Epitope-specific antibody design using diffusion models on the latent space of ESM embeddings

Tomer Cohen
School of Computer Science and Engineering
The Hebrew University of Jerusalem
Jerusalem, Israel
tomer.cohen13@mail.huji.ac.il

Dina Schneidman-Duhovny
School of Computer Science and Engineering
The Hebrew University of Jerusalem
Jerusalem, Israel
dina.schneidman@mail.huji.ac.il

Abstract

There was a significant progress in protein design using deep learning approaches. The majority of methods predict sequences for a given structure. Recently, diffusion approaches were developed for generating protein backbones. However, *de novo* design of epitope-specific antibody binders remains an unsolved problem due to the challenge of simultaneous optimization of the antibody sequence, variable loop structures, and antigen binding. Here we present, EAGLE (Epitope-specific Antibody Generation using Language model Embeddings), a diffusion-based model that does not require input backbone structures. The full antibody sequence (constant and variable regions) is designed in the continuous space using protein language model embeddings. Similarly to denoising diffusion probabilistic models for image generation that condition the sampling on a text prompt, here we condition the sampling of antibody sequences on antigen structure and epitope amino acids. The model is trained on the available antibody and antibody-antigen structures, as well as antibody sequences. Our Top-100 designs include sequences with 55% identity to known binders for the most variable heavy chain loop. EAGLE's high performance is achieved by tailoring the method specifically for antibody design through integration of continuous latent space diffusion and sampling conditioned on antigen structure and epitope amino acids. Our model enables generating a wide range of diverse, unique, variable loop length antibody binders using straightforward epitope specifications.

1 Introduction

Antibody-based biotherapeutics represent a rapidly growing class of biologics that have significantly transformed the landscape of the biopharmaceutical industry. There are over 100 approved antibody-based therapeutics and over 1,000 in clinical studies for a wide range of diseases, including cancer, autoimmunity, inflammatory diseases, and viral infections [1]. Antibodies consist of two chains (light and heavy), with conserved frame regions and three variable loops (Complementarity Determining Regions - CDRs) on each chain (Fig. S1). In a typical antibody discovery project, animal immunization or display libraries [2–4] are used to generate antibodies for a specific target. Neutralizing antibodies can also be isolated from virus outbreak survivors, as in Ebola [5] or SARS-CoV-2 [6]. One major challenge is to identify antibodies that target specific epitopes with high affinity out of multiple candidates for further development [7, 8]. While antigens possess multiple epitopes, certain ones among them may serve as more favorable targets from a therapeutic standpoint. For example, binding to a highly conserved epitope reduces the risk of viral escape and extends the effectiveness of antibodies [9–11].

NeurIPS 2023 Generative AI and Biology Workshop.

Deep learning approaches have been highly successful in protein design relying on generative models [12]. The antibody design field has begun to investigate deep generative models because of their computational efficiency, which surpasses that of conventional physics-based models [13–15]. The first group of models uses protein language models (pLMs), both trained on all protein sequences [16–18] or only on antibody sequences [19–21]. These models enable sampling antibody sequences from the space of naturally observed antibodies, thus increasing the likelihood of expression and folding of the designs. The second group of models aims to find sequences for a given backbone structure [22–24]. This is applicable in a setting where some initial antibody binder and its structure with an antigen is available and our goal is to further optimize the binder [25]. Finally, the third group of models attempts to simultaneously co-design antibody sequence and structure. These methods most often rely on graph neural networks, designing antibody sequences in an autoregressive manner. The main drawbacks are that they are limited to CDRs of a fixed length and are not epitope-specific [26, 27]. When epitopes are considered, only the CDR H3 sequence is redesigned while the rest of the antibody sequence is fixed [28–30].

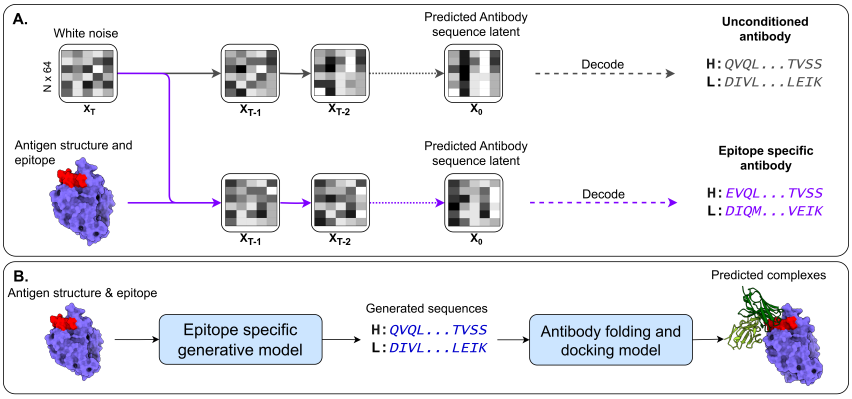

Figure 1: **A.** Sampling antibody sequences without (top) and with epitope conditioning (bottom). **B.** A pipeline for epitope-specific sequence and structure generation.

Most recent co-design approaches rely on diffusion models that have proven themselves in text-to-image generation, to co-design sequence and structure [31–33]. Diffusion models are deep generative models that work by adding Gaussian noise to the available training data (also known as the forward diffusion process) and then reversing the process (known as denoising or the reverse diffusion process) to recover the data [34]. The model gradually learns to remove the noise. The reverse process can be used for generating new data points. Similar to text-to-image generation, where image denoising is conditioned on a text prompt, in protein design, sequence generation can be conditioned on shape, symmetry, or binder. Despite the impressive performance, these methods work well for designing structured regions that include helices or sheets [31–33]. Similarly to antibody-antigen structure prediction [35–37], tailored models are needed for antibodies that bind through highly variable loops. DiffAb [38] relies on a multinomial sequence diffusion [39] to co-design antibody CDR sequences and their structure but requires a starting structure of antibody framework oriented relative to antigen. While AbDiffuser can co-design sequence and structure of variable length without a need for a starting structure, it does not consider the antigen or the epitope [40]. Despite recent advances, *de novo* design of epitope-specific antibody binders including constant regions and variable length CDR loops is not possible with current methods.

Here, we will apply techniques that are used for text conditioned image generation to generate antibody sequences (Fig. 1, 2). We condition antibody sequence generation on the antigen structure and epitope amino acids. To account for inter-dependencies between the frame and CDR loops, as well as light and heavy chains, we design the full antibody sequence using diffusion in the continuous space of ESM embeddings. Our designs are not limited to fixed-length CDR loops. Moreover, no starting structure of the antibody (or frame) relative to antigen is required. To validate epitope-specificity of the designed sequences, we use a structure generation model that simultaneously folds the antibody and docks it to the antigen [37].

2 Methods

Problem definition. Here we address the most general setting of *de novo* design: given an antigen structure and the epitope site (list of antigen amino acids), generate antibody sequences that bind to the given epitope. In the context of denoising diffusion models, instead of generating images conditioned by a text prompt, we would like to generate antibody sequences conditioned on an epitope sequence and structure.

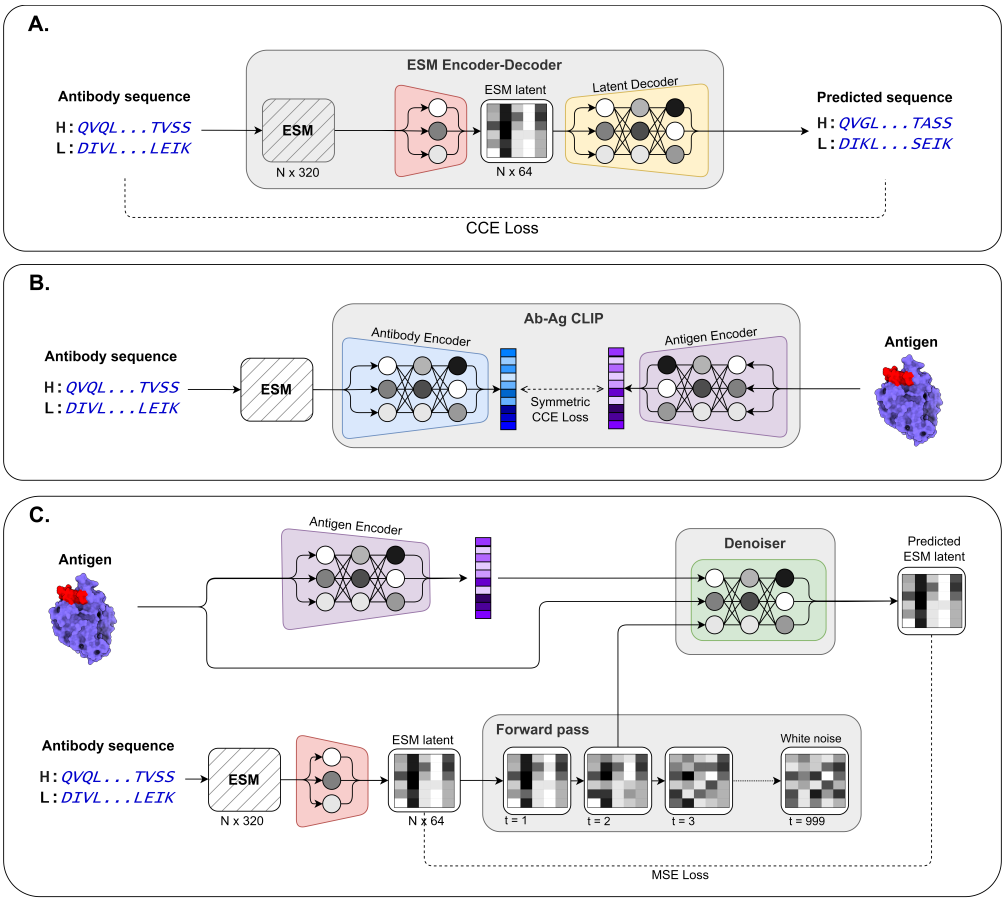

Figure 2: **Training setup. A.** ESM encoder-decoder model. **B.** CLIP model for antibody-antigen pairs. **C.** Denoiser model: the antigen CLIP encoder (purple) and ESM encoder (red) are fixed during the training.

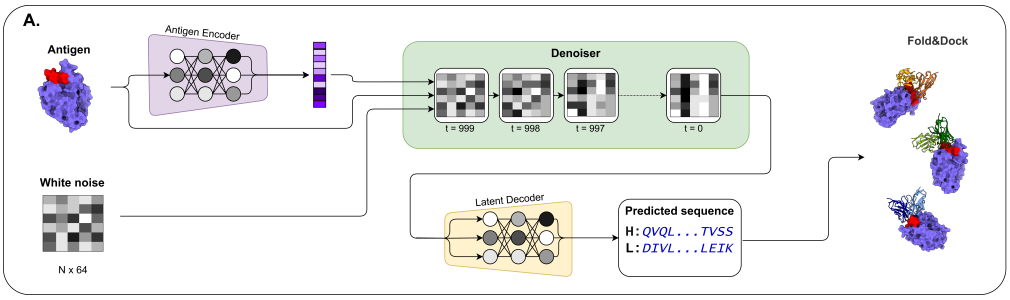

Figure 3: **Sampling setup.** The sampling relies on the pre-trained antigen CLIP encoder (purple), denoiser (green), and ESM decoder (yellow) modules. The generated antibody sequences are folded and docked to an antibody using a Fold&Dock model [37].

Denoising diffusion for latent space generation of sequences. We rationalized that a diffusion process will work better in the continuous pLMs embeddings space, that capture information about the local and global contexts of amino acids rather than in the discrete categorical space of 20 amino acids. For this purpose, we represent the antibody sequences by their ESM [41] residue embeddings which are continuous and are essentially like a color in images. Diffusion models for image generation have been currently used to generate images with smaller dimensions (about 64x64), that are later converted to a high resolution image [42, 43]. The smallest ESM amino acid embedding dimension currently available is 320, so we first train an encoder-decoder model (A.1.1, Fig. 2A). The encoder reduces an antibody ESM embedding (Nx320) to a latent dimension of Nx64 in the range of [-1,1]. The diffusion model operates on this smaller dimension representation. The decoder model converts the latent representation back to the original antibody sequence.

In the forward process we add Gaussian noise to the antibody latent space embeddings with a noise scheduler until we reach a standard normal distribution (A.1.5). In the reverse process we train a denoiser model to gradually move from a standard normal distribution (white noise) to the antibody ESM latent embedding space (Nx64) and use the trained decoder to convert to an amino acid sequence.

Epitope conditioning. To condition the sampling on an antigen structure and a specific epitope, we use a 'classifier-free guidance' technique [44] that trains the denoiser model without the antigen epitope for 10% of the time. Inspired by text-to-image generative models such as DALL·E [43], we designed a Contrastive Language-Image Pretraining (CLIP) [45] - like model that greatly improved the results of diffusion models for conditional image generation. Our CLIP model involves training two encoders to generate embeddings for antibody sequence and antigen structure, respectively, with the goal of maximizing the cosine similarities between interacting antibody-antigen pairs and minimizing the cosine similarities between non-interactive pairs in a contrastive manner (Fig. 2B, A.1.2). This is achieved by training two encoders simultaneously with symmetrical cross-entropy loss. We later use the antigen embedding as an additional input for the denoiser model to provide information about antibody sequences that can bind to the antigen, guiding the diffusion process in the right direction. The CLIP-like model was trained prior to the denoiser model.

Denoiser architecture. The denoiser model consists of four Transformer modules, each accounting for antibody-antibody, antibody-antigen, antigen-antibody, and antigen-antigen interactions (A.1.3). The input for the denoiser model is the noised antibody ESM latent embeddings, the timestep t, the antigen sequence and structure, epitope amino acids, and CLIP epitope representation (A.1.4). The output is the predicted noise added at time t. The loss is defined as a mean squared error (MSE) loss between the predicted and the actual noise (Fig. 2C).

Docking and scoring. The generated sequences are folded and docked to antigens, followed by ranking (Fig. 1B, 3) using the fast 'Fold&Dock' model [37] .

Datasets. Structures for the training of the Denoiser and CLIP models were obtained from the SabDAB database [46]. A total of 8,411 structures were used for training and validation. For test set, we retrieved all the antibodies from SabDAB that were published after the data for training was obtained and had at least three different amino acids in CDR3 from each CDR3 in both the training and validation sets, resulting in 71 structures. For the training of the ESM encoder-decoder model which requires only antibody sequences without structures, we used sequences from the OAS database [47] (A.1.6).

3 Results

Evaluation metrics. In nature, multiple diverse antibody sequences bind to the same epitope [9–11]. When we evaluate design methods, it is common to compare designed sequences to a single known binder, neglecting the fact that the space of possible binder sequences is large. Despite that, metrics that measure how similar the designed sequences to a single known binder are used due to a lack of better options. The similarity is measured by amino acid recovery (AAR), that is the fraction of correctly recovered CDR amino acids. Our model is the first one that can produce CDRs of different length for a given epitope. Therefore, we define variable length AAR (VAAR) to account for CDR length variability, dividing the number of identical aligned positions by the maximal length of the two CDRs (A.2.2). In addition, because VAAR tends to decrease for longer CDR designs compared to a true one, we also use sequence identity metrics (A.2.3). While most methods report the mean

sequence recovery over a large number of generated sequences, due to the limitation of comparison to a single binder, here we also report the maximal sequence recovery for 1,000 generated sequences per test case.

Test set VAAR. As a baseline, for each of the 1,000 generated sequence, a random sequence of the same length was generated by sampling CDRs with uniform amino acids probability over each CDR position. In addition, 1,000 sequences were generated using the same architecture trained without CLIP embeddings. We calculate maximal and mean VAAR over all generated sequences (Fig. 4A, S4A, Tab. S1). As expected, the maximal VAAR for all CDRs except for CDR H3 is high, in the range of 80%. For CDR H3, the maximal VAAR is ∼60%. Without CLIP training, the performance is slightly lower. We have also explored the effect of the weight of classifier guidance (Fig. S2, S3) and selected the value of 2.0 for the final model.

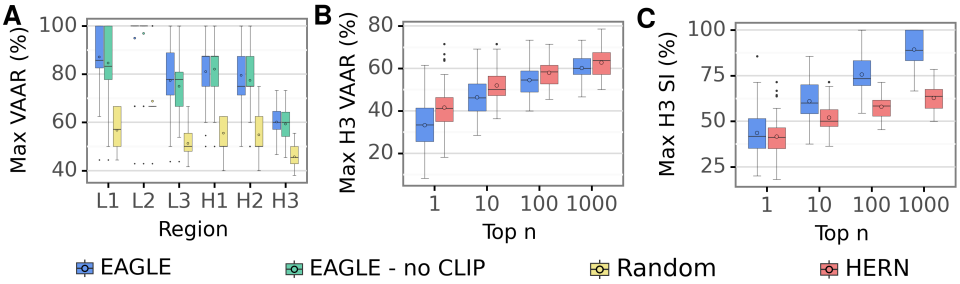

Figure 4: Test set performance. **A.** Maximal VAAR (%) by CDR loop. **B.** Maximal CDR H3 VAAR (%) compared to HERN. **B.** Maximal CDR H3 sequence identity (%) compared to HERN.

Docking score correlates with CDR H3 VAAR. The 'Fold&Dock' model typically generates hundreds of models for each antibody sequence. We focus only on models that have at least 80% overlap with the input epitope and select the best scoring one. We rank the 1,000 generated sequences using this score. To support this ranking, we test if this score correlates with CDR H3 sequence recovery. Indeed, we find such a dependency for our test set cases (Fig. S5).

Comparison to other methods. We compare our method to HERN (A.2.4), the only approach that can co-design antibody sequence and structure given only the structure of the antigen and epitope amino acids [28]. HERN only designs CDR H3 with fixed length and without considering frame and other CDR loops. We find that HERN performs slightly better for the VAAR metrics (Fig. 4B, S4B, Tab. S2) while our model performs slightly better for the sequence identity metrics (Fig. 4C, S4C, Tab. S3). The difference can be attributed to the fact that our model can produce variable length CDR loops. While the performance is comparable, EAGLE solves the most general antibody design settings, while HERN only designs fixed length CDR H3 loops.

4 Conclusion

We present a model for the most general setting of *de novo* epitope-specific antibody design by adapting ideas that advanced text-to-image generation, such as CLIP training. However, unlike in images, we suffer from small size of the training set of only a few thousands antibody-antigen structures. Another difficulty in training and validating antibody design models vs. image generation is that sequence recovery of a single known binder is the main approach to assessment of designs without labor-intensive lab experiments. We anticipate that further progress in the accuracy of antibody-antigen docking and scoring models will aid in addressing those bottlenecks. Further experimental validation of EAGLE designs will help to improve the method.

References

[1] Hélène Kaplon, Silvia Crescioli, Alicia Chenoweth, Jyothsna Visweswaraiah, and Janice M Reichert. Antibodies to watch in 2023. In *MAbs*, volume 15, page 2153410. Taylor & Francis, 2023.

[2] Sirid-Aimée Kellermann and Larry L Green. Antibody discovery: the use of transgenic mice to generate human monoclonal antibodies for therapeutics. *Current opinion in biotechnology*, 13(6):593–597, 2002.

[3] Juan C Almagro, Martha Pedraza-Escalona, Hugo Iván Arrieta, and Sonia Mayra Pérez-Tapia. Phage display libraries for antibody therapeutic discovery and development. *Antibodies*, 8(3):44, 2019.

[4] Andreas H Laustsen, Victor Greiff, Aneesh Karatt-Vellatt, Serge Muyldermans, and Timothy P Jenkins. Animal immunization, in vitro display technologies, and machine learning for antibody discovery. *Trends in Biotechnology*, 39(12):1263–1273, 2021.

[5] Z. A. Bornholdt, H. L. Turner, C. D. Murin, W. Li, D. Sok, C. A. Souders, A. E. Piper, A. Goff, J. D. Shamblin, S. E. Wollen, T. R. Sprague, M. L. Fusco, K. B. Pommert, L. A. Cavacini, H. L. Smith, M. Klempner, K. A. Reimann, E. Krauland, T. U. Gerngross, K. D. Wittrup, E. O. Saphire, D. R. Burton, P. J. Glass, A. B. Ward, and L. M. Walker. Isolation of potent neutralizing antibodies from a survivor of the 2014 ebola virus outbreak. *Science*, 351(6277):1078–83, 2016. ISSN 1095-9203 (Electronic) 0036-8075 (Linking). doi: 10.1126/science.aad5788. URL `http://www.ncbi.nlm.nih.gov/pubmed/26912366`.

[6] Seth J Zost, Pavlo Gilchuk, Rita E Chen, James Brett Case, Joseph X Reidy, Andrew Trivette, Rachel S Nargi, Rachel E Sutton, Naveenchandra Suryadevara, and Elaine C Chen. Rapid isolation and profiling of a diverse panel of human monoclonal antibodies targeting the sars-cov-2 spike protein. *Nature medicine*, 26(9):1422–1427, 2020. ISSN 1546-170X.

[7] T. Jain, T. Sun, S. Durand, A. Hall, N. R. Houston, J. H. Nett, B. Sharkey, B. Bobrowicz, I. Caffry, Y. Yu, Y. Cao, H. Lynaugh, M. Brown, H. Baruah, L. T. Gray, E. M. Krauland, Y. Xu, M. Vasquez, and K. D. Wittrup. Biophysical properties of the clinical-stage antibody landscape. *Proc Natl Acad Sci U S A*, 114(5): 944–949, 2017. ISSN 1091-6490 (Electronic) 0027-8424 (Linking). doi: 10.1073/pnas.1616408114. URL `https://www.ncbi.nlm.nih.gov/pubmed/28096333`.

[8] Jie Zhou, Chau Q. Le, Yun Zhang, and James A. Wells. A general approach for selection of epitope-directed binders to proteins. *bioRxiv*, 2023. doi: 10.1101/2022.10.24.513434. URL `https://www.biorxiv.org/content/early/2023/09/23/2022.10.24.513434`.

[9] Y. Xiang, W. Huang, H. Liu, Z. Sang, S. Nambulli, J. Tubiana, Jr. Williams, K. L., W. P. Duprex, D. Schneidman-Duhovny, I. A. Wilson, D. J. Taylor, and Y. Shi. Superimmunity by pan-sarbecovirus nanobodies. *Cell Rep*, 39(13):111004, 2022. ISSN 2211-1247 (Electronic). doi: 10.1016/j.celrep.2022.111004. URL `https://www.ncbi.nlm.nih.gov/pubmed/35738279`.

[10] Adam S Dingens, Dana Arenz, Haidyn Weight, Julie Overbaugh, and Jesse D Bloom. An antigenic atlas of hiv-1 escape from broadly neutralizing antibodies distinguishes functional and structural epitopes. *Immunity*, 50(2):520–532, 2019.

[11] Nicholas C Wu and Ian A Wilson. Influenza hemagglutinin structures and antibody recognition. *Cold Spring Harbor perspectives in medicine*, 10(8):a038778, 2020.

[12] Michael Chungyoun and Jeffrey J Gray. Ai models for protein design are driving antibody engineering. *Current Opinion in Biomedical Engineering*, page 100473, 2023.

[13] B. D. Weitzner, D. Kuroda, N. Marze, J. Xu, and J. J. Gray. Blind prediction performance of rosettaantibody 3.0: grafting, relaxation, kinematic loop modeling, and full cdr optimization. *Proteins*, 82(8):1611–23, 2014. ISSN 1097-0134 (Electronic) 0887-3585 (Linking). doi: 10.1002/prot.24534. URL `https://www.ncbi.nlm.nih.gov/pubmed/24519881`.

[14] J. Adolf-Bryfogle, O. Kalyuzhniy, M. Kubitz, B. D. Weitzner, X. Hu, Y. Adachi, W. R. Schief, and Jr. Dunbrack, R. L. Rosettaantibodydesign (rabd): A general framework for computational antibody design. *PLoS Comput Biol*, 14(4):e1006112, 2018. ISSN 1553-7358 (Electronic) 1553-734X (Linking). doi: 10.1371/journal.pcbi.1006112. URL `https://www.ncbi.nlm.nih.gov/pubmed/29702641`.

[15] Shira Warszawski, Aliza Borenstein Katz, Rosalie Lipsh, Lev Khmelnitsky, Gili Ben Nissan, Gabriel Javitt, Orly Dym, Tamar Unger, Orli Knop, Shira Albeck, et al. Optimizing antibody affinity and stability by the automated design of the variable light-heavy chain interfaces. *PLoS computational biology*, 15(8): e1007207, 2019.

[16] 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.

[17] Noelia Ferruz, Steffen Schmidt, and Birte Höcker. Protgpt2 is a deep unsupervised language model for protein design. *Nature communications*, 13(1):4348, 2022.

[18] Vladimir Gligorijević, Daniel Berenberg, Stephen Ra, Andrew Watkins, Simon Kelow, Kyunghyun Cho, and Richard Bonneau. Function-guided protein design by deep manifold sampling. *bioRxiv*, pages 2021–12, 2021.

[19] Richard W Shuai, Jeffrey A Ruffolo, and Jeffrey J Gray. Generative language modeling for antibody design. *bioRxiv*, pages 2021–12, 2021.

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

[21] 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.

[22] John Ingraham, Vikas Garg, Regina Barzilay, and Tommi Jaakkola. Generative models for graph-based protein design. *Advances in neural information processing systems*, 32, 2019.

[23] Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J Ragotte, Lukas F Milles, Basile IM Wicky, Alexis Courbet, Rob J de Haas, Neville Bethel, et al. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378(6615):49–56, 2022.

[24] 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. *ICML*, 2022. doi: 10.1101/2022.04.10.487779. URL `https://www.biorxiv.org/content/early/2022/04/10/2022.04.10.487779`.

[25] Sai Pooja Mahajan, Jeffrey A Ruffolo, Rahel Frick, and Jeffrey J Gray. Hallucinating structure-conditioned antibody libraries for target-specific binders. *Frontiers in immunology*, 13:999034, 2022.

[26] 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.

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

[28] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Antibody-antigen docking and design via hierarchical equivariant refinement. *arXiv preprint arXiv:2207.06616*, 2022.

[29] Kaiyuan Gao, Lijun Wu, Jinhua Zhu, Tianbo Peng, Yingce Xia, Liang He, Shufang Xie, Tao Qin, Haiguang Liu, Kun He, et al. Incorporating pre-training paradigm for antibody sequence-structure co-design. *bioRxiv*, pages 2022–11, 2022.

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

[31] 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*, pages 1–3, 2023.

[32] Alexander E Chu, Lucy Cheng, Gina El Nesr, Minkai Xu, and Po-Ssu Huang. An all-atom protein generative model. *bioRxiv*, pages 2023–05, 2023.

[33] John Ingraham, Max Baranov, Zak Costello, Vincent Frappier, Ahmed Ismail, Shan Tie, Wujie Wang, Vincent Xue, Fritz Obermeyer, Andrew Beam, et al. Illuminating protein space with a programmable generative model. *BioRxiv*, pages 2022–12, 2022.

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

[35] B. D. Weitzner, J. R. Jeliazkov, S. Lyskov, N. Marze, D. Kuroda, R. Frick, J. Adolf-Bryfogle, N. Biswas, Jr. Dunbrack, R. L., and J. J. Gray. Modeling and docking of antibody structures with rosetta. *Nat Protoc*, 12(2):401–416, 2017. ISSN 1750-2799 (Electronic) 1750-2799 (Linking). doi: 10.1038/nprot.2016.180. URL `http://www.ncbi.nlm.nih.gov/pubmed/28125104`.

[36] Francesco Ambrosetti, Brian Jiménez-García, Jorge Roel-Touris, and Alexandre MJJ Bonvin. Modeling antibody-antigen complexes by information-driven docking. *Structure*, 28(1):119–129, 2020.

[37] Tomer Cohen and Dina Schneidman-Duhovny. End-to-end accurate and high-throughput modeling of antibody-antigen complexes. In *NeurIPS Workshop on Machine Learning for Structural Biology*, 2022.

[38] 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.

[39] Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. Argmax flows and multinomial diffusion: Learning categorical distributions. *Advances in Neural Information Processing Systems*, 34:12454–12465, 2021.

[40] Karolis Martinkus, Jan Ludwiczak, Kyunghyun Cho, Wei-Ching Lian, Julien Lafrance-Vanasse, Isidro Hotzel, Arvind Rajpal, Yan Wu, Richard Bonneau, Vladimir Gligorijevic, et al. Abdiffuser: Full-atom generation of in-vitro functioning antibodies. *arXiv preprint arXiv:2308.05027*, 2023.

[41] 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.

[42] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

[43] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[44] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[45] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[46] J. Dunbar, K. Krawczyk, J. Leem, T. Baker, A. Fuchs, G. Georges, J. Shi, and C. M. Deane. Sabdab: the structural antibody database. *Nucleic Acids Res*, 42(Database issue):D1140–6, 2014. ISSN 1362-4962 (Electronic) 0305-1048 (Linking). doi: 10.1093/nar/gkt1043. URL `https://www.ncbi.nlm.nih.gov/pubmed/24214988`.

[47] 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, 2022.

[48] 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.

[49] Annemarie Honegger and Andreas PluÈckthun. Yet another numbering scheme for immunoglobulin variable domains: an automatic modeling and analysis tool. *Journal of molecular biology*, 309(3):657–670, 2001.

[50] Steven Henikoff and Jorja G Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22):10915–10919, 1992.

[51] M. P. Lefranc, C. Pommie, M. Ruiz, V. Giudicelli, E. Foulquier, L. Truong, V. Thouvenin-Contet, and G. Lefranc. Imgt unique numbering for immunoglobulin and t cell receptor variable domains and ig superfamily v-like domains. *Dev Comp Immunol*, 27(1):55–77, 2003. ISSN 0145-305X (Print) 0145-305X (Linking). doi: 10.1016/s0145-305x(02)00039-3. URL `https://www.ncbi.nlm.nih.gov/pubmed/12477501`.

[52] J. Dunbar and C. M. Deane. Anarci: antigen receptor numbering and receptor classification. *Bioinformatics*, 32(2):298–300, 2016. ISSN 1367-4811 (Electronic) 1367-4803 (Linking). doi: 10.1093/bioinformatics/btv552. URL `https://www.ncbi.nlm.nih.gov/pubmed/26424857`.

A Appendix

A.1 Extended Methods

A.1.1 ESM encoder-decoder architecture

Our proposed model performs the diffusion process on the space of ESM latent embeddings. We first tried to perform the diffusion process directly on the ESM embedding (with a dimension of Nx320) [41], but found it difficult to train with lower performance compared to diffusing in a smaller dimension space. This is consistent with image diffusion models which usually perform the diffusion process on 64x64 images. Because we are performing the diffusion process in the space of ESM latent embeddings, during the sampling our model will generate those embeddings and not antibody sequences. To translate these latent embeddings back to amino acid sequence, we trained a simple transformer based encoder-decoder model that converts ESM embeddings to a lower dimensional latent space and then convert those latent embeddings back to amino acid sequences (Fig. 2A). The architecture contains an encoder and decoder that are trained simultaneously (Fig. 2A). The encoder reduces antibody chain (heavy/light) ESM embeddings (Nx320) to a latent dimension of 64 using a single linear layer, which is then scaled to the range of [-1, 1] using a $tanh$ activation. The decoder converts this representation ('ESM latent embedding') to amino acid probabilities for each position (Nx22) using a simple BERT module [48]. The loss is defined as a categorical cross entropy between a true antibody sequence and the decoder output. The encoder-decoder model was trained on a $\sim$100M antibody sequence from a database of observed antibody sequences (OAS) [47] until high sequence recovery was reached (99.99%). Because the encoder-decoder model is trained on a huge dataset, it achieves a nearly perfect sequence recovery rate even when adding some noise to the ESM latent embedding.

A.1.2 CLIP architecture

The CLIP architecture contains two modules, antibody and antigen encoders (Fig. 2B). The antibody encoder receives an antibody sequence as an ESM embedding and outputs an antibody vector representation of dimension 128. The antibody encoder is a BERT-like model that contains a sequence of transformers. The output of the last transformer is projected and normalized into a 128 vector representation. The antigen encoder receives an antigen representation identical to the denoiser model but without the CLIP representation. The antigen encoder architecture is the same as the antibody encoder but instead of regular transformers, it uses a transformer with added distance as bias (same as the denoiser antigen-antigen module). The loss for training the CLIP is symmetrical categorical cross-entropy loss between the corresponding antigen and antibody vectors representations in the batch.

A.1.3 Denoiser architecture

The denoiser architecture is based on the previous work [37] which uses four Transformer modules (a GPA block), each responsible for a different aspect of the antibody-antigen interaction. Here, for each of the four interactions we use a simple transformer where the queries, keys and values are either the antigen amino acids representation or the antibody amino acid representation. For the antibody-antigen and the antigen-antibody transformers we use only the antigen amino acids that are part of the epitope as queries, keys, and values. For the antigen-antigen transformer we provide the antigen structure information by storing distances of amino acid atoms (N, Cα, C, O, and, Cβ atoms and five additional side chain atoms that define the χ_{1-5} angles) from Cβ atom of all other amino acids (LxLx10). This distance matrix is added to the antigen-antigen attention logits as bias before the softmax activation. For the antigen-antigen transformer we also attend only on amino acids that have Cβ-Cβ distance lower than 10Å. There are a total of three GPA blocks, each updating the antibody and antigen representations.

A.1.4 Denoiser input and output

Antibody representation. The input for the denoiser model includes the antibody noised sequence in the form of latent ESM embeddings (Nx64) with 5 additional channels for chain (light/heavy) and antibody type (mAb/heavy chain only/light chain only) resulting in a matrix of size Nx69. To support generation of CDRs with variable lengths, we represent the antibody sequence with the AHo

numbering scheme [49] that defines 149 amino acid positions, including gaps, for each chain. We define N=298 for light and heavy chains and treat a gap '-' as an additional amino acid (22 amino acids in total: 20 standard amino acids, unknown, and gap).

Antigen and epitope representation. The input to the denoiser also includes the antigen information: the one-hot encoded sequence (Lx22), the sequence in BLOSUM62 [50] representation where each amino acid is represented by a corresponding row from the matrix (Lx22), the antigen CLIP embeddings (Lx128), two additional channels for specifying binary epitope information, and one channel for surface accessible area. This results in a Lx175 matrix for antigen 1D representation. The antigen 3D structure information is represented by a Nx10x3 matrix which includes the 3D coordinates of the backbone N, Cα, C, O, and, Cβ atoms and five additional side chain atoms that define the χ_{1-5} angles. The antigen acts as a conditional same as a text prompt for image generation.

The denoiser input also includes the time step t. The output of the denoiser model is a matrix (Nx64) which is trained to match the added noise at timestep t by minimizing the MSE loss between the network prediction and the true noise added at timestep t (Fig. 2C). This output is converted to an antibody sequence using the ESM decoder (Fig. 3).

A.1.5 Denoising diffusion for antibody sequence generation

We use the same forward diffusion process as defined in (Ho at el.) [34] with the addition of classifier-free guidance [44]. Where $q(x_0)$ is the latent ESM embedding (Nx64) in the range of [-1,1]. We use $T = 1000$ timesteps and a 0.0001 to 0.02 linear β scheduler. For sampling, we use guidance weight of 2.0 (Fig. S2 ,S3) and dynamic thresholding as described in Imagen [42].

A.1.6 Datasets

Structures for the training of the Denoiser and CLIP models were obtained from the SabDAB database [46]. We used only structures with a resolution of 3.5Å or better, resulting in a total of 8,411 structures (6,375 antibodies, 1686 heavy chain only antibodies (nanobodies), 350 light chain only antibodies) for training and validation (92%, 8% respectively). 4,963 of the sequences were solved with an antigen structure, for simplicity and memory efficiency we used only antigens with up to 500 amino acids. The sequences that were solved without an antigen were used for the denoiser training as the 10% of inputs without a conditional antigen (classifier-free guidance). For test set, after finishing training our models we retrieved all the antibodies from SabDAB that were published after 08.03.2023 (the date the data for training the models was obtained) and had at least three different amino acids in CDR3 from each CDR3 in both the training and validation sets after alignment of the CDRs. This resulted in 51 antibodies and 20 nanobodies. For the training of the ESM encoder-decoder model which requires only antibody sequences without structures, we used millions of sequences obtained from the OAS database [47].

A.2 Evaluation

A.2.1 CDR definitions

To calculate amino acid recovery and sequence identity, we defined the CDRs following the IMGT numbering scheme [51] using the abnumber python package which relies on ANARCI [52].

A.2.2 Variable length amino acid recovery (VAAR)

VAAR between a true CDR, x, and a predicted CDR, y is defined as follows:

$$VAAR(x, y) = AlignmentScore(x, y)/max(|x|, |y|) \tag{1}$$

where $AlignmentScore$ is the number of matching amino acids after performing global sequence alignment.

A.2.3 CDR sequence identity

CDR sequence identity between a true CDR, x, and a predicted CDR, y is defined as follows:

$$SeqID(x, y) = AlignmentScore(x, y)/min(|x|, |y|) \tag{2}$$

where $AlignmentScore$ is the number of matching amino acids after performing global sequence alignment.

A.2.4 Comparison to HERN

We used the trained model provided in the github repository of HERN [28]. We converted our test set to the format needed for HERN using the provided scripts in the repository and generated 1,000 sequences ranked by HERN score.

A.2.5 Runtimes

Generation of 1,000 sequences takes about 2.5 hours on a RTX2080 GPU with 8Gb.

A.3 Figures

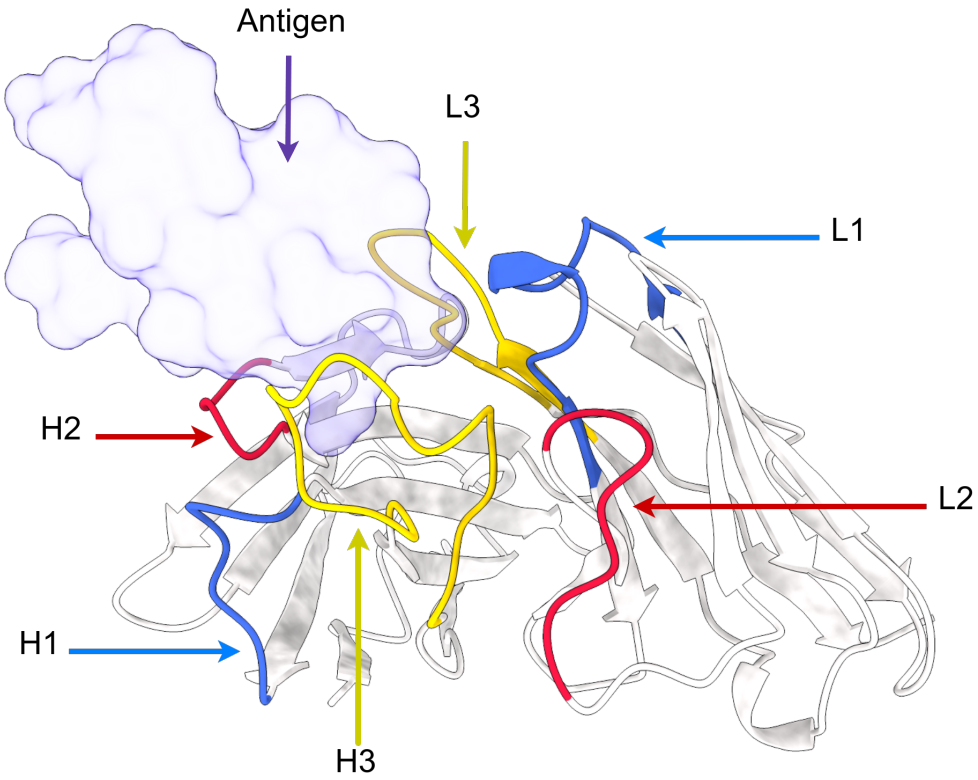

Figure S1: **Antibody recognition**. An antibody bound to an antigen (PDB 6xm2). The variable CDR loops are labeled L1, L2, L3 and H1, H2, H3 for light and heavy chains respectively.

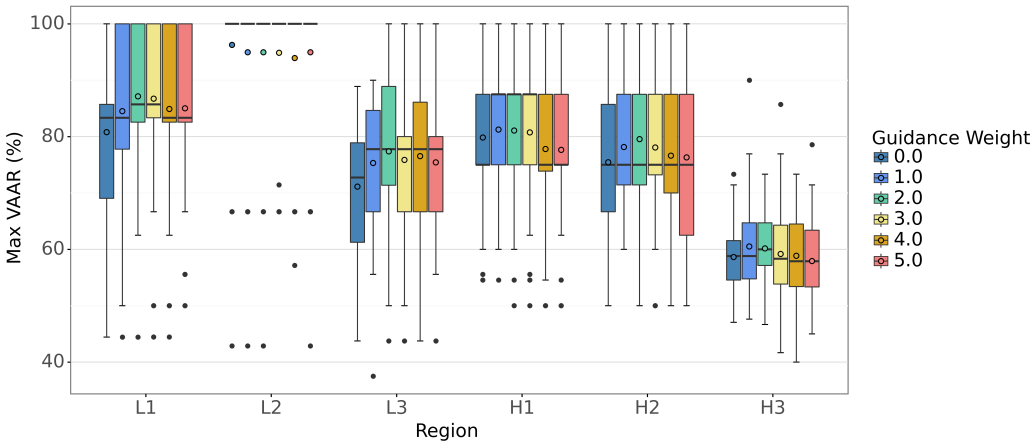

Figure S2: Maximal VAAR (%) for 1,000 generated sequences for each antibody in the test set by CDR loop. Guidance weight of 0.0 corresponds to an antibody sequence design without an epitope information

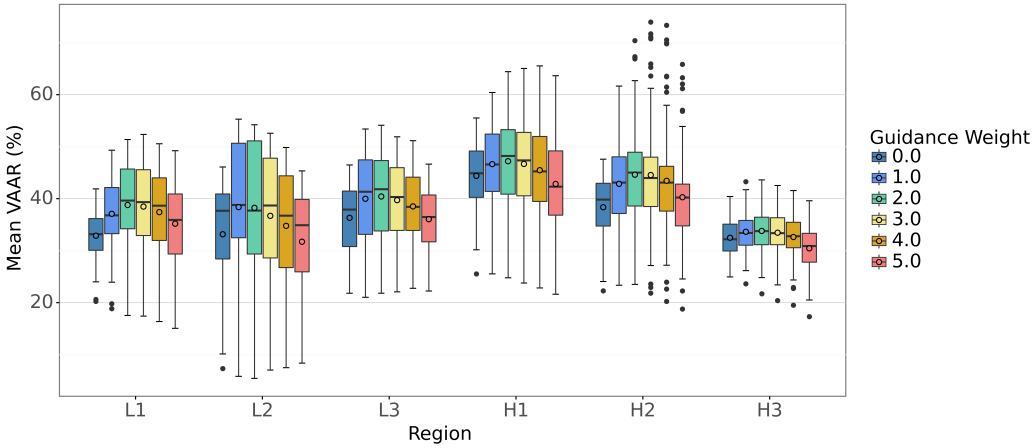

Figure S3: Average VAAR (%) for 1,000 generated sequences for each antibody in the test set by CDR loop. Guidance weight of 0.0 corresponds to an antibody sequence design without an epitope information

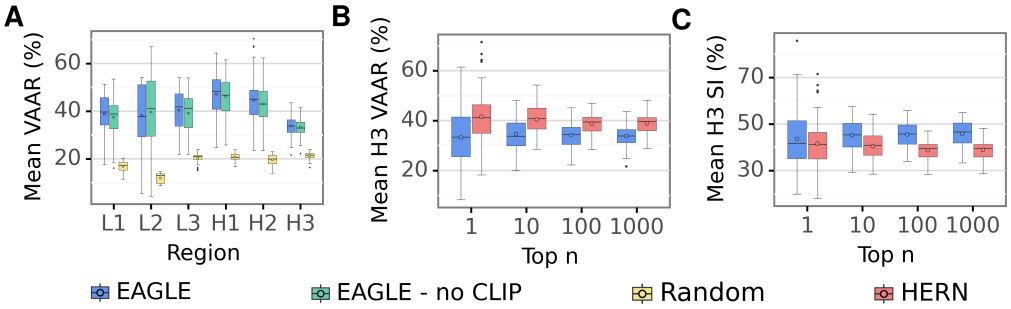

Figure S4: Test set performance. **A.** Average VAAR (%) by CDR loop. **B.** Average CDR H3 VAAR (%) compared to HERN. **B.** Average CDR H3 sequence identity (%) compared to HERN.

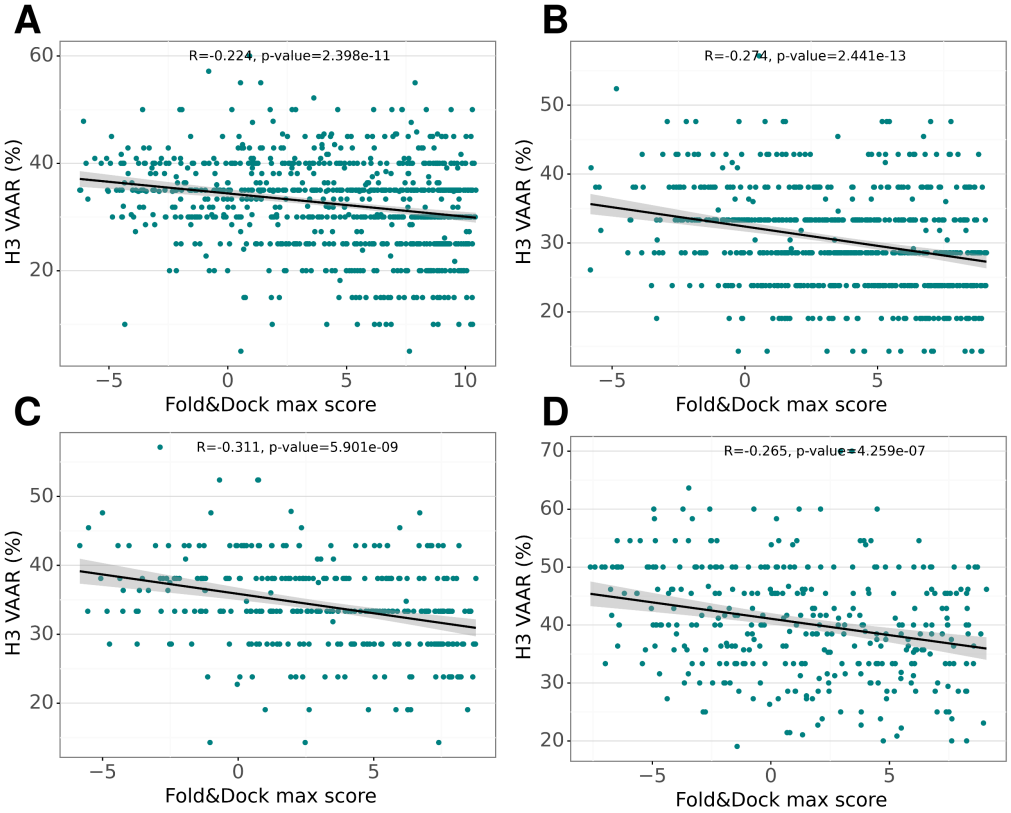

Figure S5: Correlation between Fold&Dock maximal score for a complex with >80% true epitope overlap and CDR H3 VAAR (%) for four test set antibodies. **A.** PDB 8ELO. **B.** PDB 8TCO. **C.** PDB 8G3P. **D.** PDB 8GS9.

A.4 Tables

	H1		H2		H3		L1		L2		L3	
Method	Max	Mean	Max	Mean	Max	Mean	Max	Mean	Max	Mean	Max	Mean
Random	55.55	20.76	54.84	19.64	45.71	21.34	56.75	16.85	68.81	11.94	51.30	20.45
EAGLE - no CLIP	**82.10**	46.17	77.46	43.00	59.32	33.21	84.63	37.46	**96.92**	**39.66**	74.90	39.18
EAGLE	81.07	**47.19**	**79.55**	**44.60**	**60.16**	**33.78**	**87.14**	**38.77**	94.96	38.23	**77.37**	**40.41**

Table S1: Averages over the 71 test cases of the maximal and average VAAR (%) of 1,000 generated sequences by CDR loop.

	T1	T10		T100		T1000	
Method	Max/Mean	Max	Mean	Max	Mean	Max	Mean
HERN	**41.56**	**51.96**	**40.48**	**57.93**	**38.75**	**62.75**	**38.88**
EAGLE	33.27	46.40	34.53	54.45	34.16	60.16	33.78

Table S2: Averages over the 71 test cases of the maximal and average H3 VAAR (%) of top-n generated sequences.

	T1	T10		T100		T1000	
Method	Max/Mean	Max	Mean	Max	Mean	Max	Mean
HERN	41.56	51.96	40.48	57.93	38.75	62.75	38.88
EAGLE	**43.57**	**60.93**	**45.18**	**75.60**	**45.48**	**89.31**	**45.88**

Table S3: Averages over the 71 test cases of the maximal and average H3 sequence identity (%) of top-n generated sequences.