An Energy Based Model for Incorporating Sequence Priors for Target-Specific Antibody Design

Yining Huang[†]
Biomedical Informatics,
Harvard Medical School

Steffanie Paul[† *]
Systems Biology,
Harvard Medical School

Debora S. Marks[*]
Harvard Medical School
Broad Institute

Abstract

With the growing demand for antibody therapeutics, there is a great need for computational methods to accelerate antibody discovery and optimization. Advances in machine learning on graphs have been leveraged to develop generative models of antibody sequence and structure that condition on specific antigen epitopes. However, the data availability for training models on structure ($\sim$5k antibody binding complexes Schneider et al. [2022]) is dwarfed by the amount of antibody sequence data available ($>$ 550M sequences Olsen et al. [2022]) which have been used to train protein language models useful for antibody generation and optimization Here we motivate the combination of well-trained antibody sequence models and graph generative models on target structures to enhance their performance for target-conditioned antibody design. First, we present the results of an investigation into the sitewise design performance of popular target-conditioned design models. We show that target-conditioned models may not be incorporating target information into the generation of middle loop residues of the complementarity determining region of the antibody sequence. Next, we propose an energy-based model framework designed to encourage a model to learn target-specific information by supplementing it with pre-trained marginal-sequence information. We present preliminary results on the development of this model and outline future steps to improve the model framework.

1 Introduction

Antibodies are the fastest-growing class of biologics, creating a great need for technologies to accelerate antibody discovery and optimization. The heavy and light variable chains of an antibody bind to its target (or antigen), and the complementarity-determining regions (CDRs) comprise most of the paratope (the antigen-binding region) [Kunik et al., 2012]. Many antibody engineering approaches focus on diversifying these regions to find novel binders or to increase the affinity of a candidate binding sequence. Machine learning has been deployed to improve various aspects of the antibody development pipeline, including library design [Shin et al., 2021], binder selection from affinity maturation campaigns [Saka et al., 2021, Liu et al., 2020] and antibody optimization [Shan et al., 2022, Hie et al., 2023].

Recently, there has been a growing interest in methods to computationally design *de novo* antibodies that bind to a specified epitope on a target. Many groups have proposed deep graph generative models that take a target structure as input and sample the structure and corresponding sequence of the CDR loops that would bind to the target [Jin et al., 2022, Kong et al., 2022, Luo et al., 2022, Kong et al., 2023]. We term this model class *target-conditioned models* as they aim to learn a distribution of antibody sequences (and structures) conditioned on a binding target. We contrast this with models

[*]Correspondence: steffanpaul@g.harvard.edu, debbie@hms.harvard.edu ; † Equal contribution

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

trained on only antibody sequences, which we term *marginal-sequence models*, as they learn a marginal sequence distribution, unconditional of a target.

Most target-conditioned models suffer from a limited amount of training data as they are trained on the small number of binding complex structures in the PDB ($\sim$ 5k structures in the Structural Antibody Database Schneider et al. [2022]). Datasets of antibody sequence repertoires are significantly larger than structural repertoires (the Observed Antibody Space contains >500M sequences [Olsen et al., 2022]). Groups have developed large protein language models (PLMs) trained on these datasets. For example, Shuai et al. [2022b] developed the Immunoglobulin Language Model (IgLM), a PLM trained on all of OAS that takes in an antibody framework sequence, information about the antibody species origin and heavy or light chain identity, and generates likely CDR sequences. This model was shown to be able to generate expressible and natural-appearing antibody sequences.

Evaluation of target-conditioned models' sequence design performance has been largely limited to benchmarking their Amino Acid Recovery (AAR) and test set perplexity (PP). AAR is the frequency with which sampled sequences match the native amino acid at a particular site in a test sequence. PP is the exponentiated, average negative log-likelihood of a test sequence under the model's learned distribution. Both of these metrics describe how highly a model is prioritizing the known test sequence in its designs when conditioned on the corresponding target.

The CDR3 region of an antibody is considered to contribute most to its binding specificity [Kunik et al., 2012] and models are often compared based on their CDR3 design performance. Evaluation metrics are often reported as the mean across all residue positions in the CDR3. However, middle loop residues are more frequently in contact with the antigen than the flanking residues, and they are much more diverse as they determine the antibody binding specificity. Thus, to evaluate whether a model's designs would bind a target, it is important to compare models based on their performance on these sites.

Here we begin by investigating the sitewise design performance of target-conditioned models and comparing them to marginal-sequence baselines. We find that on middle loop residues of the CDR3, the models have convergent performance, suggesting that the target-conditioned models may not be leveraging target information in their design. This motivated us to develop a model framework that encourages a target-conditioned model to learn target information by supplementing the model with sequence information from a pre-trained marginal-sequence model. We present preliminary findings from benchmarking our initial approach and outline further development goals.

2 Converging performance of target-conditioned models and sequence models

	Average AAR	
Model	All CDR3 sites	Middle loop sites
PSSM	25.66%	10.71%
HERN	**32.73%**	**13.39%**
MEAN	29.17%	10.47%

Table 1: **Summary AAR statistics** The average AAR for the 60 sequences in the RAbD test set. Bolded values are the top scoring model at each sampling temperature and region of the CDR3 being considered. "All CDR3 sites" means the AAR is calculated on the full CDR3 sequence. "Middle loop sites" means the AAR is calculated on only the middle loop sites (excluding the first two and last three sites in CDR3 sequence).

We compared the sitewise CDR3 AAR of two target-conditioned models (HERN [Jin et al., 2022] and MEAN [Kong et al., 2022]) (See Appendix for more details on these models). We used the 60 antibody-complexes in the Rosetta Antibody Design (RAbD) benchmark set [Adolf-Bryfogle et al., 2018] as this is widely used test set for this task. We generated 500 sequences for each target from each model with the sampling temperature set to 1.0 and selected the top 100 sequences by perplexity. We calculated the generated sequences' AAR for each site of the native PDB sequence. We also compared these to Position Specific Scoring Matrices (PSSMs) of the training data for HERN (which is a subset of the training data for MEAN) grouped by CDR3 length. A PSSM is a $L \times D$ matrix, where L is the length of a sequence set and $D = 20$ for the number of amino acid choices. Each

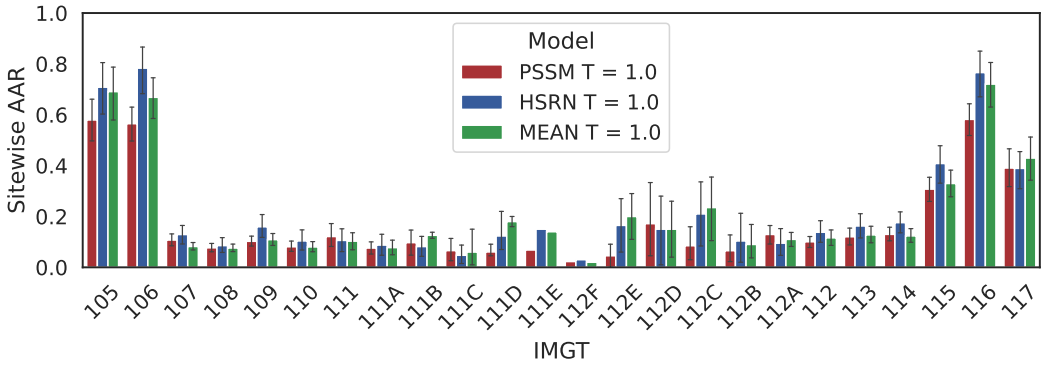

Figure 1: **Sitewise AAR** The mean AAR of each site in the CDR3. Sequences were aligned using ANARCI such that sites from sequences of all lengths could be aggregated. Each site is labelled using IMGT numbering. Error bars are SEM.

entry, $p_{i,a}$ in the PSSM is the frequency of each amino acid a at site i in the sequence. The sitewise AAR of a PSSM is essentially the frequency of the test amino acid at that CDR3 site in the training data. This constitutes a simple sequence model baseline that doesn't use any target information.

We found that, while HERN and MEAN have higher AAR than the PSSM on the flanking residues, they have similar AAR on the middle residues of the CDR3 (Fig 1). When calculating the average AAR without including the flanking residues (IMGT no. 105, 106, 115, 116 and 117 [Lefranc, 1997]), the AAR of all models drops significantly, and MEAN's performance falls under the PSSM baseline (Table 1).

As the middle loop residues more frequently comprise the actual binding interface of the antibody, we would expect that a model that properly leverages target information to design binding sequences would have higher recovery on these residues. The fact that target-conditioned models have convergent recovery compared to a PSSM on antigen-contacting residues suggests that they may only be learning a marginal sequence distribution for the middle loop sites. It would be beneficial to be able to diagnose when a target-specific model is able to leverage target information, and when it is only learning marginal sequence information. Identifying this can motivate the development of the target-conditioned model architecture and training regime such that the model properly learns a conditional sequence distribution.

3 An EBM for combining sequence priors with target conditioned energies

We propose an energy-based model (EBM) [LeCun et al., 2006] framework for target-conditioned antibody sequence design that explicitly models the contribution of marginal sequence information separately from target-specific information.

We define an antibody sequence as X, consisting of residues $(X_1, X_2, \cdots, X_L)$, and the structure of the target epitope as Y. We seek to learn a conditional generative model $P(X|Y) = \prod_{i=1}^{L} P(X_i|Y, X_{<i})$. We model the distribution of each residue as an EBM, where the total energy of a residue is the sum of energy contributions from a marginal-sequence term $E_X(X_i)$ and target-conditioned energy term $E_Y(Y, X_i)$:

$$
\begin{aligned}
P(X_i|Y, X_{<i}) &= \frac{1}{Z_\theta} \exp[-\alpha_\theta E_Y(Y, X_i) - (1 - \alpha_\theta) E_X(X_i)] \\
E_Y(Y, X_i) &= F_\theta(Y) \\
E_X(X_i) &= \log(\pi(X_i))
\end{aligned}
$$

Where Z_θ is the sum of exponentiated negative energies for the different amino acid choices (i.e the denominator of a softmax operation), α_θ is a learned parameter that controls how much each energy term contributes to the final probability, F_θ is a model operating on the target structure with trained parameters θ, and $\pi(X_i)$ is a pre-trained marginal sequence model with amortized parameters.

We hypothesize that by explicitly providing the model with pre-trained sequence information, $F_\theta(Y)$ is encouraged to leverage target information in its energy prediction. Furthermore, the combination parameter α_θ allows us to diagnose how much target information the model is using. If α_θ is low, this suggests the model is just defaulting to a marginal sequence distribution.

Our framework is flexible to using any marginal sequence model and target model architecture. For our initial experiments, we used length-matched PSSMs fit on the same sequences used to train F_θ as the marginal-sequence model. However, the true value of our model comes from leveraging sequence models trained on larger sequences corpuses. This will be explored in future work.

We used the architecture of HERN to model F_θ and we set α_θ to be a single learned variable that tells us globally how much the model is using either marginal-sequence or target-conditioned information. To begin with, we trained models to generate only the CDR3 sequence, however, our framework can be extended to sample all 3 CDR regions.

4 Experiments

	Median Perplexity	
Model	**All CDR3 sites**	**Middle loop sites**
HERN	8.05	↑14.04
PSSM	10.81	↑18.24
IgLM	17.74	↓15.80
HERN (amortized) + PSSM	8.88	↑15.04
HERN (amortized) + IgLM	8.88	↑12.94
EBM (PSSM)	9.67	↑16.07

Table 2: **Perplexity** We report the median perplexity of each model across all test set sequences (lower is better). Perplexity is calculated across all of the CDR3 residues or just the middle loop residues. Here we show the median of the results from 10 cross-validation folds.

Data We benchmarked our model on binding complex structures in the Structural Antibody Database (SAbDab) [Schneider et al., 2022]. After filtering by CDR sequence redundancy, we had 3150 binding complexes. We clustered the sequences according to CDR sequence similarity using mmseqs2, and divided the clusters into training (80%), validation (10%) and test sets (10%). We employed 10-fold cross-validation.

Baselines In addition to our model, we evaluated HERN re-trained on the training splits and the pre-fit PSSMs by themselves. We also benchmarked IgLM to test how well a marginal sequence distribution learned from a larger sequence repertoire can approximate the target-conditional distribution. We also explored using an amortized version of HERN within our EBM framework (i.e we trained HERN on the training split without including any marginal-sequence information, and then combined the target-conditioned and marginal-sequence contributions without any further training). This baseline tests whether providing the target-conditioned model with marginal sequence information during training encourages it to learn more target information.

Metrics We used test perplexity (PP) on the test set sequences as our evaluation metric. Both AAR and PP test whether a model can identify the ground truth sequence from a test binding complex, however, a model's AAR can be augmented depending on how sequences are sampled (See Appendix for further discussion). As a preliminary result, we only present PP here. Future work will investigate our model's AAR as well.

Results We calculated each model's PP on the entire CDR3 and only the CDR3 middle loop residues of the test data (Table 2). We found that our EBM approach had lower perplexity than the marginal-sequence model baselines (PSSM and IgLM), suggesting that the model was able to some learn target-conditioned information. However, HERN by itself had better perplexity on both the full CDR3 and just the middle sites. This suggests that, under our framework and on this task, the marginal-sequence models were confounding the target-conditioned module of the EBM. Looking at the per-site perplexities of all the residues in one fold of the test data, we found that there were

many sites for which the EBM had lower test perplexity than HERN, and these are enriched for sites in which the pre-fit PSSM had high probability (Fig A2). This suggests that there are gains to be made from incorporating information from a marginal-sequence distribution if the EBM can correctly distinguish when the sequence model is reliable.

Remarkably, simply combining the contributions from an amortized HERN model with a marginal-sequence model outperformed our EBM. This suggests that training a target-conditioned model with external pre-trained sequence energies confounds the trained target model. It may be beneficial to start with a pre-trained target-conditioned model, and then fine-tune the model with energy contributions from a marginal-sequence model.

Interestingly, out of all the models tested, IgLM was the only model that had lower perplexity on the middle loop sites with respect to the full CDR3. Notably, the best-performing model on the middle loop sites (which are highly important for target binding) was HERN (amortized) + IgLM. This suggests that, on these sites, including information from IgLM improves the model's designs. This is encouraging and adds some corroboration to our hypothesis that combining marginal-sequence information from larger sequence repertoires could be beneficial for designing binders.

5 Discussion and future work

These preliminary results show that there is great room for improvement with our approach. As it seems that a linear combination of the energy contributions only serves to confound the model, we may have greater success combining the target-conditioned and marginal-sequence information earlier in the model's architecture (eg. with hidden layers that take in both representations of the target structure and representations from a marginal-sequence model). Recently, Wu and Li [2023] developed a hierarchical paradigm for training a model on protein universe datasets, large antibody sequence datasets and structural binding complexes, demonstrating great gains for target-conditioned CDR generation. The ideas motivating their approach are similar to ours, however, this model is fine-tuned end-to-end and information from different data sources is incorporated as features directly in the model's graph representation. In future work, we will benchmark their model and explore incorporating ideas from their approach into ours.

The fact that our EBM underperforms with respect to HERN suggests that the model cannot self-correct when the marginal-sequence contribution is unreliable. It would be favorable for the model to be able to calculate $\alpha_\theta(E_X, E_Y)$ as a function of the energy contributions, so that it can upweight a particular energy on the fly. For example, if the energy from a marginal-sequence model is too high owing to distribution mismatches in the training set of the pre-trained model, the EBM can prioritize the target-conditioned model. We will explore this more in future work.

Lastly, while our approach has low performance under the PP metric, incorporating sequence information may have gains under other evaluation metrics that assay other antibody design goals (eg. designability, expressibility). Metrics such as AlphaFold2 plDDT [Jumper et al., 2021] and ProteinMPNN log probabilities [Dauparas et al., 2022] have been used to rank the designability of computationally generated sequences [Alamdari et al., 2023, Johnson et al., 2023], and scores from large protein language models have been found to correlate well with protein expression [Notin et al., 2022]. We will explore the performance of our model under these metrics in future work.

In conclusion, this work motivates combining sequence and target structure information by investigating the convergent performance of popular target-conditioned generative models with marginal-sequence models. We propose an energy-based model formulation of a model to combine these data sources and present preliminary results on benchmarking our model. While our initial model does not outperform the baselines, these results give us many directions for future work.

References

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 Comput. Biol.*, 14(4):e1006112, April 2018.

Sarah Alamdari, Nitya Thakkar, Rianne van den Berg, Alex Xijie Lu, Nicolo Fusi, Ava Pardis Amini, and Kevin K Yang. Protein generation with evolutionary diffusion: sequence is all you need. September 2023.

J Dauparas, I Anishchenko, N Bennett, H Bai, R J Ragotte, L F Milles, B I M Wicky, A Courbet, R J de Haas, N Bethel, P J Y Leung, T F Huddy, S Pellock, D Tischer, F Chan, B Koepnick, H Nguyen, A Kang, B Sankaran, A K Bera, N P King, and D Baker. Robust deep learning-based protein sequence design using ProteinMPNN. *Science*, 378(6615):49–56, October 2022.

Brian L Hie, Varun R Shanker, Duo Xu, Theodora U J 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. *Nat. Biotechnol.*, April 2023.

Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Antibody-Antigen docking and design via hierarchical equivariant refinement. July 2022.

Sean R Johnson, Xiaozhi Fu, Sandra Viknander, Clara Goldin, Sarah Monaco, Aleksej Zelezniak, and Kevin K Yang. Computational scoring and experimental evaluation of enzymes generated by neural networks. April 2023.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andrew J Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, July 2021.

Xiangzhe Kong, Wenbing Huang, and Yang Liu. Conditional antibody design as 3D equivariant graph translation. August 2022.

Xiangzhe Kong, Wenbing Huang, and Yang Liu. End-to-End Full-Atom antibody design. February 2023.

Vered Kunik, Bjoern Peters, and Yanay Ofran. Structural consensus among antibodies defines the antigen binding site. *PLoS Comput. Biol.*, 8(2):e1002388, February 2012.

Yann LeCun, Sumit Chopra, Raia Hadsell, Aurelio Ranzato, and Fu Jie Huang. A tutorial on energy-based learning. 2006. URL `https://api.semanticscholar.org/CorpusID:8531544`.

M P Lefranc. Unique database numbering system for immunogenetic analysis. *Immunol. Today*, 18(11):509, November 1997.

Ge Liu, Haoyang Zeng, Jonas Mueller, Brandon Carter, Ziheng Wang, Jonas Schilz, Geraldine Horny, Michael E Birnbaum, Stefan Ewert, and David K Gifford. Antibody complementarity determining region design using high-capacity machine learning. *Bioinformatics*, 36(7):2126–2133, 2020.

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. October 2022.

Pascal Notin, Mafalda Dias, Jonathan Frazer, Javier Marchena-Hurtado, Aidan Gomez, Debora S Marks, and Yarin Gal. Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval. May 2022.

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 Sci.*, 31(1):141–146, January 2022.

Koichiro Saka, Taro Kakuzaki, Shoichi Metsugi, Daiki Kashiwagi, Kenji Yoshida, Manabu Wada, Hiroyuki Tsunoda, and Reiji Teramoto. Antibody design using LSTM based deep generative model from phage display library for affinity maturation. *Sci. Rep.*, 11(1):5852, March 2021.

Constantin Schneider, Matthew I J Raybould, and Charlotte M Deane. SAbDab in the age of biotherapeutics: updates including SAbDab-nano, the nanobody structure tracker. *Nucleic Acids Res.*, 50(D1):D1368–D1372, January 2022.

Sisi Shan, Shitong Luo, Ziqing Yang, Junxian Hong, Yufeng Su, Fan Ding, Lili Fu, Chenyu Li, Peng Chen, Jianzhu Ma, Xuanling Shi, Qi Zhang, Bonnie Berger, Linqi Zhang, and Jian Peng. Deep learning guided optimization of human antibody against SARS-CoV-2 variants with broad neutralization. *Proc. Natl. Acad. Sci. U. S. A.*, 119(11):e2122954119, March 2022.

Jung-Eun Shin, Adam J Riesselman, Aaron W Kollasch, Conor McMahon, Elana Simon, Chris Sander, Aashish Manglik, Andrew C Kruse, and Debora S Marks. Protein design and variant prediction using autoregressive generative models. *Nat. Commun.*, 12(1):2403, April 2021.

Richard W Shuai, Jeffrey A Ruffolo, and Jeffrey J Gray. Generative language modeling for antibody design. December 2022a.

Richard W Shuai, Jeffrey A Ruffolo, and Jeffrey J Gray. Generative language modeling for antibody design. December 2022b.

Fang Wu and Stan Z Li. A hierarchical training paradigm for antibody structure-sequence co-design. November 2023.

A Appendix

A.1 Further model explication

Here we provide further detail on the models used in this paper.

Target-conditioned models De novo antibody design models focus on designing the antibody given the target antigen. Hierarchical Equivariant Refinement Network (HERN) generates CDR3 (paratope) sequence autoregressively given the antigen residues that closely interact with antibody (epitope) Jin et al. [2022]. At each autoregressive step, HERN uses a message-passing network to predict the docking complex of the generated CDR3 sequence and epitope. The docking complex is then used to predict the next CDR3 residues in the sequence.

Multi-channel Equivariant Attention Network (MEAN) employs an attention mechanism that generates the whole sequence and structure together Kong et al. [2022]. The full-shot scheme to generate the full CDR sequence at each step is less prone to error accumulation compared to autoregressive models. Generating 1D sequence and 3D structure of CDR together also maintains the full geometric relationship within residues.

However, the above models only focus on designing CDR sequences, they both ignore the framework region and side-chains outside CDR3 sequences. The dynamic Multi-channel Equivariant grAph Network (dyMEAN) incorporates framework sequence into the model to better capture the distribution of the whole antibody Kong et al. [2023].

Sequence models The position-specific scoring matrix (PSSM) represents the frequency of amino acids at each CDR3 site. We fit PSSM of different CDR3 lengths on the training data. We use it here as a baseline CDR3 sequence model.

The Immunoglobulin Language Model (IgLM) is a deep generative language model for designing antibody sequences Shuai et al. [2022a]. It formulates antibody sequence designing tasks as text infilling and generates amino acids autoregressively. IgLM can generate both full-length antibody sequences and infilled CDR loops, conditioned on chain-type (light or heavy) and species (human, mouse, rat, rabbit, rhesus, and camel).

A.2 Limitations of the AAR metric

Both authors of MEAN and HERN use AAR as the model evaluation metric. They evaluate their models by generating a large sample of sequences and selecting the top 100 by perplexity; this is equivalent to sampling from a model with lower temperature as you are selecting for sequences around the mode of the model's distribution. We looked at the AAR across the full CDR3 and for just the middle loop sites for model generations sampled at different temperatures (Table A1, Fig A1). We find that AAR increases for all the models at lower temperatures. Notably, at T=0.0, the PSSM outperforms both target-specific models. This demonstrates how the AAR metric is influenced by any differences in the sampling procedure between models which could confound any comparisons. Given the confounding that sampling schemes can have on model evaluation, we opted to use test perplexity (PP) to evaluate the design methods for our intial experiments.

	Average AAR			
	All CDR3 sites		**Middle loop sites**	
Model	T=1.0	T=0.0	T=1.0	T=0.0
PSSM	25.66%	**39.06%**	10.71%	**19.56%**
HERN	**32.73%**	35.67%	**13.39%**	16.97%
MEAN	29.17%	36.77%	10.47%	16.66%

Table A1: **Summary AAR statistics** The average AAR for the 60 sequences in the RAbD test set. Bolded values are the top scoring model on each region of the CDR3 being considered. "All CDR3 sites" means the AAR is calculated on the full CDR3 sequence. "Middle loop sites" means the AAR is calculated on only the middle loop sites (excluding the first two and last three sites in CDR3 sequence).

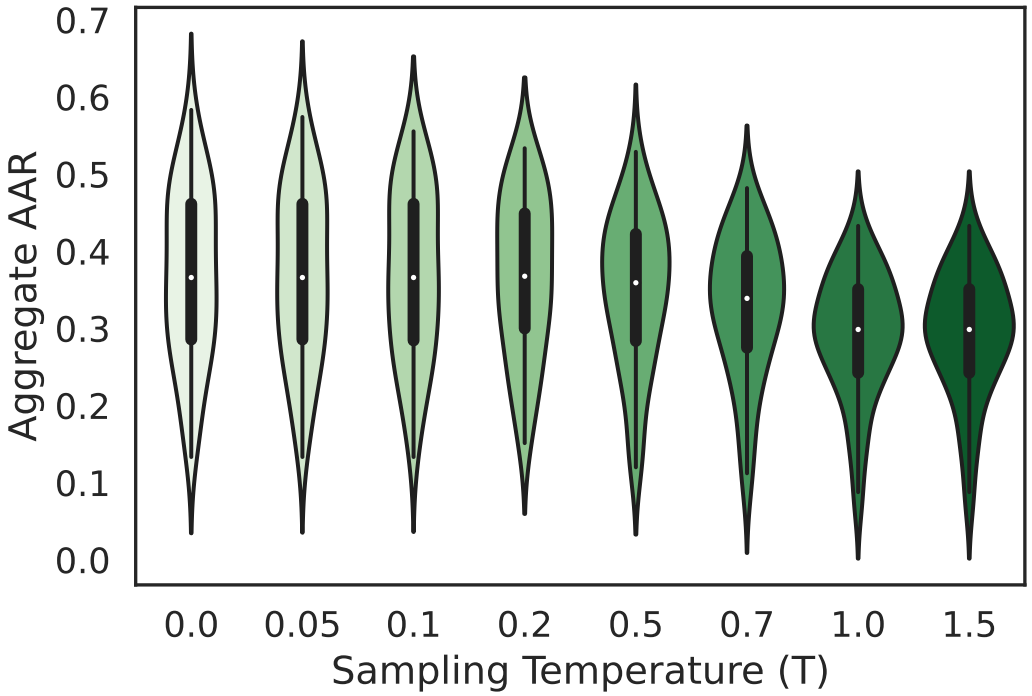

Figure A1: **MEAN AAR by sampling temperature** The average AAR across the whole CDR3 for MEAN [Kong et al., 2022] sampled using varying temperature settings.

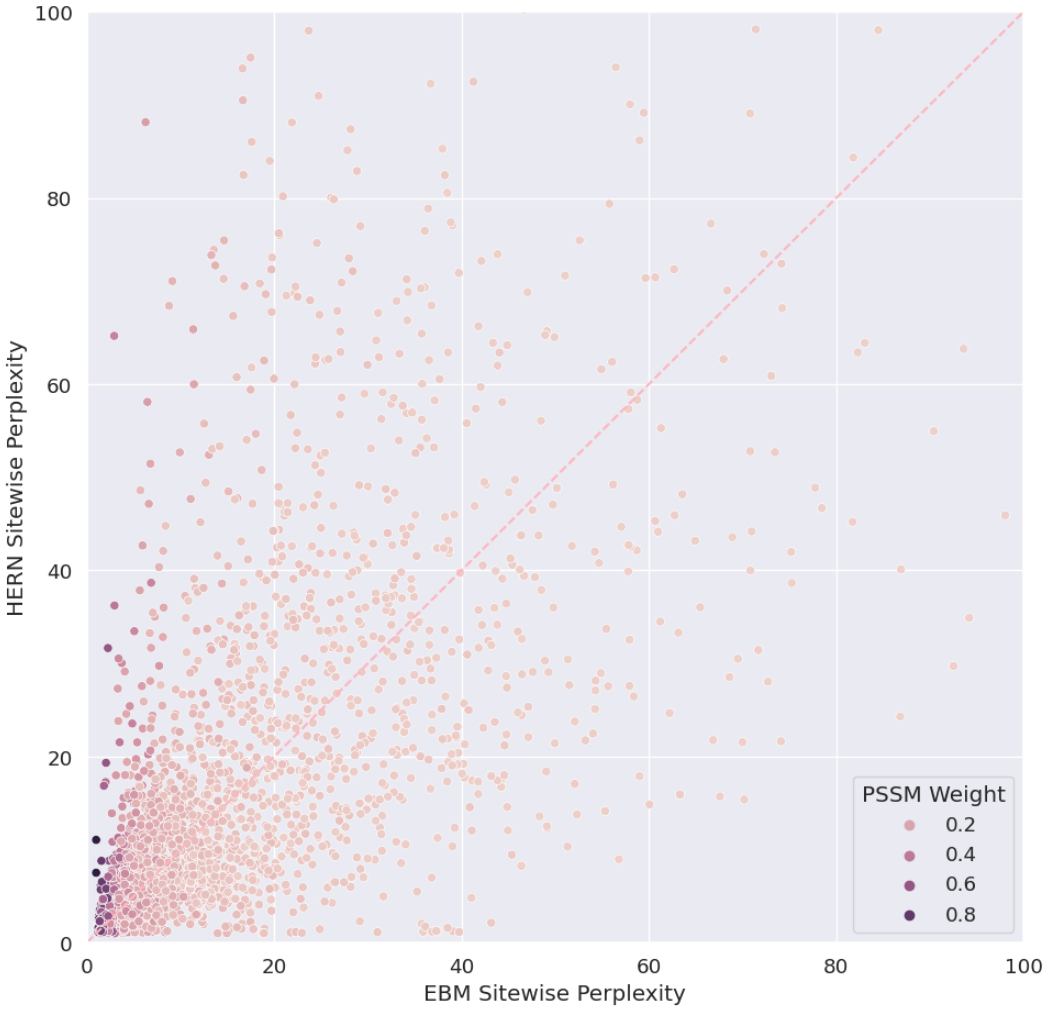

Figure A2: Here we show the HERN sitewise perplexity against EBM sitewise perplexity. Each point in the plot is a site in a CDR3 sequence in the test dataset. The color of the point represents the probability of that residue under the PSSM (PSSM weight). The dotted line corresponds to the identity line.