# GENERATIVE HUMANIZATION FOR THERAPEUTIC ANTIBODIES

**Cade Gordon**[*] [†]
University of California, Berkeley

**Aniruddh Raghu**[*]
BigHat Biosciences

**Peyton Greenside**
BigHat Biosciences

**Hunter Elliott**
BigHat Biosciences
helliott@bighatbio.com

## ABSTRACT

Antibody therapies have been employed to address some of today's most challenging diseases, but must meet many criteria during drug development before reaching a patient. *Humanization* is a sequence optimization strategy that addresses one critical risk called immunogenicity — a patient's immune response to the drug — by making an antibody more 'human-like' in the absence of a predictive lab-based test for immunogenicity. However, existing humanization strategies generally yield very few humanized candidates, which may have degraded biophysical properties or decreased drug efficacy. Here, we re-frame humanization as a conditional generative modeling task, where humanizing mutations are sampled from a language model trained on human antibody data. We describe a sampling process that incorporates models of therapeutic attributes, such as antigen binding affinity, to obtain candidate sequences that have both reduced immunogenicity risk *and* maintained or improved therapeutic properties, allowing this algorithm to be readily embedded into an iterative antibody optimization campaign. We demonstrate *in silico* and in lab validation that in real therapeutic programs our generative humanization method produces diverse sets of antibodies that are both (1) highly-human and (2) have favorable therapeutic properties, such as improved binding to target antigens.

## 1 INTRODUCTION

Antibodies are the fastest growing drug class, with approved molecules treating a breadth of disorders ranging from cancer to autoimmune disease to infectious disease (Carter & Lazar, 2018). Many candidate therapeutic antibodies are derived from non-human e.g., murine or camelid sources, and modern antibody formats such as multi-specifics or antibody-drug conjugates can require heavy sequence engineering after discovery. This increases the risk of immunogenicity, where Anti-Drug Antibodies (ADAs) result in either fast clearance of the drug or adverse events (Hwang & Foote, 2005). While antibody sequence humanness is only roughly correlated with immunogenicity, humanization is widely employed to decrease immunogenicity risk (Prihoda et al., 2022). However, even targeted humanizing changes may alter drug efficacy, and any such engineering must be undertaken while maintaining or improving the "drug-likeness" (developability) and intended function of the resulting antibody.

Canonically, humanization is a time-consuming, manual or semi-manual "cut-and-paste" process requiring specialized expertise. Machine learning-based approaches to humanization have been proposed to address this challenge – many of these (Marks et al., 2021; Prihoda et al., 2022; Ramon et al., 2023) use models built on large antibody datasets (Kovaltsuk et al., 2018) to suggest specific mutations which greedily optimize humanness. Published methods employ various mitigating measures aimed at preserving antibody function such as avoiding changes in the Complementarity Determining Regions (CDRs), which are key for target binding, or relying on structural features

[*] Equal contribution

[†] Work performed while at BigHat Biosciences

(Choi et al., 2015; Tennenhouse et al., 2023)(related work in Appendix A). However these methods may still alter antibody function (Chirino et al., 2004; Harding et al., 2010), and generally yield one or a small number of humanized candidates which, if unsuccessful, can prevent the drug from advancing in development.

In this work, we propose re-framing humanization as a conditional generative sequence modeling task, and outline a new humanization algorithm that, given a starting antibody, generates *multiple* humanized candidates enriched for therapeutic properties of interest. When this method is combined with a modern high-throughput wet lab, it allows reliable and efficient humanization that can be performed either in a single step, or incrementally during an iterative optimization campaign: at each round of optimization, we generate many humanized sequences enriched for properties of interest, validate them quickly in the lab, and select top performers for further development in the next round.

From a technical standpoint, our humanization algorithm operates by introducing controlled, humanizing mutations to antibody sequences. These mutations are obtained by *sampling* from a masked language model (MLM) trained on human antibody sequences. Sampling from an MLM allows us to generate many more diverse, human-like sequences than what existing (mostly deterministic) methods produce. We outline a product of experts formulation for sampling that incorporates oracle models trained to predict therapeutic attributes of antibodies, such as binding affinity. This allows us to generate *multiple* humanized candidates that maintain (or even improve) the therapeutic properties of the starting antibody.

In experiments, we first evaluate our algorithms *in silico*, demonstrating we can obtain a large number of highly human candidates with favorable therapeutic attributes. Secondly, we conduct lab validation and demonstrate that in two real therapeutic programs, our method generates humanized antibodies that have improved binding to a target antigen as compared to baselines.

## 2 Methods

### 2.1 Background

**Antibody humanization.** To be a viable therapeutic candidate, an antibody must be "developable" - *e.g.*, it must be synthesizable, thermostable, and of specific relevance here, non-immunogenic (Jarasch et al., 2015). Humanization aims to address this requirement, by taking an engineered antibody and altering its amino acid sequence such that it resembles human antibodies more closely, while preserving its structure, function, and developability.

**Notation.** Denote the set of 20 amino acids by $\mathcal{A} = \{r_0, r_1, \ldots, r_{19}\}$. Let $\mathbf{x} = [x_0, \ldots, x_{L-1}] \in \mathcal{A}^L$ denote an antibody, representing a sequence of $L$ amino acids with each $x_i \in \mathcal{A}$.

Let $f_k : \mathcal{A}^L \mapsto \mathbb{R}$ be a scoring function, or *oracle*, that evaluates an antibody $x$ on some metric of interest. Examples include binding affinity $f_{K_D}$, melting temperature $f_{T_m}$, and humanness $f_h$.

Further, let $m : \mathcal{A}^L \mapsto \mathbb{R}^{L \times 20}$ denote a *masked language model* (MLM) that takes an input antibody $\mathbf{x}$ (potentially with some residues replaced by `<MASK>`) and outputs a $L \times 20$ matrix $\mathbf{Z} = [\mathbf{z}_0, \ldots, \mathbf{z}_{L-1}]^\top$, where each $\mathbf{z}_l$ contains log probabilities over the set of residues in $\mathcal{A}$. The probability distribution over the residues at location $l$ is obtained using the `softmax` function with temperature $\tau$, giving the probability of each residue $r_i$ at location $l$ as:

$$p(r_i) = \texttt{softmax}(\mathbf{z}_l/\tau)[i] = \frac{\exp(\mathbf{z}_l[i]/\tau)}{\sum_{k=1}^{20} \exp(\mathbf{z}_l[k]/\tau)}. \tag{1}$$

### 2.2 Humanization via Sampling

**Concept.** We humanize a starter antibody by mutating it over a series of steps. Each step's mutation is obtained by *sampling* from the output distribution of a masked language model (MLM) conditioned on the current sequence. Importantly, the MLM is trained on a large human antibody dataset (The Observed Antibody Space, OAS (Kovaltsuk et al., 2018)). Samples from such an MLM will tend to introduce more human-like mutations, which can be confirmed with an orthogonal humanness metrics such as the OASis percentile (Prihoda et al., 2022).

Our algorithm proceeds as follows. We first propose a starting antibody sequence $\mathbf{x}^{(0)}$, mutable residue locations $\mathbf{I}$, and a sampling temperature $\tau$. For each step $j$ of the iterative humanization, we pass the sequence $\mathbf{x}^{(j)}$ through the MLM $m$ (potentially after masking), generating a matrix of log probabilities $\mathbf{Z}$. We take $\mathbf{z}_{\mathbf{I}[j]}$, the log probabilities at location $\mathbf{I}[j]$, and sample from the resulting distribution following equation 1. The mutated location $\mathbf{I}[j]$ is infilled with the sampled residue, resulting in $\mathbf{x}^{(j+1)}$. This process then repeats until all indices $\mathbf{I}$ are filled.

**On the choice of masking strategy.** The simplest variant is to not mask at all – *Unmasked Sampling*. If we mask all of $\mathbf{I}$ at the start and progressively infill, we obtain *Autoregressive Denoising Sampling* (similar to Autoregressive Diffusion Models (Hoogeboom et al., 2021)). If at each iteration $j$, we only mask $x_{\mathbf{I}[j]}$, we obtain *Gibbs-like Sampling*. We compare these different strategies in our experiments. Appendix B has full details for all three variants.

**On `argmax` Humanization.** Our algorithm samples from the MLM output distribution at mutable locations, rather than infilling mutable locations with the residue that has the maximal MLM output (i.e., taking an `argmax`). This choice results in many more humanized candidates (Section 3.2). `argmax` infilling (possibly over multiple rounds) without masking gives the Sapiens algorithm (Prihoda et al., 2022), a baseline with which we compare with in experiments. We also define two new variants on Sapiens that do incorporate masking:

- Random Masking Argmax: Mask the mutable input residues, pass the sequence through the MLM, infill all masked locations with the `argmax` over the MLM outputs.
- Iterative Masking Argmax: For each mutable residue, mask it out, pass the sequence through the MLM and infill using the `argmax` operator. Repeat for all mutable residues.

### 2.3 Guided Sampling for Attribute-Aware Iterative Humanization

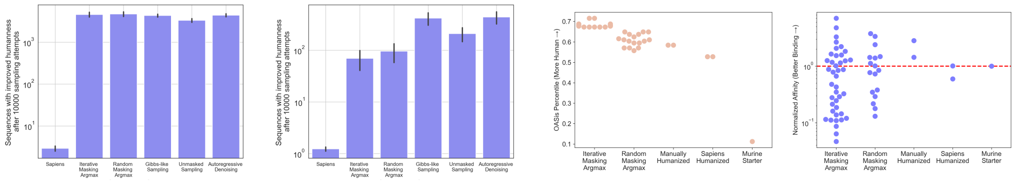
One important drawback with the approach described so far is that the process is completely agnostic to important functional or developability attributes of the starting sequence such as strong binding affinity and thermostability. The humanized candidates can therefore have substantially worse properties along these axes, as we demonstrate in Section 3.2. Furthermore, it assumes that humanization will be undertaken as a single step at either the beginning or end of a campaign, rather than as part of an iterative optimization process.

To address this, we modify the sampling to enrich for attributes of therapeutic importance while iteratively improving humanness. Assume access to a set of oracles $\{f_k\}_{k=1}^{K}$, each of which scores an antibody $\mathbf{x}$ based on some attribute. Then, at each mutable location $l$, instead of sampling from the MLM distribution, we sample from a product of experts (PoE) (Hinton, 2002) distribution. We define $\mathbf{s}_{k,x_l} \in \mathbb{R}^{20}$ to be a vector of scores that oracle $f_k$ produces for all 20 possible mutations of $x_l$. Interpreting these scores as log probabilities, the PoE distribution at this location has the following log probability of residue $r_i$:

$$\log p_{\text{PoE}}(r_i) = \mathbf{z}_l[i]/\tau + \sum_{k=1}^{K} \mathbf{s}_{k,x_l}[i]/\tau_k - \log Z, \tag{2}$$

where $\mathbf{z}_l$ are the MLM's log probabilities at location $l$, $\tau_k$ represents each oracle's temperature, and $Z$ is a normalizing constant. Here, high probability is assigned to a residue $r_i$ that has both: (1) high likelihood under the MLM, representing 'humanness'; and (2) high scores under the oracles, representing other desired therapeutic attributes, weighted by the temperature terms. By sampling from $r' \sim \texttt{softmax}(\mathbf{z}_{\mathbf{I}[j]}/\tau + \sum_{k=1}^{K} \mathbf{s}_{k,x_{\mathbf{I}[j]}}/\tau_k)$ instead of equation 1, we obtain our updated algorithm: *Guided Sampling Humanization*. In our experiments, we demonstrate that sampling in this way generates sequences which progressively improve both humanness *and* desired therapeutic attributes.

**Comparison to PoE in existing work.** Each step of our guided sampling algorithm constructs a *local PoE* distribution over mutations at a single location, which allows for tractable sampling. This implicitly assumes a notion of local functional smoothness in sequence space, which we show is effective in practice. Prior work (Emami et al., 2023; Hie et al., 2022a) has formulated PoE distributions over multiple-mutation trajectories; however, sampling from these distributions necessitates approximations since calculating the partition function exactly is intractable – see Appendix A.



(a) Multiple Murine Starters (b) Multiple Human Starters (c) Humanness assessment (d) Lab validating affinity

Figure 1: **Sampling-based approaches generate many candidates with increased humanness** from both murine (a) and manually-humanized (b) starter antibodies. Allowing up to 2 CDR mutations but measuring resulting affinity in the lab gives improved humanness (c) and a range of affinities (d) including some equal to or better than the starter.

# 3 Experiments

## 3.1 Experimental Setup

**Datasets.** For Masked Language Model (MLM) training, following several earlier works (Ruffolo et al., 2021; Prihoda et al., 2022), we use a dataset of 4 million sequences from Open Antibody Space (Kovaltsuk et al., 2018) (OAS). OAS data processing details are in the appendix. Oracle models were first pre-trained on in-house Next Generation Sequencing (NGS) datasets with between 500k and 1M sequences, and then fine-tuned to 8k binding and thermostability measurements.

**Model and training details.** We train a BERT-style (Devlin et al., 2018) masked language model (MLM) with ∼25 million parameters on the OAS subset described above. We train two oracles – one to predict binding affinity, and one to predict melting temperature. Each of these is an ensemble of 1D CNNs with a ByteNet/CARP architecture (Yang et al., 2022) – details in Appendix C.

**Humanness metrics.** We use an established humanness metric, the OASis percentile (Prihoda et al., 2022), which is well-correlated with immunogenicity risk (with higher humanness indicating lower immunogenicity). When assessing the contribution of experts to guided sampling, we also use the log likelihood of a sequence under the MLM as a continuous humanness proxy, which is correlated with the OASis percentile score (Appendix C).

## 3.2 Sampling Yields Highly Diverse Human Candidates

Our first goal is to understand what impact our sampling-based approaches to humanization have on the number of unique, highly human candidates that can be generated for a given a starting antibody.

First, we take a set of 10 murine (mouse) clinical antibodies (see Appendix C) with low starting humanness (average OASis percentile score of 0.06), and humanize them with different strategies, including a manual (germline grafted) baseline – results are in Figure 1a. In all cases humanizing mutations were limited to the framework regions, with the number of mutations ranging from 22 to 35. Sapiens generates only a few, highly-human candidates; in contrast, the masked `argmax` methods and sampling algorithms all generate orders of magnitude more.

Next, we evaluate a more challenging situation – further humanization of a set of 30 already-humanized clinical antibodies (average OASis percentile score of 0.41) – results are in Figure 1b. On this more difficult task, the sampling-based approaches generate far more highly human candidates as compared to the `argmax` methods, demonstrating the value of sampling-based humanization for challenging starter molecules.

**Lab validating binding affinity.** We lab validate a subset of humanized candidates from a murine starter sequence (details in Appendix C). We synthesize sequences from Random Masking Argmax, Iterative Masking Argmax, Sapiens, HuMAb (Marks et al., 2021) (failed synthesis), and manual humanization, and measure affinities with Bio Layer Interferometry (BLI). Because we have many humanized candidates and will measure affinity in high-throughput, we allow up to 2 humanizing CDR mutations. This gives even more human sequences with a range of affinities including many with equal or better binding (Figure 1c).

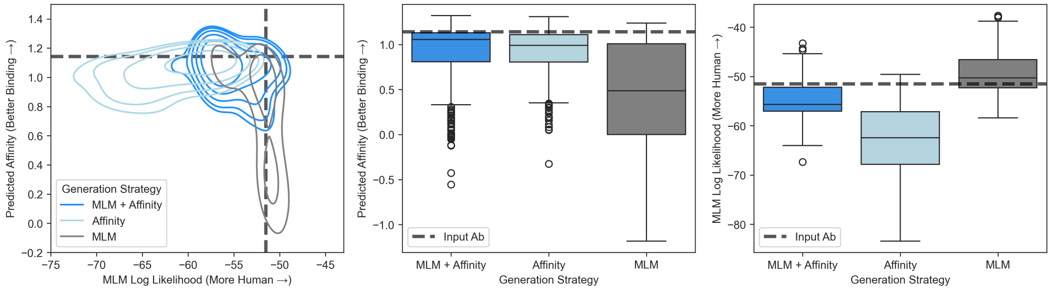
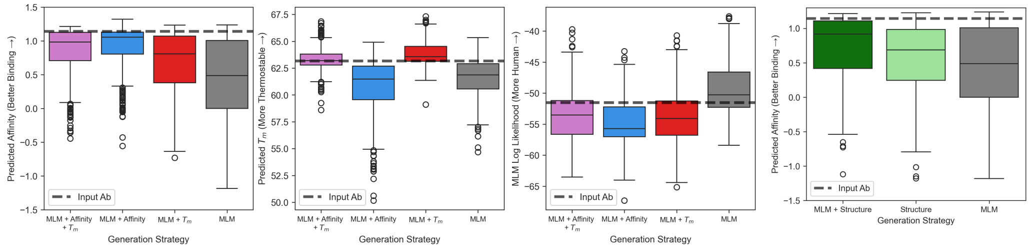


Figure 2: **Guided sampling generates sequences which balance MLM log likelihood ($\sim$ humanness) and predicted target binding affinity** yielding higher-affinity samples compared to unguided sampling (left, contours indicate areas of high sample density, middle and right affinity and log likelihood marginals respectively).



(a) Multioracle Guidance

(b) Predicted Affinity with Structure Guidance

Figure 3: **Guided sampling can flexibly capture multiple desired properties.** Multioracle guidance (a) improves predicted affinity (left), thermostability (middle) and maintains log likelihood (right). Even in the absence of a trained oracle, sampling guided to minimize structural perturbation (b) yields samples which less frequently ablate target binding.

### 3.3 Guided Sampling Allows for Joint Iterative Optimization

We now investigate whether sampling-based humanization can generate candidates with increased humanness that are also enriched for favourable therapeutic properties, such as target binding affinity and thermostability, as part of an iterative antibody optimization campaign. We use the Unmasked Sampling algorithm and use log likelihood as a continuous humanness proxy. For efficiency, we employ a cached oracle approximation, where oracles are first evaluated for all point mutations of the starter sequence and this is used to compute the PoE distribution at each step of sampling. Further details for all experiments are in Appendix C.

**Single-oracle guidance.** We first study guided sampling with a binding affinity oracle, with results in Figure 2. Guided sampling results in sequences that are enriched for being highly human (high log likelihood) and have high predicted binding affinity.

**Multi-oracle guidance.** Extending the above, we generate samples guided by *both* the affinity and thermostability oracles – Figure 3a shows the distribution of samples. Multi-oracle sampling results in a greater density of sequences with high log likelihood, predicted affinity, and predicted thermostability, as compared to unguided sampling or when just using one oracle.

In this scenario we are humanizing while iteratively optimizing *e.g.* affinity over several rounds, so we adjust the number and location of mutations introduced accordingly: First, we introduce only a limited number of mutations (up to 6) at each step, and we allow only a subset of these to be within the CDRs (up to 2).

**Structure-guided sampling.** In situations where we do not have data to train attribute oracles or such oracles are otherwise unavailable, we can use an off-the-shelf antibody structure predictor (e.g., IgFold Ruffolo & Gray (2022)) in guided sampling to generate candidates that have high log likelihood *and* minimize structural deviation from the starting antibody. These samples are enriched for high predicted binding affinity as compared to unguided sampling (Figure 3b), even though an affinity oracle was not used during sampling.

### 3.4 Guided Sampling is Effective in Lab Validation

We now validate our sampling approaches in the lab, confirming they enrich for sequences with desirable therapeutic properties.

**Setup.** We evaluate both Guided and Unguided Unmasked Sampling. For Guided Unmasked Sampling, we generate 500 candidate sequences that have increased humanness and oracle-predicted binding affinity as compared to the starter. For Unguided Unmasked Sampling, we generate 500 candidate sequences that have increased humanness as compared to the starter. We also included 2 sequences humanized by the Sapiens algorithm. Sequences are then filtered for developability liabilities (Appendix C.6).

We study two variants of each method:

1. Unranked, where we sample at random 20 sequences.
2. Ranked, where we rank the sequences with the affinity oracle and then select the top 10.

The final sets of sequences are then synthesized in the lab and characterized for binding affinity.

**Results.** Figure 4 presents the lab validation results – we see that guided sampling methods outperforms unguided sampling, and ranked guided sampling performs best. The Sapiens humanized sequences failed to bind. This demonstrates the value *in vitro* of biasing the sampling process for predicted high affinity variants, rather than only filtering with an oracle after generation.

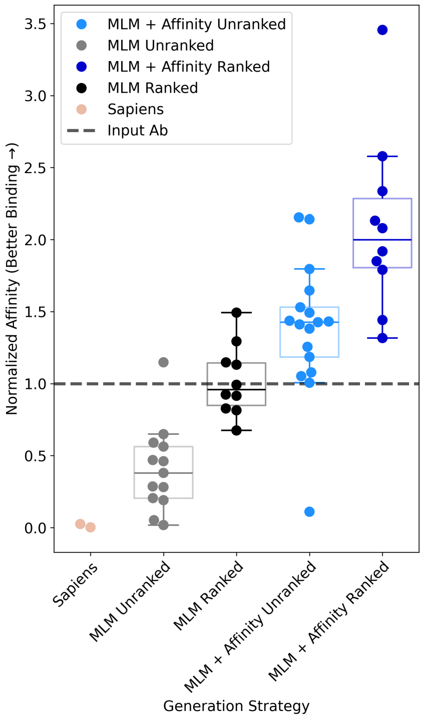


Figure 4: **In lab validation, guided generative humanization yields improved binding affinity.** Affinity guidance outperforms both unguided sampling, as well as unguided samples which are ranked *post-hoc* by an affinity oracle.

## 4 Conclusion

In this paper, we reframed the problem of antibody humanization as one of conditional sampling from a generative model. This represents a departure from prior work that considered humanization as building a functional mapping between an input antibody and a (often single) humanized output antibody. We demonstrate through *in silico* and lab validation that our new sampling based approach can generate many highly human candidate antibodies that are also enriched for properties of therapeutic interest, such as strong binding affinity to a target antigen.

Our method for humanization can be readily embedded into an iterative optimization loop, which could lead to accelerating the development of antibody therapies with reduced immunogenicity risk.

## Acknowledgements

We would like to thank the BigHat Production team, especially Elisabeth Visser, Caleb Ling, AJ Toledo, and Zach Ormsby and the Data Science team, especially Lindsey Osimiri, Joe Davis, and Dan Anderson. We'd also like to thank Ryan Henrici, Jamie Haddon, and the rest of the BigHat ML team for productive discussions and insightful suggestions.

## REFERENCES

Paul J Carter and Greg A Lazar. Next generation antibody drugs: pursuit of the'high-hanging fruit'. *Nature Reviews Drug Discovery*, 17(3):197–223, 2018.

Arthur J Chirino, Marie L Ary, and Shannon A Marshall. Minimizing the immunogenicity of protein therapeutics. *Drug discovery today*, 9(2):82–90, 2004.

Yoonjoo Choi, Casey Hua, Charles L Sentman, Margaret E Ackerman, and Chris Bailey-Kellogg. Antibody humanization by structure-based computational protein design. In *MAbs*, volume 7, pp. 1045–1057. Taylor & Francis, 2015.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. Plug and play language models: A simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164*, 2019.

Anne S De Groot and Leonard Moise. Prediction of immunogenicity for therapeutic proteins: state of the art. *Current Opinion in Drug Discovery and Development*, 10(3):332, 2007.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Patrick Emami, Aidan Perreault, Jeffrey Law, David Biagioni, and Peter St John. Plug & play directed evolution of proteins with gradient-based discrete mcmc. *Machine Learning: Science and Technology*, 4(2):025014, 2023.

Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6): 721–741, 1984.

Ulf Grenander and Michael I Miller. Representations of knowledge in complex systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(4):549–581, 1994.

Nate Gruver, Samuel Stanton, Nathan C Frey, Tim GJ Rudner, Isidro Hotzel, Julien Lafrance-Vanasse, Arvind Rajpal, Kyunghyun Cho, and Andrew Gordon Wilson. Protein design with guided discrete diffusion. *arXiv preprint arXiv:2305.20009*, 2023.

Fiona A Harding, Marcia M Stickler, Jennifer Razo, and Robert DuBridge. The immunogenicity of humanized and fully human antibodies: residual immunogenicity resides in the cdr regions. In *MAbs*, volume 2, pp. 256–265. Taylor & Francis, 2010.

W Keith Hastings. Monte carlo sampling methods using markov chains and their applications. 1970.

Brian Hie, Salvatore Candido, Zeming Lin, Ori Kabeli, Roshan Rao, Nikita Smetanin, Tom Sercu, and Alexander Rives. A high-level programming language for generative protein design. *bioRxiv*, pp. 2022–12, 2022a.

Brian L Hie, Kevin K Yang, and Peter S Kim. Evolutionary velocity with protein language models predicts evolutionary dynamics of diverse proteins. *Cell Systems*, 13(4):274–285, 2022b.

Brian L Hie, Varun R Shanker, Duo Xu, Theodora UJ Bruun, Payton A Weidenbacher, Shaogeng Tang, Wesley Wu, John E Pak, and Peter S Kim. Efficient evolution of human antibodies from general protein language models. *Nature Biotechnology*, 2023.

Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.

Emiel Hoogeboom, Alexey A Gritsenko, Jasmijn Bastings, Ben Poole, Rianne van den Berg, and Tim Salimans. Autoregressive diffusion models. *arXiv preprint arXiv:2110.02037*, 2021.

William Ying Khee Hwang and Jefferson Foote. Immunogenicity of engineered antibodies. *Methods*, 36(1):3–10, 2005.

Alexander Jarasch, Hans Koll, Joerg T Regula, Martin Bader, Apollon Papadimitriou, and Hubert Kettenberger. Developability assessment during the selection of novel therapeutic antibodies. *Journal of pharmaceutical sciences*, 104(6):1885–1898, 2015.

Aleksandr Kovaltsuk, Jinwoo Leem, Sebastian Kelm, James Snowden, Charlotte M Deane, and Konrad Krawczyk. Observed antibody space: a resource for data mining next-generation sequencing of antibody repertoires. *The Journal of Immunology*, 201(8):2502–2509, 2018.

Jinwoo Leem, Laura S Mitchell, James HR Farmery, Justin Barton, and Jacob D Galson. Deciphering the language of antibodies using self-supervised learning. *Patterns*, 3(7), 2022.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

Claire Marks, Alissa M Hummer, Mark Chin, and Charlotte M Deane. Humanization of antibodies using a machine learning approach on large-scale repertoire data. *Bioinformatics*, 37(22):4041–4047, 2021.

Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. *arXiv preprint arXiv:1711.00937*, 2017.

C Poiron, Y Wu, C Ginestoux, F Ehrenmann, P Duroux, and MP Lefranc. Imgt/mab-db: the imgt® database for therapeutic monoclonal antibodies.

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, pp. 2020203. Taylor & Francis, 2022.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Aubin Ramon, Montader Ali, Misha Atkinson, Alessio Saturnino, Kieran Didi, Cristina Visentin, Stefano Ricagno, Xing Xu, Matthew Greenig, and Pietro Sormanni. Abnativ: Vq-vae-based assessment of antibody and nanobody nativeness for hit selection, humanisation, and engineering. *bioRxiv*, pp. 2023–04, 2023.

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.

Jeffrey A Ruffolo and Jeffrey J Gray. Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies. *Biophysical Journal*, 121(3):155a–156a, 2022.

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.

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

Ariel Tennenhouse, Lev Khmelnitsky, Razi Khalaila, Noa Yeshaya, Ashish Noronha, Moshit Lindzen, Emily K Makowski, Ira Zaretsky, Yael Fridmann Sirkis, Yael Galon-Wolfenson, et al. Computational optimization of antibody humanness and stability by systematic energy-based ranking. *Nature biomedical engineering*, pp. 1–15, 2023.

Alex Wang and Kyunghyun Cho. Bert has a mouth, and it must speak: Bert as a markov random field language model. *arXiv preprint arXiv:1902.04094*, 2019.

Kevin K Yang, Alex Xijie Lu, and Nicolo Fusi. Convolutions are competitive with transformers for protein sequence pretraining. In *ICLR2022 Machine Learning for Drug Discovery*, 2022.

Honggang Zou, Rongqing Yuan, Boqiao Lai, Yang Dou, Li Wei, and Jinbo Xu. Antibody humanization via protein language model and neighbor retrieval. *bioRxiv*, pp. 2023–09, 2023.

# A Extended Related Work

Here, we describe additional related work.

**Protein Language Models.** Language models such as BERT or GPT (Devlin et al., 2018; Radford et al., 2019) operate over discrete tokenized spaces making them appropriate modeling tools for proteins (also sequences of discrete tokens, namely amino acid residues). There have been many works developing language models specifically for proteins. ESM trained a series of large-scale transformers that were able to learn biological properties of general proteins in an unsupervised manner (Rives et al., 2021; Lin et al., 2023). In line with this work, others have followed for antibody-specific protein language models such as AntiBERTy, AntiBERTa, and IgLM (Ruffolo et al., 2021; Leem et al., 2022; Shuai et al., 2023). Protein language models have been shown to model the evolutionary trajectory of proteins and mature antibodies without conditioning on the antigen (Hie et al., 2022b; 2023). In this work, we train a protein language model on human antibody data and use it to suggest humanizing mutations to a starter antibody sequence, motivated by the fact that such language models effectively model the human antibody repertoire.

**Classical Humanization.** Early work to reduce the immune risk of an antibody involved combining a non-human antibody's variable region with a fully human antibody's constant region, in a process known as grafting. In an effort to increase the human content of the antibody even more, many classical humanization methods preserve only the CDRs of the non-human molecule, since these regions are important in antigen binding. In such an approach, a scientist will take antibody sequences found in the human genome, known as germline antibodies, and graft the non-human CDRs onto the human framework. These methods are manual, require a lot of time, and often put other features of the antibody at risk (Chirino et al., 2004; Harding et al., 2010). Further, often only a single humanized candidate is produced for a single input antibody sequence.

**Machine Learning-based Humanization.** As antibody datasets grew in size and methods improved, humanization has become an important machine learning task. Marks et al. (2021) train random forest models to classify antibodies as human or nonhuman based on their V gene and uses those models to greedily humanize an antibody using framework mutations. Sapiens trained a masked language model (MLM) on human antibody data, and then used this to suggest humanized sequences by taking the most probable amino acid at each location as determined by the model (Prihoda et al., 2022). PLAN uses a protein language model to embed antibody $k$-mers, takes the closest $k$-Nearest-Neighbors, and finally votes on mutations of the initial $k$-mer using the retrieved $k$-mers (Zou et al., 2023). Most recently, AbNatiV trains VQ-VAEs on human and camelid antibody data (Ramon et al., 2023; Oord et al., 2017). It then chooses liable positions and suggests mutations at those locations using observed frequency seen at the residues in existing human or camelid samples, while retaining only those that satisfy humanness or camelid-ness criterion. Compared to these methods, our sampling-based humanization approach: (1) generates many more humanized candidates for a starter antibody sequence; and (2) generates highly human candidates that are also enriched for desirable therapeutic properties.

**Sampling from Language Models.** Protein language models can provide transition probabilities from one sequence to the next, framing the problem as a Markov random field. If the mutational space has stationary distribution, the problem opens up to Markov Chain Monte Carlo sampling algorithms like Gibbs, Metropolis-Hastings, or the Metropolis-adjusted Langevin algorithm Geman & Geman (1984); Metropolis et al. (1953); Hastings (1970); Grenander & Miller (1994). Notably, Wang & Cho (2019) treated the English MLM Bert as a Markov Random Field to generate text using Gibbs sampling.

When sampling mutations, designers often want multiple different factors to influence sampling. The aforementioned methods only consider a single unsupervised oracle. To overcome this in English language modeling, Plug and Play Language Models (PPLMs) modified the sampling distribution of a model by performing gradient ascent on the final representation of a token using a supervised classifier Dathathri et al. (2019). NOS from Gruver et al. (2023) extended PPLM to diffusion models using both the gradient of supervised classifier and a KL-divergence penalty between the original sequence and the current proposal.

The PPLM line of solutions necessitate a supervised oracle that shares the latent space of the language model of interest, an often untrue assumption in practice and in the context of our experiments. Plug and Play Directed Evolution proposed a general solution to mirror PPLMs, while not requiring gradients Emami et al. (2023). The work proposed a Metropolis Hastings algorithm that uses the gradients of the experts to arrive at proposal distributions. With unsupervised $f_i$s and supervised $g_i$s as experts, they viewed log odds of the distribution as $\pi(x) = \log p(x) = \sum_i f_i(x) + \lambda \sum_j g_j(x) - \log Z$, $Z$ being a normalizing constant. They then take the gradients w.r.t. $x$ to create a forward proposal distribution. After proposing $n$ mutations to $x^{(0)}$ making it $x^{(n)}$ the probabilities of mutating from $x^{(n)}$ to $x^{(0)}$ are calculated. A Metropolis-Hastings criteria accepts the sample with probability equal to the product of $\exp(\pi(x^{(n)}) - \pi(x^{(0)}))$ and the ratio of the product of forward chain probabilities and reverse chain probabilities. As another multi-objective sampler, Hie et al. (2022a) proposed a programming language for proteins by creating sums of energy functions from different oracles then sampling using an algorithm with qualities of both Metropolis-Hastings and Simulated Annealing.

## B Additional Methods Details

**Algorithm descriptions.** Algorithms 1, 2, and 3 present Unguided Unmasked Sampling, Unguided Gibbs-like Sampling and Unguided Autoregressive Denoising Sampling respectively.

**Algorithm 1** Unguided Humanization via Unmasked Sampling

```
**Inputs:** starting antibody $\mathbf{x}$, mutation indices $\mathbf{I}$, sampling temperature $\tau$, and an MLM $m$.
**Output:** a final humanized sequence.

Initialize $\mathbf{x}^{(0)} = \mathbf{x}$ and shuffle $\mathbf{I}$.
**for** $j$ in range(len($\mathbf{I}$)) **do**
    Compute $\mathbf{Z} = m(\mathbf{x}^{(j)})$
    Extract row $\mathbf{I}[j]$ from $\mathbf{Z}$, obtaining log probabilities $\mathbf{z}_{\mathbf{I}[j]}$.
    Sample $r' \sim \texttt{softmax}(\mathbf{z}_{\mathbf{I}[j]}/\tau)$
    Set $x_{\mathbf{I}[j]}$ = $r'$
    Set $\mathbf{x}^{(j+1)} = \mathbf{x}^{(\mathbf{j})}$
**end for**
**Return:** $\mathbf{x}^{(\texttt{len}(\mathbf{I}))}$
```

**Algorithm 2** Unguided Humanization via Gibbs Sampling

```
**Inputs:** starting antibody $\mathbf{x}$, mutation indices $\mathbf{I}$, sampling temperature $\tau$, and an MLM $m$.
**Output:** a final humanized sequence.
Initialize $\mathbf{x}^{(0)} = \mathbf{x}$ and shuffle $\mathbf{I}$.
**for** $j$ in range(len($\mathbf{I}$)) **do**
    Set $x_{\mathbf{I}[j]}$ = <MASK>
    Compute $\mathbf{Z} = m(\mathbf{x}^{(j)})$
    Extract row $\mathbf{I}[j]$ from $\mathbf{Z}$, obtaining log probabilities $\mathbf{z}_{\mathbf{I}[j]}$.
    Sample $r' \sim \texttt{softmax}(\mathbf{z}_{\mathbf{I}[j]}/\tau)$
    Set $x_{\mathbf{I}[j]}$ = $r'$
    Set $\mathbf{x}^{(j+1)} = \mathbf{x}^{(\mathbf{j})}$
**end for**
**Return:** $\mathbf{x}^{(\texttt{len}(\mathbf{I}))}$
```

**Algorithm 3** Unguided Humanization via Autoregressive Denoising Sampling

```
**Inputs:** starting antibody $\mathbf{x}$, mutation indices $\mathbf{I}$, sampling temperature $\tau$, and an MLM $m$.
**Output:** a final humanized sequence.

**for** $i$ in $\mathbf{I}$ **do**
    Set $x_i$ = <MASK>
**end for**
Initialize $\mathbf{x}^{(0)} = \mathbf{x}$ and shuffle $\mathbf{I}$.
**for** $j$ in range(len($\mathbf{I}$)) **do**
    Compute $\mathbf{Z} = m(\mathbf{x}^{(j)})$
    Extract row $\mathbf{I}[j]$ from $\mathbf{Z}$, obtaining log probabilities $\mathbf{z}_{\mathbf{I}[j]}$.
    Sample $r' \sim \texttt{softmax}(\mathbf{z}_{\mathbf{I}[j]}/\tau)$
    Set $x_{\mathbf{I}[j]}$ = $r'$
    Set $\mathbf{x}^{(j+1)} = \mathbf{x}^{(\mathbf{j})}$
**end for**
**Return:** $\mathbf{x}^{(\texttt{len}(\mathbf{I}))}$
```

# C Additional Experimental Details

## C.1 Dataset details

**Open Antibody Space (OAS).** To train masked language models (MLMs), we use approximately 4 million total unpaired heavy and light chain sequences from the Open Antibody Space (Kovaltsuk et al., 2018) (OAS), a large, unlabelled dataset of antibody sequences. This follows several earlier works (Ruffolo et al., 2021; Prihoda et al., 2022). We extracted human studies from OAS up to 2019, and derived a training set of 2 million heavy and 2 million light chain antibody sequences, randomly sampling studies until the desired training set size was obtained.

**Oracle training datasets.** The affinity and thermostability oracles were pre-trained on Next-Generation Sequencing (NGS) data from phage display selections containing 1M and 1.9M sequences respectively. Round-over-round enrichments as log-transformed reads-per-million (RPM) ratios were used as regression targets. The affinity models were fine-tuned to 6k BLI (Octet) $K_D$ measurements. The thermostability models were fine-tuned to 7k nanoDSF (Uncle) measurements of both $T_m$ and $T_{agg}$ as a multi-objective regression task. Fine-tuning data was derived from antibodies produced using Cell-Free Protein Synthesis (CFPS).

## C.2 Model and training details

**Masked language model.** We train a masked language model (MLM) on the OAS subset described above. Our MLM is an 8 layer transformer encoder with 8 attention heads, an embedding dimension of 512, and a feedforward dimension of 2048. It has approximately 25 million trainable parameters in total. This model is trained using a masked language modelling objective, following a similar procedure to Sapiens (Prihoda et al., 2022) and the original BERT model (Devlin et al., 2018). During training, we use the same masking configuration as BERT with 15% of the input amino acid tokens being corrupted overall. Of these, 80% become `<MASK>`, 10% are randomly corrupted to other residues, and 10% remain the same. We use train this model for 100 epochs using a cross entropy loss with label smoothing of 0.1 and the AdamW optimizer with a learning rate of 3e-4 and weight decay of 0.01. We use a cosine decay learning rate scheduler with linear warmup for 1000 steps.

**Oracle models.** We train two oracle model ensembles. The first predicts an input antibody's binding affinity to a target antigen, $K_D$. The second predicts an input antibody's melting and aggregation temperature $T_m$ and $T_{agg}$, which is a proxy for the molecule's thermostability (to be developable an antibody must have high thermostability). Note that in the experiments shown here only the $T_m$ predictions were used. Each ensemble consists of 10 models that are trained on different subsets of the input data. At inference time, we obtain a single point prediction for an input antibody by taking the minimum prediction over the antibody (representing a lower bound on the property of interest). Each model in the ensemble is a 1D CNN with a CARP/ByteNet architecture (Yang et al., 2022), modified for regression by removing causal masking and adding a linear projection layer on the output. The affinity and thermostability models had 48 and 64 residual ByteNet blocks respectively and a model dimension of 64, giving a total of 4.1M and 5.5M parameters.

## C.3 Humanness metrics

**OASis Percentile.** The central goal of our work is to increase the 'humanness' of antibodies, thereby reducing the immunogenicity risk. As current *in vitro* immunogenicity assays are minimall predictive of immunogenicity in human patients, prior work has developed various surrogates to predict immunogenicity, ranging from metrics such as the OASis percentile (Prihoda et al., 2022), to models of the underlying biology to predict anti-drug antibody (ADA) response (De Groot & Moise, 2007). We adopt the OASis Percentile from Prihoda et al. (2022) since it is is well correlated with ADA response on clinical antibody data and has the best AUROC on the human-non human discriminative task amongst open-source humanness metrics. This score is computed by calculating the number of overlapping 9-mer amino acid sequences within an antibody as compared to human 9-mers within the OAS database (Kovaltsuk et al., 2018) at different prevalence levels. This number

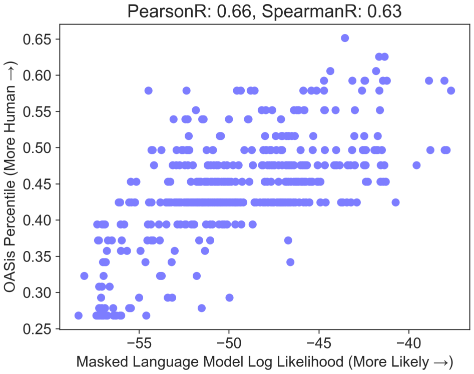


Figure 5: Log-likelihood under the MLM is correlated with the OASis percentile score.

is normalized and then assigned a percentile based on a set of 544 therapeutic antibodies from the IMGT/mAb-DB database (Poiron et al.).

**Masked Language Model Log Likelihood.** The OASis percentile score is a discrete quantity and is only indirectly related to the likelihood of a sequence under an OAS language model. Therefore when combining MLMs with other oracles we use the log likelihood of a sequence under our OAS-trained MLM as a measure of that model's contribution to the sampling, as well as a proxy for the humanness of a given antibody. To check that this is a suitable approximation, we compute the OASis score and MLM log likelihood for 500 test sequences and compare them in Figure 5. We see a correlation between the two metrics.

### C.4 Additional Details on Unguided Sampling Experiments

#### C.4.1 Experimental Setup

**Data.** For the *in silico* evaluation the starter murine and humanized antibodies are sampled from a set of existing clinical antibodies taken from Prihoda et al. (2022). We randomly sample 10 murine starters as a simple test case, and 30 humanized starters as a challenging test case.

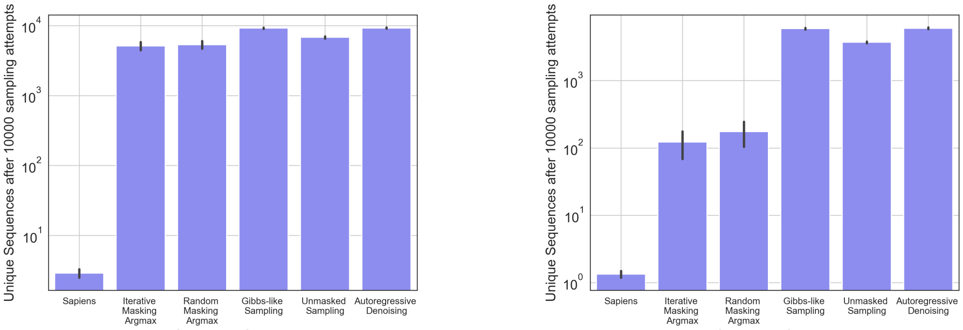
For the lab validation, we synthesized the 30 most human random masking argmax framework-region-only humanized versions of a murine starter sequence. Because we are able to validate binding in high-throughput, we also synthesized 230 iterative random masking argmax humanized sequences where we allowed up to 2 mutations in CDRs in addition to framework humanization. HumAb(Marks et al., 2021) and manually humanized forms of the same starter sequence were also included, although the former failed at the synthesis step. Affinities of all synthesizable variants were measured via BLI. This demonstrates that we are able to produce many variants with high humanness, and which are also synthesizable and bind the target (albeit with variable affinity).

**Sample generation hyperparameters.** We set the softmax temperature for the MLM to be equal to 1 for all sampling methods. We set all non-CDR locations in the input to be mutable indices (that is, all framework region residues) in order to obtain the most diversity possible, without mutating regions of the sequence that might affect binding.

#### C.4.2 Additional results

In the main text, we showed that all methods except Sapiens generated large numbers of unique samples with increased humanness from the highly non-human murine starter antibodies. With the partly-humanized starter antibodies, the sampling methods generated many more highly human samples.

Figure 6 additionally presents the raw number of unique samples generated by the different methods for murine starters and humanized starters, without filtering for increased humanness. We see that



(a) Murine Starter

(b) Human Starter

Figure 6: **Sampling-based approaches generate many unique humanized candidates for both Murine and Humanized starter antibodies.**

the sampling based techniques produce far more unique sequences overall. We observed that for the partly-humanized starters, a large number of generated samples from our the methods have humanness *equal* to that of the starter – this is perhaps expected, since the starter sequences have already passed through a humanization process, and it is difficult to humanize them further.

## C.5 Additional Details on Guided Sampling Experiments

### C.5.1 Experimental Setup

**Starter antibody.** We use an already-optimized nanobody (VHH) sequence from an ongoing therapeutic optimization program as the starter sequence. This represents a typical real use-case for humanization, when an antibody sequence with favourable therapeutic properties must undergo a humanization step to reduce the immunogenicity risk.

**Sample generation hyperparameters.** We set temperature hyperparameters based on visual inspection of the distribution of predicted affinity and log likelihood of samples. When sampling from the MLM with no guidance, we use a temperature of 0.6. When sampling from oracle models (affinity, thermostability, structure) without the MLM, we use a temperature of 0.2. When sampling with guidance (MLM and one or more oracles) we use an MLM temperature of 1.2 and an oracle temperature of 0.4.

In these experiments, we choose at random at most 6 locations in the sequence as mutable, with maximally 2 of these being in the CDRs. This allows us to generate a diverse set of samples while also maintaining minimal changes in the CDRs to prevent changes that might affect properties of therapeutic importance such as antigen binding.

For each of the visualizations, we sample 500 unique antibodies from each method.

**Cached oracle approximation.** At every mutable location, our algorithm constructs the product of experts (PoE) sampling distribution by evaluating the oracle for every possible point mutation (20 total) at that location, keeping the rest of the sequence fixed. This evaluation can be computationally intensive, especially with an ensemble of large oracle models.

As an approximation, following Emami et al. (2023), before running the sampling algorithm, we evaluate the oracle for all possible single mutations of the starter antibody at every location and store this matrix (for a length 100 antibody, this amounts to 2000 oracle evaluations). We then use this pre-computed matrix to obtain the oracle's scores for the different mutations and compute the PoE distribution in equation 2 at each mutable location. This approximation is exact for mutations with hamming distance 1 (i.e., single point mutations) from the starting sequence.

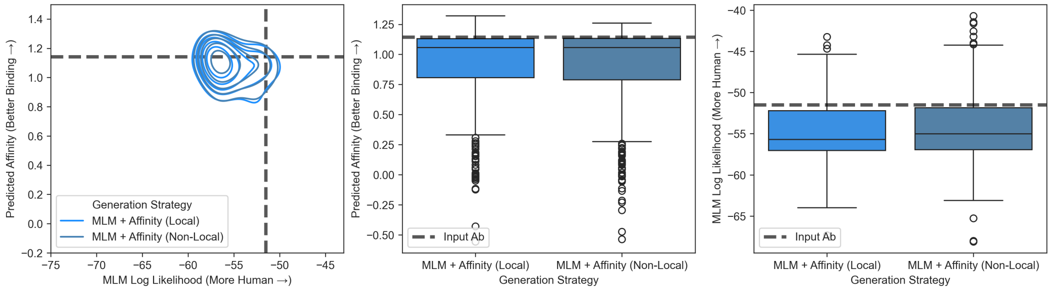


Figure 7: **Using Local PoE results in a similar disribution of generated samples to using Non-Local PoE.**

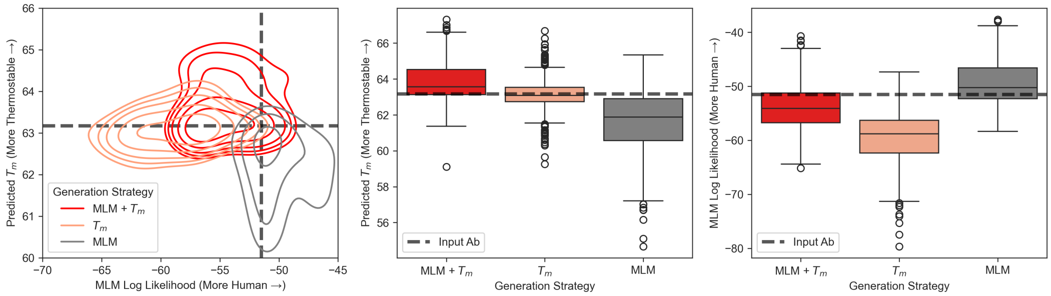


Figure 8: **Guided sampling with a thermostability oracle generates sequences enriched for high melting point.**

Since we consider only 6 total mutations from the starter, we find that this approximation is effective in practice. Figure 7 shows that the distribution of samples with and without this approximation for guided sampling with an affinity oracle, showing that the two distributions are similar.

### C.5.2 Additional results

**Guidance with Thermostability Oracle.** In the main paper, we showed how guided sampling with an affinity oracle generates antibodies enriched for high MLM log likelihood (a proxy for humanness) and high oracle-predicted binding affinity. Figure 8 presents guided sampling with a thermostability oracle, showing we obtain sequences with high MLM likelihood and high oracle-predicted melting temperature, demonstrating again that our algorithm effectively samples sequences enriched for properties of interest.

**Guidance with Structure Prediction Oracle.** As discussed in the main text, there may be situations where attribute prediction oracles are not available for guided sampling. As an example – early in an antibody optimization campaign, there may not be sufficiently large datasets of antibody sequences and lab-measured attributes to train such an oracle model. One option in this case is to guide the sampling using an off-the-shelf antibody structure predictor, such as such as IgFold (Ruffolo & Gray, 2022) – this serves as a proxy for similarity to the starter sequence. We now describe this approach.

Given a candidate antibody $\mathbf{x}$, we define:

$$S(\mathbf{x}) = \begin{bmatrix} x_{0_i} & x_{0_j} & x_{0_k} \\ \vdots & \vdots & \vdots \\ x_{(L-1)_i} & x_{(L-1)_j} & x_{(L-1)_k} \end{bmatrix} \in \mathbb{R}^{L\times 3}$$

to be the antibody's predicted backbone structure. Each row of this matrix represents the 3D coordinates of each alpha carbon atom for each amino acid in the sequence after rigid alignment to the corresponding alpha carbons of the starter antibody $\mathbf{x}^{(0)}$.

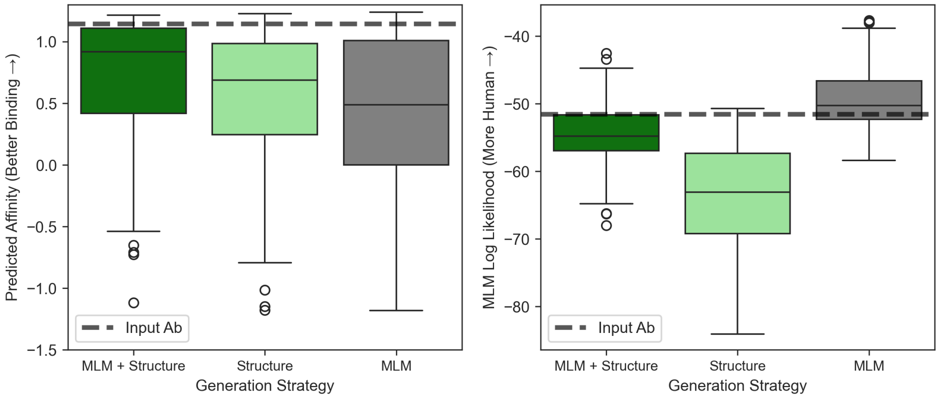


Figure 9: Guided sampling with a structural score function results in fewer samples that have poor binding to a target antigen when compared to sampling without guidance (only using the MLM).

We define the structural score function $f_s$ between the candidate and the starter antibody as follows:

$$f_s(\mathbf{x}, \mathbf{x}^{(0)}) = -\frac{1}{9L^2}||S(\mathbf{x}) - S(\mathbf{x}^{(0)})||_{\text{Fr}},$$

which is the negative of the Froebenius norm of the difference between the two structures, normalized by dimensionality. Intuitively, under this definition, a score of 0 represents the highest similarity to the starter antibody, and a highly negative score represents large structural change from the starter.

We use this score function to guide sampling, which selects for samples that are structurally close to the starter antibody. We then evaluate the predicted affinity of the sampled sequences using our affinity oracle. These results are visualized in Figure 9. We observe that incorporating structure guidance results in fewer samples with low predicted binding affinity, suggesting that this score function can help reduce the number of samples obtained that may bind very poorly to the target antigen.

### C.6 Additional Details on Guided Sampling Lab Validation

#### C.6.1 Experimental Setup

**Data.** We use the same starter antibody as in the previous section.

**Sample generation details.** We use the same sampling hyperparameters and settings as in the previous section.

Humanizing mutations were primarily limited to the framework regions, with only up to 2 humanizing mutations allowed within CDRs (except for Sapiens humanization where CDRs were completely excluded). CDRs were defined as the union of the IMGT, Kabat and Chothia definitions. Sequences were filtered to exclude DDD isomerization and glycosylation motifs, as well as any non-canonical Cysteines.

**Wet-lab validation details.** VHHs were synthesized from the humanized sequences via CFPS. Affinities of the variants which were synthesizable in sufficient quantities were measured via BLI (Octet).