$0.35 \pm 0.05^{***}$
	ESM	$0.29 \pm 0.03^{***}$	0.01 ± 0.02	0.25 ± 0.08	-0.18 ± 0.12	0.18 ± 0.12	0.03 ± 0.12	$0.27 \pm 0.06^{**}$
Inverse Folding	**Antifold**	$0.43 \pm 0.03^{***}$	$0.22 \pm 0.01^{***}$	$0.40 \pm 0.07^{**}$	$-0.55 \pm 0.11^{**}$	$-0.47 \pm 0.08^{**}$	-0.27 ± 0.09	$0.38 \pm 0.04^{***}$
	ESM-IF	0.06 ± 0.04	$-0.27 \pm 0.02^{***}$	0.09 ± 0.10	-0.28 ± 0.10	$-0.53 \pm 0.09^{**}$	-0.31 ± 0.12	$0.42 \pm 0.06^{***}$
Diffusion	**AbX**	$0.28 \pm 0.04^{***}$	$0.19 \pm 0.09^{***}$	NA	NA	NA	0.03 ± 0.00	0.08 ± 0.02
	DiffAb	$0.34 \pm 0.01^{***}$	$0.21 \pm 0.01^{***}$	0.21 ± 0.04	-0.24 ± 0.04	-0.14 ± 0.10	-0.07 ± 0.07	$0.22 \pm 0.02^{*}$
	DiffAbXL-H3-DN	$\mathbf{0.49 \pm 0.00^{***}}$	0.05 ± 0.01	$0.52 \pm 0.01^{**}$	$\mathbf{0.23 \pm 0.05}$	-0.08 ± 0.06	-0.22 ± 0.02	$0.37 \pm 0.02^{**}$
	DiffAbXL-H3-SG	$0.48 \pm 0.00^{***}$	0.02 ± 0.00	$0.40 \pm 0.01^{*}$	0.06 ± 0.08	$-0.41 \pm 0.01^{*}$	-0.30 ± 0.04	$0.29 \pm 0.00^{**}$
	DiffAbXL-A-DN	$0.43 \pm 0.00^{***}$	$\mathbf{0.22 \pm 0.00^{***}}$	$0.62 \pm 0.01^{**}$	$-0.79 \pm 0.01^{***}$	$\mathbf{0.37 \pm 0.07^{*}}$	-0.11 ± 0.01	$0.41 \pm 0.00^{**}$
	DiffAbXL-A-SG	$0.46 \pm 0.00^{***}$	$\mathbf{0.22 \pm 0.00^{***}}$	$\mathbf{0.64 \pm 0.01^{***}}$	$-0.80 \pm 0.01^{***}$	-0.38 ± 0.01	-0.02 ± 0.00	$\mathbf{0.43 \pm 0.00^{***}}$

Target-1 and K_D for Target-2. For both targets, we used their crystal structures for models that require structure.

4.2 Models

We utilized a variety of baseline models from the literature, categorized by protein vs. antibody design, modeling approach, and input modality. For protein-based models, we included ESM [Rives et al., 2021], a sequence-only LLM model, and ESM-IF [Hsu et al., 2022], an inverse folding model. For antibody-specific models, we employed several sequence-only LLM models, including Ablang [Olsen et al., 2022b], Ablang2 [Olsen et al., 2024], and AntiBERTy [Ruffolo et al., 2021]. We also included IgBlend, an LLM that integrates both sequence and structural information for antibody and nanobody design. In the graph-based category, we evaluated MEAN [Kong et al., 2022], and dyMEAN [Kong et al., 2023], which use sequence-structure co-design for antibodies. Among diffusion-based models, we included AbX [Zhu et al.], DiffAb [Luo et al., 2022] and our scaled version, DiffAbXL, all of which utilize sequence-structure co-design. Lastly, we assessed Antifold [Høie et al., 2023], an inverse folding model for antibodies.

4.3 Results

We evaluated a broad range of generative models, including LLM-based, diffusion-based, and graph-based models, on seven real-world datasets. The datasets measured binding affinity in terms of K_D, $qAC50$, or $IC50$. Our primary goal was to assess the correlation between the models' log-likelihoods and the experimentally measured binding affinities. The results are summarized in Table 1 for Spearman and Table 3 of the Appendix for Kendall correlations respectively. Across our extensive experiments, several key observations emerged.

First, all generative models trained on antibody and/or protein data exhibited a degree of correlation between their log-likelihoods and binding affinity, although the strength of this correlation varied among models (see Table 1 in the main paper and Table 3 in Appendix). This consistent relationship between likelihood and affinity suggests that these models are capturing relevant aspects of antibody design, even when they were not specifically trained to design the libraries evaluated in this study. This is a crucial finding, as it demonstrates that the models generalize to unseen targets with varying success rates. However, we note that in cases where the target is entirely out of the model's training distribution, the correlation may diminish or disappear.

Second, the models' log-likelihood scores retain predictive power even when synthetic structures are used as input, as demonstrated with the Nature HEL, HER2, and IL7 datasets in Table 1.

Third, for models capable of leveraging epitope information, such as DiffAbXL, we observed only slight variations in correlation when experiments were repeated with and without the antigen as input (see Table 5 in the Appendix). This suggests that including antigen information may not substantially enhance the predictive performance of these models in certain cases. Moreover, structure-based models seem to perform better at ranking than sequence-based models, highlighting the importance of modeling structural information in antibody design.

Fourth, we observed that a model trained specifically to redesign the HCDR3 region of the antibody (i.e., DiffAbXL-H3) is capable of evaluating sequences with mutations outside the HCDR3 region and demonstrates a strong correlation with measured binding affinity.

Fifth, models primarily trained on one type of data (e.g., proteins or antibodies) can effectively evaluate sequences from a different data type (e.g., nanobodies in Nature HEL), showing a strong correlation with experimentally measured binding affinity.

Sixth, we observed strong negative correlations in some experiments involving few targets, particularly when the binding affinity is measured in terms of $IC50$ and $qAC50$ (see Nature HER2 and AZ Target-1 results in Table 1). The reasons for these negative correlations are not fully understood, indicating that the relationship between log-likelihood scores and binding affinity may be more complex in these scenarios and warrants further investigation.

Seventh, it is important to note that success in established in silico metrics does not necessarily translate to better correlation with experimentally measured binding affinities. A notable example is the comparison between AbX and DiffAb, where AbX demonstrates stronger performance across several in silico metrics (see Table-1 in [Zhu et al.]). However, DiffAb exhibits a better correlation with the actual binding affinity measurements in our analysis (see Table-1). This discrepancy suggests that while in silico metrics may capture certain aspects of antibody properties, they do not always align with the true binding affinity, which highlights the challenges in fully replicating biological complexity through computational metrics alone.

Finally, among the evaluated models, the scaled diffusion model, DiffAbXL, consistently outperformed others across most datasets, demonstrating the highest correlation between log-likelihood and binding affinity (see Figure 2 and Table 1). Notably, when comparing the original DiffAb model to its scaled counterpart DiffAbXL, we observed a significant improvement in performance, highlighting the impact of training the diffusion model on a much larger synthetic dataset. This scaling effect underscores the importance of data diversity and volume in enhancing model generalization and accuracy in predicting binding affinity.

5 Conclusion

In this work, we demonstrated that log-likelihood scores from generative models can reliably rank antibody sequence designs based on binding affinity. By benchmarking a diverse set of models—including LLM-based, diffusion-based, and graph-based approaches—across seven real-world datasets, we found consistent correlations between log-likelihood and experimentally measured affinities. The scaled diffusion model, DiffAbXL, particularly stood out by outperforming other models, highlighting the benefits of training on large and diverse datasets. Our findings underscore the potential of generative models not just in designing viable antibody candidates but also in effectively prioritizing them for experimental validation. The ability of structure-based models to outperform sequence-based ones emphasizes the importance of incorporating structural information in antibody design. Areas for further investigation include understanding the negative correlations observed in certain datasets, especially those involving $IC50$ and $qAC50$ measurements. This suggests that the relationship between log-likelihood scores and binding affinity can be complex and may vary depending on the target or measurement method. Future work should explore these nuances to refine predictive models and improve ranking accuracy. Overall, our study provides a practical framework for leveraging generative models in antibody engineering, potentially accelerating the discovery and development of next-generation therapeutic antibodies.

Acknowledgement

We extend our gratitude to everyone in the MLAB program at AstraZeneca, with special thanks to Massimo Sammito, Owen Vickery and Benjamin T. Porebski for their invaluable support. Additionally, we are grateful to Tom Diethe and Rebecca Croasdale-Wood for their unwavering support.

References

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, 2023.

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, pages 1–3, 2024.

Jared Adolf-Bryfogle, Oleks Kalyuzhniy, Michael Kubitz, Brian D Weitzner, Xiaozhen Hu, Yumiko Adachi, William R Schief, and Roland L Dunbrack Jr. Rosettaantibodydesign (rabd): A general framework for computational antibody design. *PLoS computational biology*, 14(4):e1006112, 2018.

Rebecca F Alford, Andrew Leaver-Fay, Jeliazko R Jeliazkov, Matthew J O'Meara, Frank P DiMaio, Hahnbeom Park, Maxim V Shapovalov, P Douglas Renfrew, Vikram K Mulligan, Kalli Kappel, et al. The rosetta all-atom energy function for macromolecular modeling and design. *Journal of chemical theory and computation*, 13(6):3031–3048, 2017.

Anonymous. Igblend: Unifying 3d structures and sequences in antibody language models. In *ICLR*, 2025. Under review.

Nathaniel R Bennett, Brian Coventry, Inna Goreshnik, Buwei Huang, Aza Allen, Dionne Vafeados, Ying Po Peng, Justas Dauparas, Minkyung Baek, Lance Stewart, et al. Improving de novo protein binder design with deep learning. *Nature Communications*, 14(1):2625, 2023.

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

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.

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.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

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.

Wengong Jin, Jeremy Wohlwend, Regina Barzilay, and Tommi Jaakkola. Iterative refinement graph neural network for antibody sequence-structure co-design. *arXiv preprint arXiv:2110.04624*, 2021.

Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Antibody-antigen docking and design via hierarchical structure refinement. In *International Conference on Machine Learning*, pages 10217–10227. PMLR, 2022.

Bowen Jing, Bonnie Berger, and Tommi Jaakkola. Alphafold meets flow matching for generating protein ensembles. *arXiv preprint arXiv:2402.04845*, 2024.

Xiangzhe Kong, Wenbing Huang, and Yang Liu. Conditional antibody design as 3d equivariant graph translation. *arXiv preprint arXiv:2208.06073*, 2022.

Xiangzhe Kong, Wenbing Huang, and Yang Liu. End-to-end full-atom antibody design. *arXiv preprint arXiv:2302.00203*, 2023.

Shitong Luo, Yufeng Su, Xingang Peng, Sheng Wang, Jian Peng, and Jianzhu Ma. Antigen-specific antibody design and optimization with diffusion-based generative models for protein structures. *Advances in Neural Information Processing Systems*, 35:9754–9767, 2022.

Karolis Martinkus, Jan Ludwiczak, WEI-CHING LIANG, Julien Lafrance-Vanasse, Isidro Hotzel, Arvind Rajpal, Yan Wu, Kyunghyun Cho, Richard Bonneau, Vladimir Gligorijevic, et al. Abdiffuser: full-atom generation of in-vitro functioning antibodies. *Advances in Neural Information Processing Systems*, 36, 2024.

Pascal Notin, Aaron Kollasch, Daniel Ritter, Lood Van Niekerk, Steffanie Paul, Han Spinner, Nathan Rollins, Ada Shaw, Rose Orenbuch, Ruben Weitzman, et al. Proteingym: Large-scale benchmarks for protein fitness prediction and design. *Advances in Neural Information Processing Systems*, 36, 2024.

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.

Benjamin T Porebski, Matthew Balmforth, Gareth Browne, Aidan Riley, Kiarash Jamali, Maximillian JLJ Fürst, Mirko Velic, Andrew Buchanan, Ralph Minter, Tristan Vaughan, et al. Rapid discovery of high-affinity antibodies via massively parallel sequencing, ribosome display and affinity screening. *Nature biomedical engineering*, 8(3):214–232, 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.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021. doi: 10.1073/pnas.2016239118. URL `https://www.pnas.org/doi/full/10.1073/pnas.2016239118`. bioRxiv 10.1101/622803.

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.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.

John L Xu and Mark M Davis. Diversity in the cdr3 region of vh is sufficient for most antibody specificities. *Immunity*, 13(1):37–45, 2000.

Zaixiang Zheng, Yifan Deng, Dongyu Xue, Yi Zhou, Fei Ye, and Quanquan Gu. Structure-informed language models are protein designers. In *International Conference on Machine Learning*, pages 42317–42338. PMLR, 2023.

Xiangxin Zhou, Dongyu Xue, Ruizhe Chen, Zaixiang Zheng, Liang Wang, and Quanquan Gu. Antigen-specific antibody design via direct energy-based preference optimization. *arXiv preprint arXiv:2403.16576*, 2024.

Tian Zhu, Milong Ren, and Haicang Zhang. Antibody design using a score-based diffusion model guided by evolutionary, physical and geometric constraints. In *Forty-first International Conference on Machine Learning*.

A Appendix

A.1 DiffAbXL Model Parameters

Table 2: Summary of the DiffAbXL

Module Name	Description		# Layers	Activation Functions	Embedding Dim.	Other Parameters
DiffAbXL	ResidueEmbedding		4	ReLU	128	-
	PairEmbedding		3	ReLU	64	-
	EpsilonNet	ResidueEncoder	2	ReLU	128	-
		ResPairformer	6	ReLU	128	-
		Prediction head	3	ReLU Softmax	128	-
ResPairformer	Attention-based encoder		6	ReLU	128	-
ResPairBlock	Attention & projection layers		-	ReLU	128	num_heads=12 q_dim=32 v_dim=32

A.2 Optimisation

The model was trained using the AdamW optimizer with an initial learning rate of 1×10^{-4}, utilizing 32-bit floating-point precision for numerical computations. A ReduceLROnPlateau learning rate scheduler was employed, configured with a reduction factor of 0.8, patience of 1 epoch, and a minimum learning rate of 1×10^{-5}. Training was conducted over 10 epochs with a batch size of 8, utilizing 8 NVIDIA A100 GPUs.

A.3 Pseudocode for computing correlations and their variance

Algorithm 1: Compute Correlations and Their Variance

Input: Number of samples $N = 10$, Evaluation data loader D, Model M
Initialize empty lists for Spearman and Kendall correlations;
for *seed from* 1 *to* N **do**
  Set random seed using `seed`;
  `// Prepare to collect data for this seed`
  Initialize empty lists for sequences, log probabilities, and labels;
  for *each* $(batch, label)$ *in* D **do**
    Initialize empty list for log probabilities;
    `// Collect log probabilities for 100 samples`
    for i *from* 1 *to* 100 **do**
      `// Compute the log probability for the batch`
      log probability$_i \leftarrow M$.get_log_probs(batch);
      Append log probability$_i$ to the list of log probabilities;
    `// Average the log probabilities`
    Compute average log probability over 100 samples for the batch:;
    avg_log_probability $\leftarrow \frac{1}{100} \sum_{i=1}^{100}$ log probability$_i$;
    `// Get the actual sequences`
    Extract sequence tokens from `batch`;
    `// Store data for later computation`
    Append sequence tokens, average log probabilities, and labels to their respective lists;
  Compute total log-likelihoods using sequences and average log probabilities;
  Compute Spearman correlation between total log-likelihoods and labels;
  Compute Kendall's Tau between total log-likelihoods and labels;
  Append correlations to their respective lists;
`// Aggregate across seeds`
Compute mean and standard deviation of Spearman correlations;
Compute mean and standard deviation of Kendall correlations;

A.4 The results for Kendall correlation

Table 3: Summary of the results for Kendall correlation. Abbreviations: DN: De Novo mode, SG: Structure Guidance mode, NA: Epitope or complex structure required, but not available. *, **, *** indicate p-values under 0.05, 0.01 and 1e-4 respectively.

Approach	Model	Dataset						
		Absci HER2		Nature			AZ	
		Zero Shot (K_D)	Control (K_D)	HEL (K_D)	IL7 ($IC50$)	HER2 (K_D)	Target-1 ($qAC50$)	Target-2 (K_D)
Graph	MEAN	$0.24 \pm 0.00^{***}$	-0.03 ± 0.00	0.18 ± 0.00	$-0.38 \pm 0.00^{*}$	0.01 ± 0.00	-0.26 ± 0.00	0.03 ± 0.00
	dyMEAN	$0.25 \pm 0.00^{***}$	$0.10 \pm 0.00^{**}$	NA	NA	NA	-0.11 ± 0.00	0.02 ± 0.00
LLM	IgBlend	$0.18 \pm 0.02^{***}$	0.02 ± 0.01	-0.07 ± 0.04	$-0.63 \pm 0.04^{***}$	-0.06 ± 0.07	0.06 ± 0.07	$0.27 \pm 0.04^{***}$
	AbLang	$0.20 \pm 0.02^{***}$	0.02 ± 0.01	0.14 ± 0.05	$-0.65 \pm 0.05^{***}$	-0.09 ± 0.06	0.06 ± 0.06	$0.26 \pm 0.04^{***}$
	AbLang2	$0.19 \pm 0.02^{***}$	0.01 ± 0.01	0.22 ± 0.05	$-0.65 \pm 0.06^{***}$	-0.06 ± 0.08	0.06 ± 0.04	$0.26 \pm 0.04^{***}$
	AntiBERTy	$0.17 \pm 0.02^{***}$	-0.00 ± 0.02	0.04 ± 0.08	$-0.65 \pm 0.04^{***}$	-0.14 ± 0.10	0.06 ± 0.08	$0.26 \pm 0.04^{***}$
	ESM	$0.19 \pm 0.03^{***}$	0.01 ± 0.01	0.18 ± 0.05	-0.10 ± 0.07	0.13 ± 0.08	0.01 ± 0.08	$0.20 \pm 0.03^{**}$
Inverse Folding	Antifold	$0.29 \pm 0.02^{***}$	$0.15 \pm 0.01^{***}$	$0.32 \pm 0.06^{**}$	$-0.37 \pm 0.08^{**}$	$-0.32 \pm 0.06^{**}$	-0.19 ± 0.06	$0.28 \pm 0.03^{***}$
	ESM-IF	0.04 ± 0.02	$-0.18 \pm 0.01^{***}$	0.07 ± 0.08	$-0.18 \pm 0.07^{***}$	-0.40 ± 0.07	$-0.24 \pm 0.08^{***}$	$0.30 \pm 0.03^{***}$
Diffusion	AbX	$0.19 \pm 0.04^{***}$	$0.12 \pm 0.06^{***}$	NA	NA	NA	0.01 ± 0.00	0.06 ± 0.01
	DiffAb	$0.23 \pm 0.01^{***}$	$0.14 \pm 0.00^{***}$	0.16 ± 0.03	-0.16 ± 0.03	-0.11 ± 0.07	-0.07 ± 0.04	$0.16 \pm 0.01^{*}$
	DiffAbXL-H3-DN	$\mathbf{0.34 \pm 0.00^{***}}$	0.03 ± 0.01	$0.39 \pm 0.01^{**}$	$\mathbf{0.12 \pm 0.03}$	-0.06 ± 0.04	-0.15 ± 0.02	$0.25 \pm 0.01^{**}$
	DiffAbXL-H3-SG	$\mathbf{0.34 \pm 0.00^{***}}$	0.02 ± 0.00	$0.32 \pm 0.01^{*}$	0.06 ± 0.03	$-0.29 \pm 0.01^{*}$	-0.19 ± 0.03	$0.21 \pm 0.00^{**}$
	DiffAbXL-A-DN	$0.29 \pm 0.00^{***}$	$\mathbf{0.15 \pm 0.00^{***}}$	$0.47 \pm 0.01^{**}$	$-0.60 \pm 0.01^{***}$	$\mathbf{0.28 \pm 0.06^{*}}$	-0.09 ± 0.01	$0.30 \pm 0.00^{**}$
	DiffAbXL-A-SG	$0.31 \pm 0.00^{***}$	$\mathbf{0.15 \pm 0.00^{***}}$	$\mathbf{0.52 \pm 0.01^{***}}$	$-0.60 \pm 0.02^{***}$	-0.27 ± 0.01	-0.02 ± 0.01	$\mathbf{0.32 \pm 0.00^{***}}$

A.5 Details of Predicted Structures

Table 4: Prediction errors for different regions of the parental nanobody used for Nature HEL and two parental antibodies used for HER2 and IL7 respectively.

Region	Prediction Error		
	Nature HEL (Nanobody)	Nature HER2 (Antibody)	Nature IL7 (Antibody)
Framework H-chain	0.87	0.36	0.37
HCDR1	1.61	0.27	0.38
HCDR2	1.57	0.36	0.72
HCDR3	2.82	1.35	2.1
Framework L-chain	-	0.38	0.4
LCDR1	-	0.51	0.92
LCDR2	-	0.28	0.55
LCDR3	-	0.30	0.91

A.6 The Effect of Incorporating Antigen Information

Table 5: Comparison of the results with and without the antigen for Spearman correlations. Abbreviations: DN: De Novo mode, SG: Structure Guidance mode.

Correlation	Model	Antigen	Dataset						
			Absci HER2		Nature			AZ	
			Zero Shot (K_D)	Control (K_D)	HEL (K_D)	IL7 ($IC50$)	HER2 (K_D)	Target-1 ($qAC50$)	Target-2 (K_D)
Spearman	DiffAbXL-H3-DN	Yes	0.49 ± 0.00	0.05 ± 0.01	0.52 ± 0.01	0.23 ± 0.05	-0.08 ± 0.06	-0.22 ± 0.02	0.37 ± 0.02
		No	0.50 ± 0.00	-0.07 ± 0.01	0.52 ± 0.01	0.23 ± 0.05	-0.08 ± 0.06	-0.33 ± 0.05	0.35 ± 0.01
	DiffAbXL-H3-SG	Yes	0.48 ± 0.00	0.02 ± 0.00	0.40 ± 0.01	0.06 ± 0.08	-0.41 ± 0.01	-0.30 ± 0.04	0.29 ± 0.00
		No	0.48 ± 0.00	-0.02 ± 0.01	0.40 ± 0.01	0.06 ± 0.04	-0.41 ± 0.01	-0.45 ± 0.03	0.29 ± 0.01
	DiffAbXL-A-DN	Yes	0.43 ± 0.00	0.22 ± 0.00	0.62 ± 0.01	-0.79 ± 0.01	0.37 ± 0.07	-0.11 ± 0.01	0.41 ± 0.00
		No	0.47 ± 0.00	0.24 ± 0.00	0.62 ± 0.01	-0.80 ± 0.01	0.37 ± 0.07	-0.09 ± 0.00	0.31 ± 0.02
	DiffAbXL-A-SG	Yes	0.46 ± 0.00	0.22 ± 0.00	0.64 ± 0.01	-0.80 ± 0.01	-0.38 ± 0.01	-0.02 ± 0.00	0.43 ± 0.00
		No	0.45 ± 0.00	0.25 ± 0.00	0.64 ± 0.01	-0.80 ± 0.01	-0.38 ± 0.01	-0.09 ± 0.00	0.41 ± 0.00
Kendall	DiffAbXL-H3-DN	Yes	0.34 ± 0.00	0.03 ± 0.01	0.39 ± 0.01	0.12 ± 0.03	-0.06 ± 0.04	-0.15 ± 0.02	0.25 ± 0.01
		No	0.35 ± 0.00	-0.04 ± 0.01	0.39 ± 0.01	0.12 ± 0.03	-0.06 ± 0.04	-0.25 ± 0.03	0.25 ± 0.01
	DiffAbXL-H3-SG	Yes	0.34 ± 0.00	0.02 ± 0.00	0.32 ± 0.01	0.06 ± 0.03	-0.29 ± 0.01	-0.19 ± 0.03	0.21 ± 0.00
		No	0.34 ± 0.00	-0.01 ± 0.00	0.32 ± 0.01	0.06 ± 0.03	-0.29 ± 0.01	-0.28 ± 0.03	0.21 ± 0.01
	DiffAbXL-A-DN	Yes	0.29 ± 0.00	0.15 ± 0.00	0.47 ± 0.01	-0.60 ± 0.01	0.28 ± 0.06	-0.09 ± 0.01	0.30 ± 0.00
		No	0.32 ± 0.00	0.16 ± 0.00	0.47 ± 0.01	-0.60 ± 0.01	0.28 ± 0.06	-0.11 ± 0.00	0.22 ± 0.01
	DiffAbXL-A-SG	Yes	0.31 ± 0.00	0.15 ± 0.00	0.52 ± 0.01	-0.60 ± 0.02	-0.27 ± 0.01	-0.02 ± 0.01	0.32 ± 0.00
		No	0.30 ± 0.00	0.17 ± 0.00	0.52 ± 0.01	-0.60 ± 0.02	-0.27 ± 0.01	-0.10 ± 0.00	0.30 ± 0.00

A.7 Model Inference Details

Below are key considerations and limitations encountered when benchmarking certain models on the binding affinity datasets.

Availability of Antibody-Antigen Complex Information Some models, such as dyMEAN and AbX, require specific input information—dyMEAN needs the epitope location, while AbX requires a bound antibody-antigen structure. Since this information was unavailable in the Nature datasets, these models could not be benchmarked on those datasets.

Multi-CDR Inference Models such as DiffAb and dyMEAN offer checkpoints that support the simultaneous generation of multiple CDRs, which were used for datasets involving modifications to more than one CDR. However, MEAN was trained exclusively on HCDR3 and its inference API only supports generating one HCDR at a time. To allow for fair comparison with other models, we modified MEAN's inference code to enable the simultaneous masking of multiple CDRs during generation. This modification, however, may have affected the model's correlation with binding affinity for datasets involving multiple CDR modifications.