A GUIDED DESIGN FRAMEWORK FOR THE OPTIMIZATION OF THERAPEUTIC-LIKE ANTIBODIES

Amy Wang[*,1], **Zhe Sang**[1,3], **Samuel D. Stanton**[1], **Jennifer L. Hofmann**[1], **Saeed Izadi**[2], **Eliott Park**[2], **Jan Ludwiczak**[1], **Matthieu Kirchmeyer**[1], **Darcy Davidson**[1], **Andrew Maier**[2], **Tom Pritsky**[2], **Nathan C. Frey**[1], **Andrew M. Watkins**[1], **and Franziska Seeger**[1]

[1]Prescient Design, Genentech
[2]Pharmaceutical Technology Development, Genentech
[3]Icahn School of Medicine at Mount Sinai, Pharmacological Sciences

ABSTRACT

Antibodies must meet stringent developability criteria for successful commercialization—a challenge for machine learning approaches given the limited available data. Selecting candidates with biophysical properties similar to clinical-stage antibodies offers an alternative to data-intensive approaches. However, such methods typically suffer from limited throughput due to structure-based calculations and can eliminate promising candidates through overly strict filtering. By benchmarking classical filtering methods against experimental datasets, with viscosity as a proof-of-concept, we identify an informative set of biophysical definitions (relevant to charge and hydrophobicity). Using these as optimization objectives for guided design, we introduce TherAbDesign, a sequence-based framework that evaluates and optimizes antibodies for developability without requiring structure prediction or physics-based computation. TherAbDesign proposes rational modifications to mimic the properties of successful therapeutic antibodies, which we demonstrate can improve known developability liabilities like high viscosity without explicitly modeling their mechanism of action.

1 INTRODUCTION

Antibodies are a powerful class of therapeutics due to their inherent specificity and efficacy. However, developing an antibody into a successful drug requires optimizing multiple "developability" properties beyond binding affinity – including thermostability, aggregation, and viscosity (Carter & Rajpal, 2022; Jain et al., 2017). These properties are crucial since they determine whether an antibody can be reliably produced and delivered to patients. Currently, assessing these properties requires substantial costs and experimental effort, making it impractical to evaluate them early in development. This delay is problematic because addressing issues with one property can create new problems with others, often leading to costly iterative optimization or project termination Frey et al. (2025). Consequently, there is strong interest in computational methods that can predict these properties early in development, enabling more efficient optimization of therapeutic candidates.

To enable earlier screening, computational metrics are often used to filter candidates with potential risks by comparing their biophysical properties to those of clinical-stage antibodies (Raybould et al., 2019; 2024; Park & Izadi, 2024). This approach draws inspiration from Lipinski's rule of 5, a set of guidelines that revolutionized small molecule drug development by identifying simple biophysical properties that predict drug-like behavior (Lipinski, 2000). However, applying similar principles to antibodies has proven challenging for two reasons. First, there is significant variability in how computational metrics are defined and calculated. Current methods vary significantly in surface definitions, structure preparation, and the incorporation of molecular dynamics (MD) simulations (Appendix A.1). Second, filtering approaches that are too stringent may exclude viable candidates, while those that are too permissive offer little practical value. This challenge is compounded by the massive scale of antibody discovery campaigns, where evaluation of millions of candidates for early-stage screening is computationally prohibitive.

[*]Correspondence to: wanga84@gene.com

We address developability challenges with our method, **TherAbDesign** – a general sequence-based framework that efficiently evaluates and optimizes antibodies for multiple biophysical properties characteristic of clinical-stage antibodies. To determine an informative set of molecular descriptors, we evaluate established developability filters on their ability to identify experimental liabilities. We focus on viscosity – a property critical for drug delivery but challenging to measure early in development (Appendix A.2). When used as a filter, TherAbDesign not only achieves results comparable to structure-based methods but also efficiently screens large libraries without the need for physics-based computation. When used for guided design, TherAbDesign makes targeted electrostatic and hydrophobic modifications to a parental antibody without requiring domain expertise. We observe that the sequences proposed by TherAbDesign overlap with known variants that reduce viscosity. This approach represents a significant advance in computational antibody design by enabling simultaneous optimization of multiple developability properties early in the discovery process.

2 BACKGROUND

ML Approaches for Developability: It may appear that the most direct approach to identifying liabilities would be to train predictive models for each property of interest. However, this strategy faces significant challenges: (1) limited data availability hinders model generalization; (2) existing datasets are often heterogeneous; and (3) diverse underlying mechanisms may prevent models from learning consistent functional principles. Despite these limitations, a substantial body of literature aims to predict function with specialized models, particularly viscosity due to its relatively greater data availability (Sharma et al., 2014; Makowski et al., 2024; Li et al., 2025). To overcome data limitations, many approaches incorporate biophysical descriptors as features. This is thought to improve predictive power as antibody function is broadly governed by biophysical mechanisms, particularly electrostatics and hydrophobicity (Esfandiary et al., 2015; Hoerschinger et al., 2023; Bashour et al., 2024). However, specialized property prediction models still suffer from poor generalization.

Biophysics-based Risk Assessment: Comparing surface properties of candidate antibodies to those of successful therapeutics provides a holistic characterization of developability (Lipinski, 2000; Ahmed et al., 2021). Several computational tools implement this comparative approach, such as Therapeutic Antibody Profiler (TAP) and MolDesk, which use clinical stage antibodies as a reference set to identify biophysical outlier thresholds (Raybould et al., 2019; Park & Izadi, 2024). While these approaches both focus on electrostatic and hydrophobic properties of antibody surfaces, they differ substantially in their precise definitions (see Appendix A.1 for details). TAP employs custom definitions, whereas MolDesk applies thresholds derived from established biomolecular methods such as APBS (Adaptive Poisson-Boltzmann Solver) and SAP (Spatial Aggregation Propensity) (Voynov et al., 2009; Holst et al., 2000). It remains unclear how different descriptor definitions generally affect the agreement between computational risk assessment and experimental outcomes (Waibl et al., 2021; Licari et al., 2023).

Our contributions: We clarify the role of biophysical parameterizations by presenting a systematic benchmark on descriptor definitions against clinically relevant viscosity measurements. By establishing a stronger connection between biophysical properties and experimental outcomes, we inform physics-based data augmentation for ML approaches. This addresses limitations in generalization typical of specialized property prediction models.

We introduce TherAbDesign, a novel physics-informed ML framework for holistic antibody developability assessment. While recent studies have applied sequence-based biophysical principles to antibody design (Ismail et al., 2024), our work establishes clear connections between physics-based principles, guided design, and clinically relevant developability outcomes. TherAbDesign combines robust biophysical foundations with computational efficiency, making it suitable for both high-throughput candidate filtering and guided optimization.

3 METHODS

3.1 DESCRIPTOR DEFINITIONS

Because physics-based biophysical descriptors can be sensitive to the input structure, we examine how (1) structure preparation approaches, (2) surface definitions, and (3) MD ensemble averaging

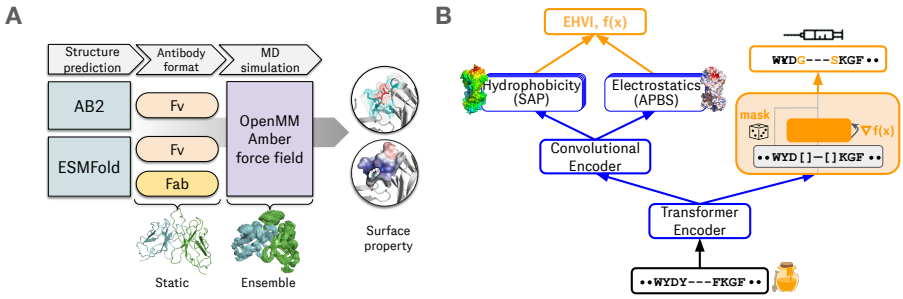


Figure 1: Analysis and optimization of antibody biophysical properties. (**A**) Workflow for biophysical descriptor benchmarking. (**B**) TherAbDesign architecture. Blue regions are pretrained on a biophysical dataset prepared from pOAS. Orange regions illustrate diffusion-based guided sampling.

affect viscosity liability identification (Figure 1A and Appendix A.1). The same set of TAP and MolDesk thresholds are applied across preparation methods, although we acknowledge that distributions and resulting thresholds may vary under different structural preparation approaches.

3.2 THERABDESIGN MODEL DETAILS

Training Data: We source sequences from pOAS to construct a dataset for general antibody biophysical property prediction (Olsen et al., 2022). To subsample, we cluster with a 95% sequence identity threshold with the Linclust algorithm (Steinegger & Söding, 2017). All Fab structures are ESMFold outputs (Lin et al., 2023). To remove misfolded structures from the dataset, we compute backbone RMSDs in PyMOL relative to a predicted herceptin Fab structure and exclude those with RMSD exceeding 5Å (Schrödinger, LLC, 2015). We compute APBS and SAP descriptors of raw, unprocessed ESMFold outputs as described above. Sequences are aligned with the AHo antibody residue numbering scheme and passed as inputs to our model (Honegger & Plückthun, 2001; Dunbar & Deane, 2015). We ensure no overlap in sequences with respect to those in the viscosity benchmark, and use a 80/20 split for training and testing (Figure 1B, Appendix A.3).

Regressor Parameterization and Regularization: We train a multi-task partial deep ensemble with four ensemble components to predict APBS electrostatics and SAP hydrophobicity directly from Fv sequence inputs (Figure 1B). We enable property prediction through an exponential family parameterization of the regression model, using canonical parameters $\theta_1 = \mu/\sigma^2$ and $\theta_2 = -1/(2\sigma^2)$ to disentangle mean and variance estimation. This parameterization better captures uncertainty in biophysical predictions, compared to standard mean and log-variance outputs. We also implement label smoothing, which interpolates between observed measurements and a prior distribution with variance σ^2_{nom} (Klarner et al., 2024). Biophysical property regressors are jointly trained with a diffusion model to learn the distribution of residue-level embeddings, which enables downstream design applications. See Appendices A.3.1, A.3.2, A.3.3 for details.

Guided Sequence Generation by Randomized Greedy Occlusion Search: We use the LaMBO-2 algorithm for multi-property optimization, which generates sequences via discrete diffusion with categorical denoising (Gruver et al., 2023). Generation is directly guided by our sequence-based APBS and SAP regressor, as designs are optimized towards the pareto front of therapeutic-like biophysical properties via the multi-objective noisy expected hypervolume improvement (nEHVI) acquisition function (Daulton et al., 2021). Notably, LaMBO-2 requires an explicit search routine to identify which positions on the sequence to alter due to its encoder-only architecture. Here, we modify LaMBO-2 by introducing randomized greedy occlusion search, which employs iterative masking to efficiently search and evaluate sequence modifications (Figure 1B, Appendix A.3.4).

4 EXPERIMENTS

Here, we consider the relationship between *in-silico* descriptors and experimental liability identification across three viscosity datasets. Antibody viscosity is a clinically important property that is difficult to measure, due to the high material requirements required for assessment. The Ab21 set consists of 21 FDA-approved antibodies (Lai et al., 2021b). The PDGF38 and GCGR datasets

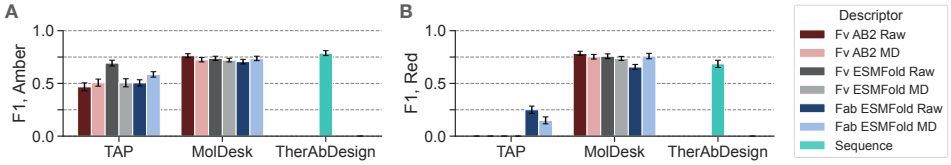


Figure 2: F1 scores of developability filters evaluated on viscosity datasets for (**A**) Amber and (**B**) Red flags (higher is better). Averaging descriptor values over MD simulation frames (light) does not typically improve filters compared to those directly computed from predicted structure (dark). TherAbDesign filters result in similar F1 scores compared to equivalent structural methods.

represent local variants around individual parent molecules (Apgar et al., 2020; Rai et al., 2023). Previous work demonstrates the electrostatic origin of the parent PDGF38 viscosity liability (Apgar et al., 2020). The GCGR dataset include rational designs to target hydrophobicity (Dai et al., 2024). Thresholds used represent the standard limits during manufacturing and drug delivery formulation, respectively (Anselmo et al., 2019; Shire et al., 2004). Roughly half of the sequences across these datasets exceed standard limits for viscosity (Appendix A.2).

4.1 Effectiveness of Descriptors in Identifying Experimental Liabilities

We evaluate how different biophysical definitions affect the agreement between experimental liabilities and developability risk proxies. To account for variance due to small sample sizes, we compute metrics over 100 random subsamples of 80% of the viscosity dataset, reporting average values with standard deviations (Figures 2, 5). Although TAP is the literature standard, we find that the published thresholds have limited utility for filtering viscosity liabilities. F1 scores are zero for the red risk flag across several structure preparation methods, due to high false negative rates (Figure 2, 5). Conversely, APBS and SAP metrics from MolDesk provide valuable experimental signal.

We also find that averaging over MD simulation frames does not globally improve descriptor effectiveness in downstream applications. To further assess the relationships between descriptor definitions, MD simulations, and structure prediction methods, we compare rank correlations between preparation methods (Figure 4). TAP parameters are particularly sensitive to structure preparation, whereas APBS and SAP parameters are more robust (Appendix B.1).

4.2 Sequence-Based Modeling of Biophysical Descriptors

We train TherAbDesign on APBS and SAP definitions because they are less sensitive to structure preparation (Figure 4) and enable better developability filters (Figure 2). Having observed minimal impact of MD from our benchmark, we predict descriptors derived from a static ESMFold-predicted Fab structures directly from a Fv sequence input.

TherAbDesign regressors generalize for antibody sequences, as spearman rank correlations between the regression model outputs and physics-based descriptors are $>$0.8 on held-out pOAS sequences (N=99,118) and the Ab21 global variant dataset (Table 1, Figure 6). Standard deviations are computed over the four partial ensemble components. Correlations are weaker for local variant datasets GCGR and PDGF38, but exceed 0.85 (shown in bold) for the biophysical properties associated with known functional mechanisms (Table 1, Figure 6).

When employed as a filter with MolDesk thresholds, TherAbDesign enables comparable F1 metrics to structure-based approaches (Figure 2). TherAbDesign also offers 3-5 orders of magnitude speed-up relative to computationally intensive physics-based descriptors (Figure 8).

Table 1: Spearman correlation between TherAbDesign predictions and physics-based calculations.

	Fv APBS neg	Fv APBS pos	Fv CAP	Fv SAP BM	Fv SAP WW
pOAS, IID	0.88 $\pm$ 0.03	0.90 $\pm$ 0.01	0.91 $\pm$ 0.03	0.84 $\pm$ 0.04	0.88 $\pm$ 0.02
Ab21	0.86 $\pm$ 0.03	0.83 $\pm$ 0.03	0.85 $\pm$ 0.03	0.89 $\pm$ 0.02	0.84 $\pm$ 0.07
GCGR	0.52 $\pm$ 0.18	0.52 $\pm$ 0.07	0.82 $\pm$ 0.01	**0.86 $\pm$ 0.01**	**0.88 $\pm$ 0.01**
PDGF38	**0.86 $\pm$ 0.03**	**0.88 $\pm$ 0.02**	0.63 $\pm$ 0.08	0.69 $\pm$ 0.07	0.50 $\pm$ 0.05

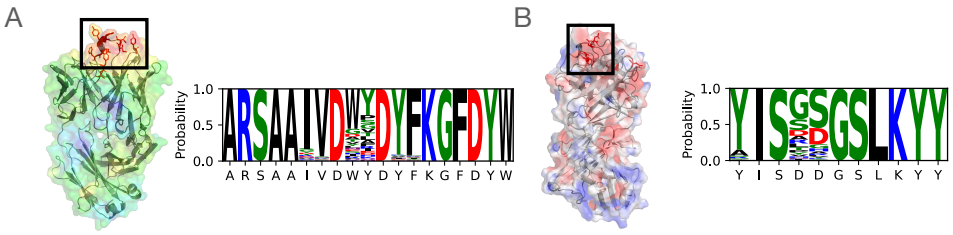


Figure 3: TherAbDesign modifies residues associated with developability risk (boxed, shown as sticks) on the parental molecule. Surfaces colored for GCGR and PDGF38 by (**A**) hydrophobicity and (**B**) electrostatics, respectively, illustrate high-scoring red regions. Sequence logo plots illustrate common design substitutions relative to the parental antibody sequence (x labels).

4.3 Guided design on GCGR and PDGF38 towards therapeutic similarility

Using TherAbDesign, we demonstrate that optimizing antibodies towards therapeutic similarity can address underlying liabilities in sequences containing hydrophobicity (GCGR) and electrostatic (PDGF38) risk flags (Table 7). Our model identified and modified relevant motifs in the complementarity-determining regions (CDRs) by sampling from the learned distribution of sequences seen in pOAS (Figure 7). Likewise, biophysical property prediction with TherAbDesign on these designs show strong correlation with computed SAP and APBS values (Table 8).

For GCGR, TherAbDesign primarily targets aromatic residues around hydrophobic patches in CDRH3 (Figure 3A). 14 sequences overlap with experimental validation data, which consisted of single point mutations from a parental antibody. Of these matches, 12 designs show reduced viscosity compared to the parent sequence, *with 9 falling below the liability threshold*.

TherAbDesign demonstrates similar intuitive behavior for PDGF38 designs, targeting the electronegative patch in the Fv. Our designs introduces compensatory mutations that replace negative and uncharged residues with positive ones, particularly in the "DD" motif of CDRH2 and CDRL2 (Figure 3B, 7D). While our designs did not overlap with the experimental benchmark dataset, this was expected as the benchmark contained more extensive modifications, including framework mutations that are less commonly observed in pOAS.

5 Discussion

TherAbDesign enables scalable evaluation and rational optimization to improve therapeutic-like qualities of antibody candidates. By bringing insights from structure-based developability filtering into a sequence-based design strategy, we introduce a significant conceptual advance that is better aligned with Lipinski's philosophy. As Lipinski (2012) stated, "the rule of five was not intended to be a metric to distinguish drugs from non-drugs; rather, the aim was to help improve the probability of success." As such, our proof of concept demonstrates that optimizing for therapeutic-like qualities allows rational modifications to a parental antibody without requiring explicit mechanistic understanding of its properties. This holistic approach streamlines drug development by reducing reliance on deep domain expertise or bespoke property prediction models.

Ultimately, the efficacy of TherAbDesign relies on how well its *in silico* descriptors correlate with experimental outcomes. Our work highlights a critical challenge in the field: the underlying mechanisms of antibody developability properties remain incompletely understood and lack universal consensus (Jain et al., 2017; Waibl et al., 2022). This challenge is further exacerbated when computational methods are benchmarked against low-fidelity experimental surrogates rather than direct measurements of clinically relevant properties. To address this gap, we recommend expanded mechanistic investigations that better connect computational descriptors with experimental observations. For instance, while our investigation found limited utility in all-atom molecular dynamics with current parameterizations, alternative approaches such as mesoscale simulations with different parameterization strategies may provide valuable insights (Lai et al., 2021a; Prass et al., 2023). By continuing to refine TherAbDesign with more robust biophysical descriptors, we can continue developing approaches to significantly accelerate therapeutic development pipelines.

REFERENCES

Brennan Abanades, Wing Ki Wong, Fergus Boyles, Guy Georges, Alexander Bujotzek, and Charlotte M. Deane. ImmuneBuilder: Deep-Learning models for predicting the structures of immune proteins. *Commun Biol*, 6(1):1–8, May 2023. ISSN 2399-3642. doi: 10.1038/s42003-023-04927-7. URL https://www.nature.com/articles/s42003-023-04927-7. Publisher: Nature Publishing Group.

Lucky Ahmed, Priyanka Gupta, Kyle P. Martin, Justin M. Scheer, Andrew E. Nixon, and Sandeep Kumar. Intrinsic physicochemical profile of marketed antibody-based biotherapeutics. *Proc. Natl. Acad. Sci. U.S.A.*, 118(37):e2020577118, September 2021. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2020577118. URL https://pnas.org/doi/full/10.1073/pnas.2020577118.

Aaron C Anselmo, Yatin Gokarn, and Samir Mitragotri. Non-invasive delivery strategies for biologics. *Nature Reviews Drug Discovery*, 18(1):19–40, 2019.

James R. Apgar, Amy S. P. Tam, Rhady Sorm, Sybille Moesta, Amy C. King, Han Yang, Kerry Kelleher, Denise Murphy, Aaron M. D'Antona, Guoying Yan, Xiaotian Zhong, Linette Rodriguez, Weijun Ma, Darren E. Ferguson, Gregory J. Carven, Eric M. Bennett, and Laura Lin. Modeling and mitigation of high-concentration antibody viscosity through structure-based computer-aided protein design. *PLOS ONE*, 15(5):e0232713, May 2020. ISSN 1932-6203. doi: 10.1371/journal.pone.0232713. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0232713. Publisher: Public Library of Science.

Habib Bashour, Eva Smorodina, Matteo Pariset, Jahn Zhong, Rahmad Akbar, Maria Chernigovskaya, Khang Lê Quý, Igor Snapkow, Puneet Rawat, Konrad Krawczyk, Geir Kjetil Sandve, Jose Gutierrez-Marcos, Daniel Nakhaee-Zadeh Gutierrez, Jan Terje Andersen, and Victor Greiff. Biophysical cartography of the native and human-engineered antibody landscapes quantifies the plasticity of antibody developability. *Commun Biol*, 7(1):1–25, July 2024. ISSN 2399-3642. doi: 10.1038/s42003-024-06561-3. URL https://www.nature.com/articles/s42003-024-06561-3. Publisher: Nature Publishing Group.

Paul J Carter and Arvind Rajpal. Designing antibodies as therapeutics. *Cell*, 185(15):2789–2805, 2022.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11315–11325, 2022.

Jing Dai, Saeed Izadi, Jonathan Zarzar, Patrick Wu, Angela Oh, and Paul J. Carter. Variable domain mutational analysis to probe the molecular mechanisms of high viscosity of an IgG1 antibody. *mAbs*, 16(1):2304282, December 2024. ISSN 1942-0862. doi: 10.1080/19420862.2024.2304282. URL https://doi.org/10.1080/19420862.2024.2304282. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/19420862.2024.2304282.

Samuel Daulton, Maximilian Balandat, and Eytan Bakshy. Parallel Bayesian Optimization of Multiple Noisy Objectives with Expected Hypervolume Improvement, October 2021. URL http://arxiv.org/abs/2105.08195. arXiv:2105.08195 [cs].

James Dunbar and Charlotte M. Deane. Anarci: antigen receptor numbering and receptor classification. *Bioinformatics*, 2, 2015. doi: 10.1093/bioinformatics/btv552. URL https://academic.oup.com/bioinformatics/article/32/2/298/1743894.

Reza Esfandiary, Arun Parupudi, Jose Casas-Finet, Dhanesh Gadre, and Hasige Sathish. Mechanism of Reversible Self-Association of a Monoclonal Antibody: Role of Electrostatic and Hydrophobic Interactions. *Journal of Pharmaceutical Sciences*, 104(2):577–586, February 2015. ISSN 0022-3549. doi: 10.1002/jps.24237. URL https://www.sciencedirect.com/science/article/pii/S0022354915302136.

Nathan C. Frey, Taylor Joren, Aya Ismail, Allen Goodman, Richard Bonneau, Kyunghyun Cho, and Vladimir Gligorijevic. Cramming Protein Language Model Training in 24 GPU Hours. *bioRxiv*, pp. 2024–05, 2024. URL https://www.biorxiv.org/content/10.1101/2024.05.14.594108.abstract. Publisher: Cold Spring Harbor Laboratory.

Nathan C Frey, Isidro Hötzel, Samuel D Stanton, Ryan Kelly, Robert G Alberstein, Emily Makowski, Karolis Martinkus, Daniel Berenberg, Jack Bevers III, Tyler Bryson, et al. Lab-in-the-loop therapeutic antibody design with deep learning. *bioRxiv*, pp. 2025–02, 2025.

Roger B Grosse, Chris J Maddison, and Russ R Salakhutdinov. Annealing between distributions by averaging moments. *Advances in Neural Information Processing Systems*, 26, 2013.

Nate Gruver, Samuel Stanton, Nathan Frey, Tim G. J. Rudner, Isidro Hotzel, Julien Lafrance-Vanasse, Arvind Rajpal, Kyunghyun Cho, and Andrew G. Wilson. Protein Design with Guided Discrete Diffusion. *Advances in Neural Information Processing Systems*, 36:12489–12517, December 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/29591f355702c3f4436991335784b503-Abstract-Conference.html.

Valentin J. Hoerschinger, Franz Waibl, Nancy D. Pomarici, Johannes R. Loeffler, Charlotte M. Deane, Guy Georges, Hubert Kettenberger, Monica L. Fernández-Quintero, and Klaus R. Liedl. PEP-Patch: Electrostatics in Protein–Protein Recognition, Specificity, and Antibody Developability. *J. Chem. Inf. Model.*, 63(22):6964–6971, November 2023. ISSN 1549-9596. doi: 10.1021/acs.jcim.3c01490. URL https://doi.org/10.1021/acs.jcim.3c01490. Publisher: American Chemical Society.

M Holst, N Baker, and F Wang. Adaptive multilevel finite element solution of the poisson-boltzmann equation. *Journal of Computational Chemistry*, 379, October 2000. doi: 10.1002/1096-987X(20001130)21:15⟨1319::AID-JCC1⟩3.0.CO;2-8. URL https://onlinelibrary.wiley.com/doi/full/10.1002/1096-987X%2820001130%2921%3A15%3C1319%3A%3AAID-JCC1%3E3.0.CO%3B2-8?casa_token=GV9yH5l0PWgAAAAA%3ACm3Gxzvh08FtZk2SFmAE-pWB4M3FztJijaS9h_M43Wh_geMifcnZ4wlxfWxvZ0vX0jV-7CqYAqtjdwU.

A. Honegger and A. Plückthun. Yet another numbering scheme for immunoglobulin variable domains: an automatic modeling and analysis tool. *J Mol Biol*, 309(3):657–670, June 2001. ISSN 0022-2836. doi: 10.1006/jmbi.2001.4662.

Aya Abdelsalam Ismail, Tuomas Oikarinen, Amy Wang, Julius Adebayo, Samuel Stanton, Taylor Joren, Joseph Kleinhenz, Allen Goodman, Héctor Corrada Bravo, Kyunghyun Cho, and Nathan C. Frey. Concept Bottleneck Language Models For protein design, December 2024. URL http://arxiv.org/abs/2411.06090. arXiv:2411.06090 [cs].

Tushar Jain, Tingwan Sun, Stéphanie Durand, Amy Hall, Nga Rewa Houston, Juergen H Nett, Beth Sharkey, Beata Bobrowicz, Isabelle Caffry, Yao Yu, et al. Biophysical properties of the clinical-stage antibody landscape. *Proceedings of the National Academy of Sciences*, 114(5):944–949, 2017.

Leo Klarner, Tim GJ Rudner, Garrett M Morris, Charlotte M Deane, and Yee Whye Teh. Context-guided diffusion for out-of-distribution molecular and protein design. *arXiv preprint arXiv:2407.11942*, 2024.

Pin-Kuang Lai, Swan , James W., , and Bernhardt L. Trout. Calculation of therapeutic antibody viscosity with coarse-grained models, hydrodynamic calculations and machine learning-based parameters. *mAbs*, 13(1):1907882, January 2021a. ISSN 1942-0862. doi: 10.1080/19420862.2021.1907882. URL https://doi.org/10.1080/19420862.2021.1907882. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/19420862.2021.1907882.

Pin-Kuang Lai, Amendra Fernando, Theresa K Cloutier, Yatin Gokarn, Jifeng Zhang, Walter Schwenger, Ravi Chari, Cesar Calero-Rubio, and Bernhardt L Trout. Machine learning applied to determine the molecular descriptors responsible for the viscosity behavior of concentrated therapeutic antibodies. *Molecular Pharmaceutics*, 18(3):1167–1175, 2021b.

Bian Li, Luo , Shukun, Wang , Wenhua, Xu , Jiahui, Liu , Dingjiang, Shameem , Mohammed, Mattila , John, Franklin , Matthew C., Hawkins , Peter G., , and Gurinder S. Atwal. PROPERMAB: an integrative framework for in silico prediction of antibody developability using machine learning. *mAbs*, 17(1):2474521, December 2025. ISSN 1942-0862. doi: 10.1080/19420862.2025.

2474521. URL https://doi.org/10.1080/19420862.2025.2474521. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/19420862.2025.2474521.

Giuseppe Licari, Kyle P. Martin, Maureen Crames, Joseph Mozdzierz, Michael S. Marlow, Anne R. Karow-Zwick, Sandeep Kumar, and Joschka Bauer. Embedding Dynamics in Intrinsic Physicochemical Profiles of Market-Stage Antibody-Based Biotherapeutics. *Mol. Pharmaceutics*, 20(2): 1096–1111, February 2023. ISSN 1543-8384. doi: 10.1021/acs.molpharmaceut.2c00838. URL https://doi.org/10.1021/acs.molpharmaceut.2c00838. Publisher: American Chemical Society.

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.

Lipinski. Chris Lipinski. *Nat Rev Drug Discov*, 11(12):900–901, December 2012. ISSN 1474-1784. doi: 10.1038/nrd3895. URL https://www.nature.com/articles/nrd3895. Publisher: Nature Publishing Group.

Christopher A Lipinski. Drug-like properties and the causes of poor solubility and poor permeability. *Journal of pharmacological and toxicological methods*, 44(1):235–249, 2000.

Emily K. Makowski, Hsin-Ting Chen, Tiexin Wang, Lina Wu, Jie Huang, Marissa Mock, Patrick Underhill, Emma Pelegri-O'Day, and Erick Maglalang. Reduction of monoclonal antibody viscosity using interpretable machine learning. *mAbs*, 16(1), 2024. ISSN 1942-0862. doi: 10.1080/19420862.2024.2303781. URL https://doi.org/10.1080/19420862.2024.2303781.

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. ISSN 1469-896X. doi: 10.1002/pro.4205. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/pro.4205. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pro.4205.

Eliott Park and Saeed Izadi. Molecular surface descriptors to predict antibody developability: sensitivity to parameters, structure models, and conformational sampling. *mAbs*, 16(1):2362788, December 2024. ISSN 1942-0862. doi: 10.1080/19420862.2024.2362788. URL https://doi.org/10.1080/19420862.2024.2362788. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/19420862.2024.2362788.

Tobias M. Prass, Patrick Garidel, Michaela Blech, and Lars V. Schäfer. Viscosity prediction of high-concentration antibody solutions with atomistic simulations. *J. Chem. Inf. Model.*, 63(19): 6129–6140, October 2023. ISSN 1549-9596. doi: 10.1021/acs.jcim.3c00947. URL https://doi.org/10.1021/acs.jcim.3c00947. Publisher: American Chemical Society.

Brajesh K Rai, James R Apgar, and Eric M Bennett. Low-data interpretable deep learning prediction of antibody viscosity using a biophysically meaningful representation. *Scientific Reports*, 13(1): 2917, 2023.

Matthew I. J. Raybould, Claire Marks, Konrad Krawczyk, Bruck Taddese, Jaroslaw Nowak, Alan P. Lewis, Alexander Bujotzek, Jiye Shi, and Charlotte M. Deane. Five computational developability guidelines for therapeutic antibody profiling. *Proc. Natl. Acad. Sci. U.S.A.*, 116(10):4025–4030, March 2019. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1810576116. URL https://pnas.org/doi/full/10.1073/pnas.1810576116.

Matthew IJ Raybould, Oliver M. Turnbull, Annabel Suter, Bora Guloglu, and Charlotte M. Deane. Contextualising the developability risk of antibodies with lambda light chains using enhanced therapeutic antibody profiling. *Communications Biology*, 7(1):62, 2024. URL https://www.nature.com/articles/s42003-023-05744-8. Publisher: Nature Publishing Group UK London.

Daniel Russo. Simple bayesian algorithms for best arm identification. In *Conference on Learning Theory*, pp. 1417–1418. PMLR, 2016.

Schrödinger, LLC. The PyMOL molecular graphics system, version 1.8. November 2015.

Vikas K. Sharma, Thomas W. Patapoff, Bruce Kabakoff, Satyan Pai, Eric Hilario, Boyan Zhang, Charlene Li, Oleg Borisov, Robert F. Kelley, Ilya Chorny, Joe Z. Zhou, Ken A. Dill, and Trevor E. Swartz. In silico selection of therapeutic antibodies for development: Viscosity, clearance, and chemical stability. *Proceedings of the National Academy of Sciences*, 111(52):18601–18606, December 2014. doi: 10.1073/pnas.1421779112. URL https://www.pnas.org/doi/10.1073/pnas.1421779112. Publisher: Proceedings of the National Academy of Sciences.

Steven J Shire, Zahra Shahrokh, and JUN Liu. Challenges in the development of high protein concentration formulations. *Journal of pharmaceutical sciences*, 93(6):1390–1402, 2004.

Martin Steinegger and Johannes Söding. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol*, 35(11):1026–1028, November 2017. ISSN 1546-1696. doi: 10.1038/nbt.3988. URL https://www.nature.com/articles/nbt.3988. Publisher: Nature Publishing Group.

Vladimir Voynov, Naresh Chennamsetty, Veysel Kayser, Bernhard Helk, and Bernhardt L Trout. Predictive tools for stabilization of therapeutic proteins. *MAbs*, 1(6):580–582, 2009. ISSN 1942-0862. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2791315/.

Franz Waibl, Monica L. Fernández-Quintero, Anna S. Kamenik, Johannes Kraml, Florian Hofer, Hubert Kettenberger, Guy Georges, and Klaus R. Liedl. Conformational Ensembles of Antibodies Determine Their Hydrophobicity. *Biophysical Journal*, 120(1):143–157, January 2021. ISSN 0006-3495. doi: 10.1016/j.bpj.2020.11.010. URL https://www.sciencedirect.com/science/article/pii/S0006349520308985.

Franz Waibl, Monica L. Fernández-Quintero, Florian S. Wedl, Hubert Kettenberger, Guy Georges, and Klaus R. Liedl. Comparison of hydrophobicity scales for predicting biophysical properties of antibodies. *Front Mol Biosci*, 9:960194, August 2022. ISSN 2296-889X. doi: 10.3389/fmolb.2022.960194. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9475378/.

A Extended Methods

A.1 Full description of biophysical properties

A.1.1 Extended Methods: Structure preparation and molecular dynamics

AbodyBuilder2 (AB2) and ESMFold were used to generate antibody Fv structures (Abanades et al., 2023; Lin et al., 2023). By default, AB2 predicted structures are refined under AMBER force fields using OpenMM. ESMFold was used to generate antibody structures in Fab format, using the Herceptin constant domain sequence. To run MD, we preprocessed structures with tleap. PDB2PQR was used with the AMBER force field FF14SB to assign partial atomic charges. Ionization states were determined using PROPKA at pH 5.5. The system was solvated in an octahedral box of TIP3P water molecules, with a minimum distance of 12 Å between the protein surface and the box edge. Hydrogen partitioning was applied to solute atoms to enable larger time steps of 4 fs. The system was equilibrated for 4 ns and the full production run was performed for 200 ns (Figure 1A).

A.1.2 Therapeutic antibody profiler (TAP)

TAP comprises of custom metrics focused on the surface exposed residues computed over the complementarity-determining regions (CDRs). Thresholds were determined from 754 post-phase-I clinical stage therapeutics. Cutoffs for amber and red risk flags were obtained from the 5% to 10% percentiles.

Table 2: Summary of metrics in TAP

Term	**Description**
CDR length	Loop length has a strong impact on binding affinity.
CDR PSH	Patches of surface hydrophobicity, using the Kyte and Doolittle scale.
CDR PPC	Patches of positive charge.
CDR PNC	Patches of negative charge.
FvSCP	Fv charge symmetry parameter, product of net VH and VL charges

Table 3: TAP risk flag thresholds

Category	**Amber Flag**	**Red Flag**
PSH, CDR	$95.82 \leq \text{PSH} \leq 110.88$	$\text{PSH} < 95.82$
	$168.61 \leq \text{PSH} \leq 204.59$	$\text{PSH} > 204.59$
PPC, CDR	$1.34 \leq \text{PPC} \leq 4.20$	$\text{PPC} > 4.20$
PNC, CDR	$2.02 \leq \text{PNC} \leq 4.43$	$\text{PNC} > 4.43$
SFvCSP	$-35.70 \leq \text{SFvCSP} \leq -6.04$	$\text{SFVCSP} < -35.70$

A.1.3 MolDesk

Park & Izadi (2024) include metrics from APBS and SAP to establish MolDesk thresholds. APBS computes electrostatics over a continuous electrostatic potential (EP), where potential values from a discretized cubic grid are assigned to the mesh vertices. However, MolDesk includes several definitions of SAP hydrophobicity.

Hydrophobicity considers the context of the aqueous environment of a protein, and as such, the interaction with water can be defined in many ways. SAP is computed by weighting an atom's solvent accessible area with respect to the residue's hydrophobicity scale. Some studies have benchmarked hydrophobicity scales with hydrophobic interaction chromatography (HIC), which is typically employed as a proxy assay of developability (Waibl et al., 2022).

Risk regions were calculated from 1D kernel density estimation (KDE) distributions of each descriptors for a dataset of 500 post-phase-I clinical-stage therapeutics from TheraSAbDab. Cutoffs for amber and red risk flags were obtained from the 5% to 10% percentiles of clinical stage antibodies.

Table 4: Summary of selected hydrophobicity scales, reproduced from Waibl et al. (2022)

Term	Principle
Black-Mould	Rekker coefficients, standard SAP definition
Eisenberg	Consensus of five scales
Kyte-Doolittle	Consensus of ΔG (water–vapor) and surface accessibility
Wimley-White	ΔG (water–lipid bilayer)

Table 5: Summary of metrics in Moldesk

Metric	Description
Fv SAP BM	Spatial aggregation propensity with Black-Mould scale
Fv SAP WW	Spatial aggregation propensity with Whimley-White scale
Fv APBS pos	Summed positive electrostatic potential over the Fv
Fv APBS neg	Summed negative electrostatic potential over the Fv
Fv CAP	Charge asymmetry

Table 6: MolDesk risk flag thresholds

Metric	Amber Flag	Red Flag
Fv SAP BM	$x \geq 106.6$	$x \geq 116.52$
Fv SAP WW	$x \geq 68.9$	$x \geq 76.51$
Fv APBS pos	$x \geq 220.17$	$x \geq 230.05$
Fv APBS neg	$x \leq -136.58$	$x \leq -143.09$
Fv CAP	$x \leq -5$	$x \leq -8$

A.2 Description of experimental viscosity datasets

Excessive viscosity limits the injectable dosage of a drug and may require patients to receive treatments at an infusion center. We focus on three public datasets containing experimental viscosities of IgG1 antibodies at high concentration. The Ab21 set consists of 21 FDA-approved antibodies, measured at a concentration of 150 mg/ml in 20mM histidine-HCl pH 6.0 buffer (Lai et al., 2021b). Molecules in this set are distinct, differing by at minimum 20 mutations in their variable domains. The PDGF38 and GCGR datasets represent local variants around individual parent molecules. The PDGF38 set contains 38 antibodies within 9 edits of an anti-PDGF parent, measured at 150 mg/ml in 20mM histidine-sucrose pH 5.8 buffer (Apgar et al., 2020; Rai et al., 2023). Previous work elucidated the electrostatic origin of the extreme parent viscosity liability, 440 cP at 150 mg/ml. The GCGR set contains 37 variants within 2 edits of an anti-GCGR parent, measured at 180 mg/ml in 20mM histidine-acetate pH 5.5 buffer ((Dai et al., 2024)). For these variants, high viscosity is believed to derive from extreme aromaticity in the CDRs.

All original measurements were obtained by cone-and-plate rheometry at 25C. Given that viscosity is highly concentration dependent, we assigned a higher liability threshold for the dataset measured at 180 mg/ml (30 cP) than those measured at 150 mg/ml (20 cP). These thresholds represent the standard limits for ultrafiltration/diafiltration operation during manufacturing (Shire et al., 2004) and formulation for subcutaneous delivery (Anselmo et al., 2019), respectively.

A.3 Model details

Protein Domain Knowledge: Our models are trained on AHo-aligned heavy and light variable domains of IgG antibodies (Honegger & Plückthun, 2001; Dunbar & Deane, 2015). AHo-alignment encodes a strong structural prior mapping certain sequence positions to known structural elements such as complementarity defining regions (CDRs). We disable corruption and sampling of all cysteines in input and output sequences since cysteines (C) play an essential role in determining the overall fold of a protein through the formation of disulfide bridges. We also prevent the model from sampling additional methionines (M) and tryptophans (W) in output sequences since those amino acids are associated with increased risk of oxidation (Sharma et al., 2014).

A.3.1 EXPONENTIAL FAMILY PARAMETERIZATION OF UNIVARIATE GAUSSIAN REGRESSION

We found that when estimating measurement noise along with the mean parameter that the standard deep learning regression parameterization, where the model outputs the mean and log-variance, leads to suboptimal training dynamics and underfitting. We find that exploiting the exponential family canonical parameterization effectively disentangles the gradients of the mean parameter and the variance parameter, leading to much better mean fits while preserving our desire to estimate measurement noise.

To see why this is the case, consider the univariate Gaussian distribution with mean μ and variance σ^2. The standard parameterization of the negative log-likelihood for regression is:

$$\mathcal{L}_{\text{std}}(\mu, \sigma^2; y) = \frac{1}{2}\log(2\pi\sigma^2) + \frac{(y-\mu)^2}{2\sigma^2} \tag{1}$$

The gradients of this loss with respect to the parameters are:

$$\frac{\partial \mathcal{L}_{\text{std}}}{\partial \mu} = -\frac{y-\mu}{\sigma^2} \tag{2}$$

$$\frac{\partial \mathcal{L}_{\text{std}}}{\partial \sigma^2} = \frac{1}{2\sigma^2} - \frac{(y-\mu)^2}{2\sigma^4} \tag{3}$$

Note that the gradient for μ is scaled by $1/\sigma^2$, creating an undesirable coupling between the parameters during optimization. More specifically, the model can easily improve the NLL by inflating the variance term, which in turn diminishes the magnitude of the mean parameter gradient and decreases mean fit quality.

Instead, we can use the exponential family parameterization with canonical parameters $\theta_1 = \mu/\sigma^2$ and $\theta_2 = -1/(2\sigma^2)$. We ensure strict negativity of θ_2 via a softplus transform ($\beta = 0.5$). The negative log-likelihood becomes:

$$\mathcal{L}_{\text{exp}}(\theta_1, \theta_2; y) = -(\theta_1 y + \theta_2 y^2) - \left(-\frac{\theta_1^2}{4\theta_2} - \frac{1}{2}\log(-2\theta_2)\right) - \frac{1}{2}\log(2\pi) \tag{4}$$

where the second term is the log-partition function. The gradients are:

$$\frac{\partial \mathcal{L}_{\text{exp}}}{\partial \theta_1} = -(y - \mathbb{E}_\theta[Y]) \tag{5}$$

$$\frac{\partial \mathcal{L}_{\text{exp}}}{\partial \theta_2} = -(y^2 - \mathbb{E}_\theta[Y^2]) \tag{6}$$

where $\mathbb{E}_\theta[Y] = -\theta_1/(2\theta_2)$ and $\mathbb{E}_\theta[Y^2] = \text{Var}_\theta[Y] + \mathbb{E}_\theta[Y]^2 = -1/(2\theta_2) + \mathbb{E}_\theta[Y]^2$. This parameterization decouples the gradients, as they now depend only on the difference between observed and expected sufficient statistics. The mean and variance can be recovered as transformed model outputs as follows: $\mu = -\theta_1/(2\theta_2)$ and $\sigma^2 = -1/(2\theta_2)$.

A.3.2 GENERALIZING LABEL SMOOTHING TO CONTINUOUS LABELS

For a univariate Gaussian distribution, we perform label smoothing by averaging sufficient statistics rather than directly interpolating distributions. Let $y \sim \mathcal{N}(\mu, \sigma^2)$ be the original label distribution and $z \sim \mathcal{N}(0, \sigma^2_{\text{nom}})$ be the prior with nominal variance σ^2_{nom}.[1] The smoothed distribution is defined by its first two moments:

$$\mathbb{E}[y_{\text{smooth}}] = (1-\alpha_t)\mu + \alpha_t \cdot 0, \tag{7}$$

$$\mathbb{E}[y^2_{\text{smooth}}] = (1-\alpha_t)(\sigma^2 + \mu^2) + \alpha_t \sigma^2_{\text{nom}}, \tag{8}$$

where α_t is the random diffusion corruption fraction sampled at train time. The variance of the smoothed distribution includes a cross-term from the difference in means:

$$\text{Var}[y_{\text{smooth}}] = \mathbb{E}[y^2_{\text{smooth}}] - (\mathbb{E}[y_{\text{smooth}}])^2, \tag{9}$$

$$= (1-\alpha_t)\sigma^2 + \alpha_t \sigma^2_{\text{nom}} + \alpha_t(1-\alpha_t)\mu^2. \tag{10}$$

[1] We assume y has been mean-subtracted and normalized to unit variance during data preprocessing.

This formulation ensures proper uncertainty propagation when interpolating between distributions, unlike geometric averaging of parameters (Grosse et al., 2013). We found that label-smoothing introduced two desirable effects on the model. First, it greatly stabilized multi-task regressor training. Without smoothed labels, weight updates from one task can create large residuals on other tasks, making the regression loss surface highly unstable. Second, we found that training the regressors on noised inputs with smoothed labels heavily regularized the model towards smooth solutions, which is very desirable for exploration and generalization (Klarner et al., 2024).

A.3.3 IMPLEMENTATION

Model Architecture: We implement our model in the `cortex` framework, an open-source package for modular deep learning model composition in PyTorch.[2] Our model is a multi-task partial deep ensemble built on top of a transformer encoder pretrained through a masked language model (MLM) loss (Frey et al., 2024). On top of this encoder we added 4 discriminative convolutional encoders (one for each partial ensemble component) with two residual blocks, convolutional filter width 9, 256 channel dimensions, and layernorm normalization. The representations of each discriminative encoder feed 7 linear regression heads, each modeling a different variant of the biophysical descriptors discussed in this paper, specifically

- **Fv SAP EI, not used in MolDesk filters (Table 6)**
- **Fv SAP KD, not used in MolDesk filters (Table 6)**
- **Fv SAP WW**
- **Fv SAP BM**
- **Fv CAP**
- **Fv APBS neg**
- **Fv APBS pos**

Training dataset: Our sequence-based model was trained on SAP and APBS values derived from pOAS sequences clustered at 95% sequence identity. Sequences were prepended to a Herceptin constant domain sequence, which were subsequentally folded with ESMFold. Malformed structures were identified by structure alignment with a reference herceptin structure predicted with ESMFold, and filtered with a 5Å all-atom RMSD cutoff which removed 0.1% of the labeled data.

Our training dataset is comprised of 892,061 unique Fvs sequences.

Training hyperparameters: Our training procedure followed the multi-task partial deep ensemble training procedure described in (Gruver et al., 2023). In brief, at each gradient step we sample a task head, then sample a minibatch of data from the corresponding dataset and compute the loss and gradients on the corresponding subgraph of the network. Since the task datasets are all different sizes, we dispense with the notion of epochs and instead directly parameterize training in terms of total number of gradient updates. We note that this style of training seems to respond well to higher than typical optimizer momentem hyperparameter values.

- Optimizer: Adam with $\eta_0 = 5 \times 10^{-4}$, $\beta_1 = 0.99$, and $\beta_2 = 0.999$
- Batch size: 128
- Warmup: Linear over first 10% of steps, then cosine annealing to $\eta_\infty = 1 \times 10^{-6}$
- Total gradient update steps: 50K
- Diffusion noise: categorical masking
- Diffusion noise schedule: cosine
- Nominal measurement noise σ^2_{nom}: 6.25×10^{-2}
- Weight decay: 0.0

[2] `https://github.com/prescient-design/cortex`

A.3.4 LaMBO-2 Coordinate Selection by Randomized Greedy Occlusion Search

LaMBO-2 is built on an encoder-only denoising architecture, which grants precise custom control over the number of edits and edit locations at test-time, with a corresponding burden on the program to choose the right combination of edit positions for masking. Gruver et al. (2023) originally proposed input gradients normalized via softmax for coordinate selection, but this approach has inherent limitations for discrete optimization problems. Let $f : \mathcal{V}^L \to \mathbb{R}$ be a neural scoring function and $\nabla_{\mathbf{x}} f(\mathbf{x})$ its gradient with respect to inputs. The gradient provides a local linear approximation of f, but for discrete perturbations such as masking or mutation, the change in $\mathbf{x}$ often exceeds the radius of convergence of the first-order Taylor expansion. Although computing occlusion impact via finite differences $s_j = f(\mathbf{x}'_j) - f(\mathbf{x})$ is more computationally intensive than gradient computation, it directly measures the effect of discrete modifications. The challenge then becomes exploring all possible combinations of k edit positions. Pure greedy selection may fail to discover important position interactions, while exhaustive search is computationally intractable. We propose randomized greedy occlusion search to identify high-impact edit position combinations via iterative masking to balance search efficiency with exploration of position combinations, noting that the theoretical question of whether this coordinate selection problem exhibits submodularity remains open.

Informally, given a sequence length L and target number of edits k our high-level procedure is as follows:

1. For each unselected position $j \in \{1, \ldots, L\}$, evaluate f on the sequence with position j masked.
2. For the positions with highest and second-highest impact scores:
   - Select highest scoring with probability $1 - p_2$.
   - Select second highest with probability p_2.
3. Keep selected position masked and repeat steps 1-2 until k positions are chosen.

The occlusion score s_j has a natural interpretation in terms of the model's learned empirical distribution. When position j is masked, $f(\mathbf{x}'_j)$ represents the expected value under the empirical marginal distribution of tokens at that position, conditioned on the surrounding context. Thus, $s_j > 0$ indicates that the empirical marginal expectation exceeds the value conditioned on the current token, providing direct evidence that position j may be suboptimal (without loss of generality, assume we want to maximize f). The magnitude of s_j quantifies the empirical potential for improvement, while maintaining the masks from previous iterations captures higher-order interactions between positions.

Note on Relation to Prior Work: Denoising position selection by scoring partially occluded examples is a technique used by denoising generative models since at least the introduction of MaskGIT (Chang et al., 2022), however MaskGIT does not use iterative search to find combinations of masks. More study is needed to determine the marginal value of iterative search compared to naive top-k ranking by element-wise occlusion. Our randomization technique is inspired by Top-Two Thompson Sampling (TTTS) (Russo, 2016). We find that setting $p_2 = 0.5$ works well for relatively small values of k, however larger values may be needed to achieve sufficient sample-diversity for larger edit budgets.

Sampling hyperparameters: We sampled designs with a sweep over different hyperparameter parameter values. Hyperparameters that were set to more than one value are denoted as sets, and each run was an element of the full Cartesian product (i.e. grid enumeration)

- Diffusion steps: $T = 2$
- Corruption schedule: Inverse square root $\alpha_t = 1/(1+t)^{0.5}$
- Activation value gradient updates per diffusion step: $\{1, 2\}$
- Guidance regularization λ: $\{0.0, 0.01\}$
- Edit Budget per diffusion step: $B_0 = 2$, $B_t = \lceil \alpha_t \times B_0 \rceil$

Note on Notation Discrepancy with Gruver et al. (2023): Gruver et al. (2023) defines λ as follows:

$$\mathcal{L}(h') = -v_\theta(h') + \lambda \mathrm{KL}[p_\theta(w|h')||p_\theta(w|h)], \tag{11}$$

where h denotes the model activations before the linear token head. We rescale λ to a convex combination,

$$\mathcal{L}(h') = -(1-\lambda)v_\theta(h') + \lambda \mathrm{KL}[p_\theta(w|h')||p_\theta(w|h)], \tag{12}$$

which means we only need to search over $\lambda \in [0, 1]$ to smoothly interpolate between no value guidance ($\lambda = 1$) and no value regularization ($\lambda = 0$).

B Extended Results

B.1 Extended results: Impact of Structure Preparation and Dynamics on Descriptors

To compare the sensitivity of descriptors on the folding models, we compare surface property descriptors computed on structures folded with ABB2 and ESMFold. Given that the liability thresholds in TAP and MolDesk differ in whether they were computed over the Fv or the Fab, we include this as analysis in our descriptor sensitivity assessment. Some reports indicate that a static model may not be adequate for modeling certain surface properties (Waibl et al. (2021); Park & Izadi (2024); Licari et al. (2023)). Therefore, we also compare descriptors computed from raw, unprocessed structures directly from the folding model against an ensemble mean computed over 200 ns simulations to investigate the effect of MD.

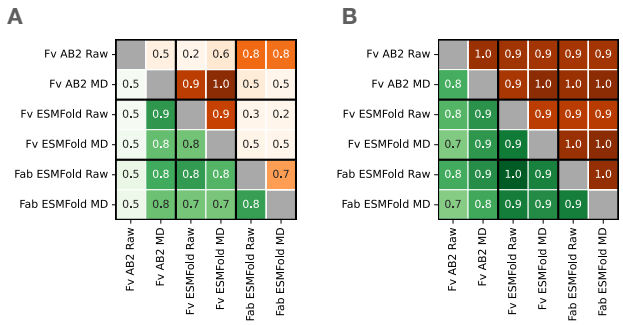


Figure 4: Pair-wise Spearman rank correlation between computed **A.** TAP and **B.** MolDesk descriptors by structure preparation method. Green regions represent surface hydrophobicity (CDR PSH and Fv SAP WW), and orange regions describe positive patches (CDR PPC and Fv APBS pos).

We find that TAP parameters are especially sensitive to structure preparation, which is consistent with published findings (Raybould et al., 2019; Licari et al., 2023). While related studies claim that the primary effect of MD ultimately serves as a correction factor to folding models, we show that it does not compensate for the variability inherent to the surface parameterization. These results are noteworthy, given that the residue-level definitions were originally proposed due to concerns that atomic-resolution descriptors would be too sensitive to use (Raybould et al., 2019). Namely, TAP defines a binary label for each residue for whether the relative accessible solvent exposure of a side chain exceeds 7.5%.

Meanwhile, correlations for APBS and SAP descriptor calculations from MolDesk are substantially less sensitive to structure preparation with rank correlations between 0.7 and 0.9.

B.2 False classification rates of developability filters evaluated on viscosity datasets

The F1 scores in Figure 2 is calculated as the harmonic mean of precision and recall. To better understand the factors contributing to these F1 score values, we discuss the false positive rates (FPR) and false negative rates (FNR). A high FPR can lead to the unnecessary exclusion of viable candidates, while high FNR implies more problematic candidates will be experimentally validated.

We find that MolDesk thresholds are effective for identifying experimental liabilities but may be overly restrictive due to its low FNR but high FPR. In contrast, TAP metrics lead to a low FNR and high FNR. A high FNR might still theoretically streamline experimental throughput, if especially unproductive candidates are filtered. However, we find that TAP metrics computed with a Fv structure input had an FNR of 1, suggesting equivalent outcomes if a filter were not considered. When TherAbDesign is used for filter applications, we observe a much lower FPR but a higher FNR than compared to MolDesk. We note that a low-pass developability filter can be advantageous by including more potentially viable candidates, while excluding the riskiest molecules.

We note that the cutoff values employed in these filters can be somewhat arbitrary, as thresholds may shift depending on the clinical reference set used. Different structure preparation methods may result in a shift in risk flag thresholds. For example, TAP metrics used in this study were derived from a reference set of Fv structures predicted from ABB2. However, preparation of the Fab could alter the positioning of the side chains in the CDRs and lead to different filtering outcomes.

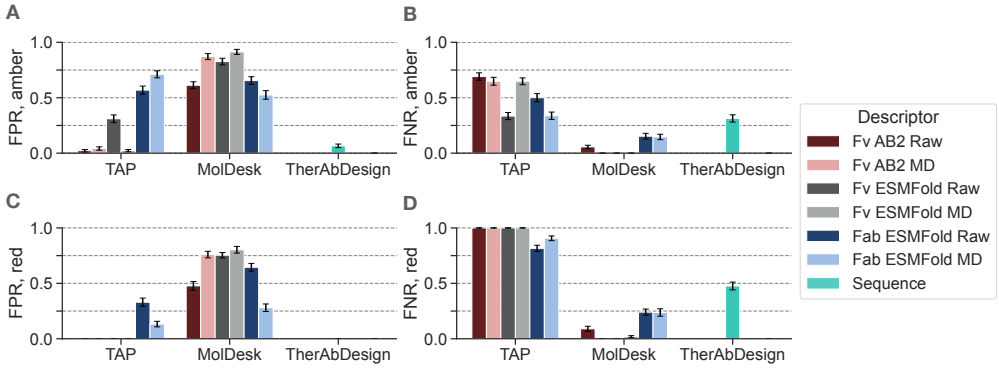


Figure 5: False rates of developability filters evaluated on viscosity datasets (lower is better). TherAbDesign filters result in lower FPR (A, C) but higher FNR (B, D) compared to MolDesk metrics. Error bars represent standard deviation of metrics computed over 100 random subsamples of 80% of the dataset.

B.3 TherAbDesign regression performance

TherAbDesign predictions tend to underestimate metrics compared to structure-based methods, which explains improved FPR when evaluated as a filter on the benchmark datasets (Figure 5). Spearman correlations are shown where predicted values from TherAbDesign are averaged over the four ensemble values, and compared against APBS and SAP values computed over a ESMFold Fab structure.

B.3.1 APBS and SAP values for PDGF38 and GCGR design tasks

Reported metric values for MolDesk are computed over a static Fab structure predicted with ESMFold. Developbility risk with sequence-based TherAbDesign agrees with that of structure-based approaches. Notably, APBS and SAP values predicted for the parental PDGF38 and GCGR raises a risk flag for electronegativity and hydrophobicity, respectively.

Table 7: Average predicted metrics for parental PDGF38 and GCGR from TherAbDesign agree with structural approaches

	Fv APBS neg	Fv APBS pos	Fv CAP	Fv SAP BM	Fv SAP WW
GCGR, MolDesk	-110.39	167.06	0	**169.11**	**118.58**
GCGR, TherAbDesign	-108.5	180.03	0.84	121.63	**76.94**
PDGF38, MolDesk	**-172.64**	78.58	-4	115.64	60.86
PDGF38, TherAbDesign	**-146.69**	121.72	-3.99	96.92	43.12

Table 8: Correlation between average predicted and APBS/SAP computed values of TherAbDesign generated sequences

	Fv APBS neg	Fv APBS pos	Fv CAP	Fv SAP BM	Fv SAP WW
GCGR	0.71 ± 0.05	0.75 ± 0.05	0.80 ± 0.08	0.71 ± 0.04	0.81 ± 0.02
PDGF38	0.87 ± 0.03	0.87 ± 0.01	0.78 ± 0.02	0.73 ± 0.06	0.50 ± 0.08

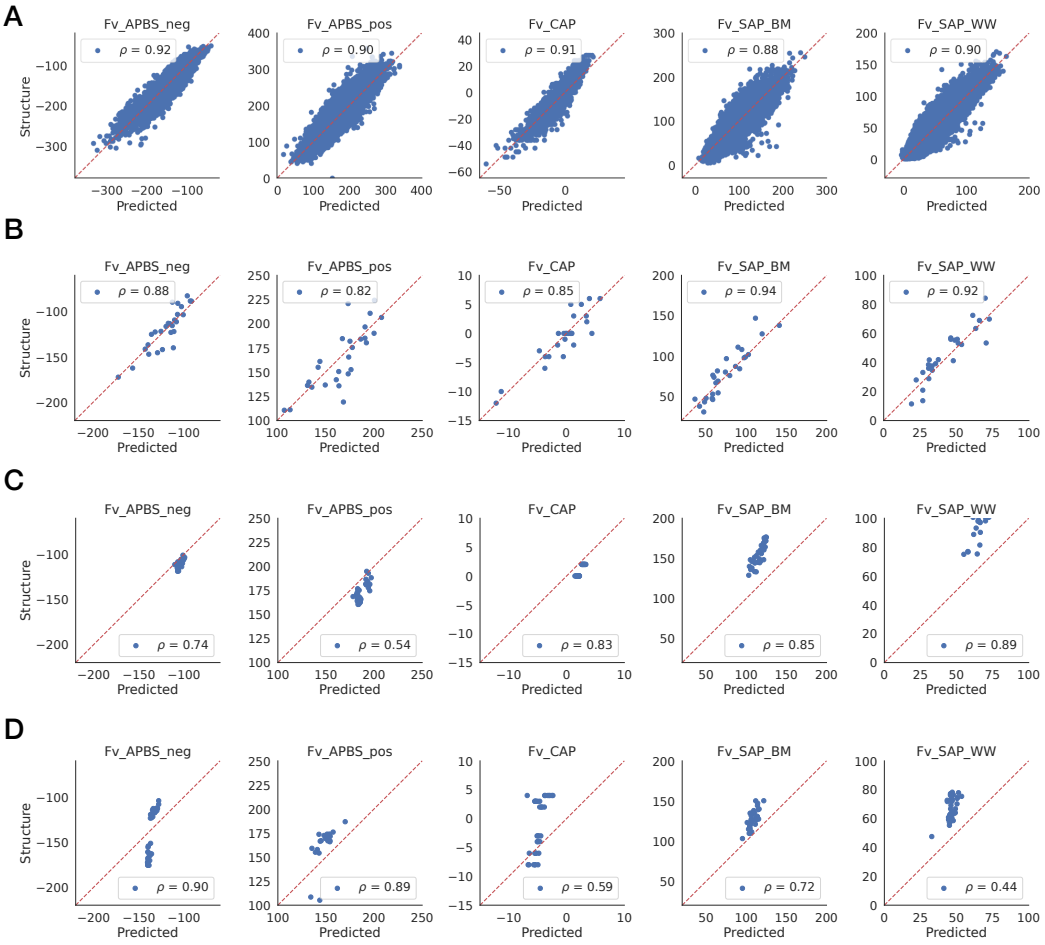
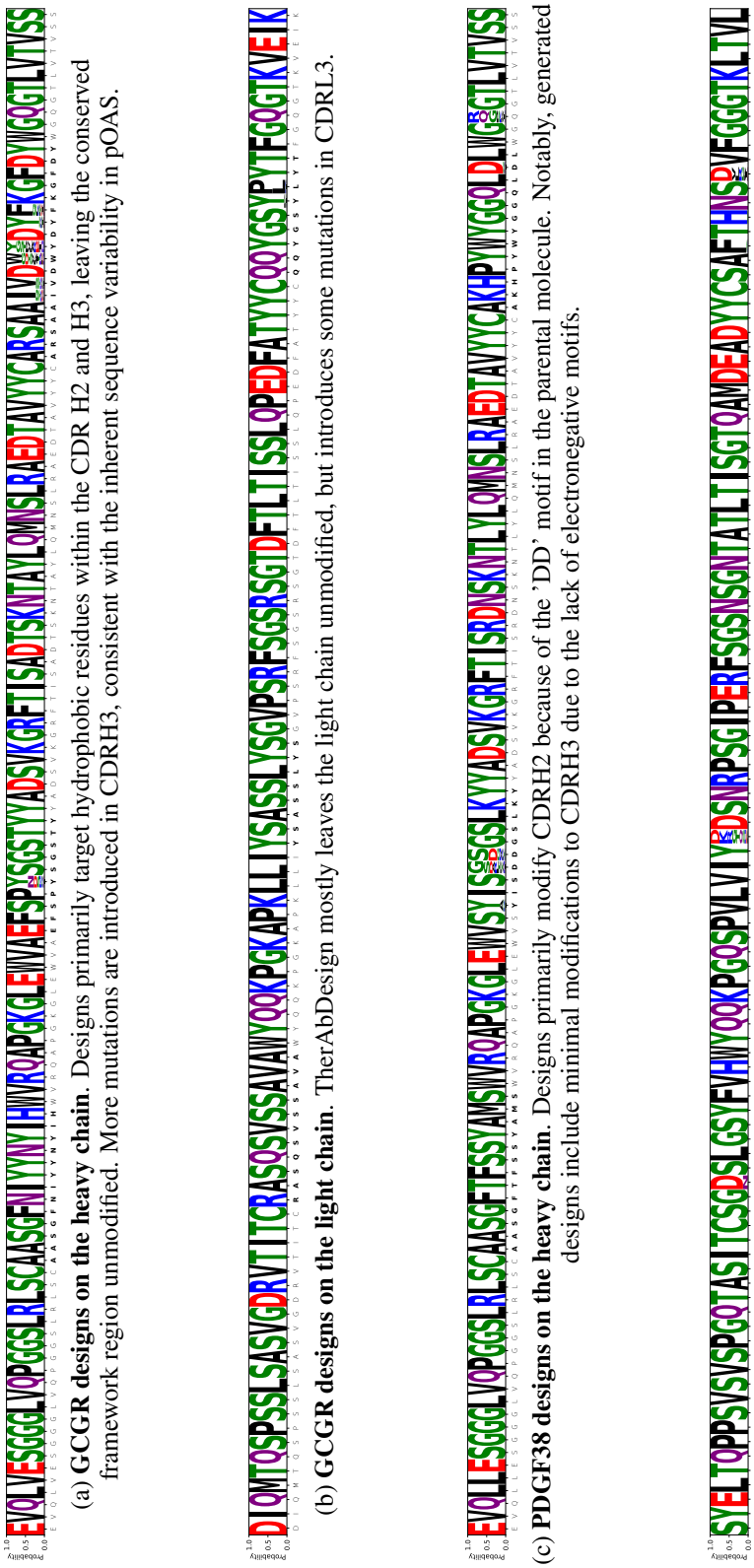


Figure 6: Correlation plots between TherAbDesign regressor outputs and physics-based biophysical properties computed over ESMFold Fab structures. **(A)** pOAS test set. **(B)** Ab21 benchmark set. **(C)** GCGR variants. **(C)** PDGF38 variants.



(a) **GCGR designs on the heavy chain**. Designs primarily target hydrophobic residues within the CDR H2 and H3, leaving the conserved framework region unmodified. More mutations are introduced in CDRH3, consistent with the inherent sequence variability in pOAS.

(b) **GCGR designs on the light chain**. TherAbDesign mostly leaves the light chain unmodified, but introduces some mutations in CDRL3.

(c) **PDGF38 designs on the heavy chain**. Designs primarily modify CDRH2 because of the 'DD' motif in the parental molecule. Notably, generated designs include minimal modifications to CDRH3 due to the lack of electronegative motifs.

(d) **PDGF38 designs on the light chain**. Designs primarily modify CDRL2 which also includes a 'DD' motif.

Figure 7: Sequence logo plots indicate design substitutions relative to the parental antibody sequence (x labels). CDR regions are indicated in bold font.

B.3.2 COMPUTATIONAL EFFICIENCY

Table 9: Computational performance comparison across different platforms for an individual sample. Standard deviations are computed over 100 independent runs.

Computation	**Processor**	**MolDesk**	**TAP**	**TherAbDesign**
ESMFold structure prediction	NVIDIA A100	8.75 ± 0.71 s	8.75 ± 0.71 s	None
Electrostatics	Xeon 9242	178.14 ± 13.15 s	0.18 ± 0.05 s	None
Hydrophobicity	Xeon 9242	27.01 ± 1.99 s	0.12 ± 0.02 s	None
Sequence alignment	-	0.45 ± 0.05 s	0.45 ± 0.05 s	0.45 ± 0.05 s
Model inference	NVIDIA A100	None	None	0.49 ± 0.20 s
Total		214.35 ± 15.9 s	9.81 ± 0.83 s	0.94 ± 0.25 s

TherAbDesign evaluates biophysical criteria more efficiently than physics-based methods, which involve structure prediction and computation of electrostatic and hydrophobicity over the surface of the folded antibody. Comparisons shown here for physics-based methods consider run times for evaluations over a static structure.

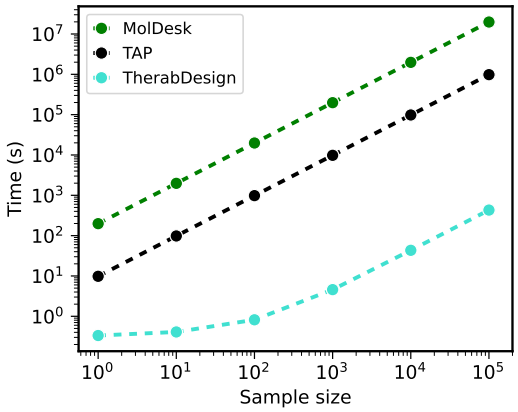


Figure 8: Runtimes for biophysical evaluation by method. TherAbDesign enables a 3-4 fold speed-up compared to MolDesk, and 2-3 speed-up compared to TAP.