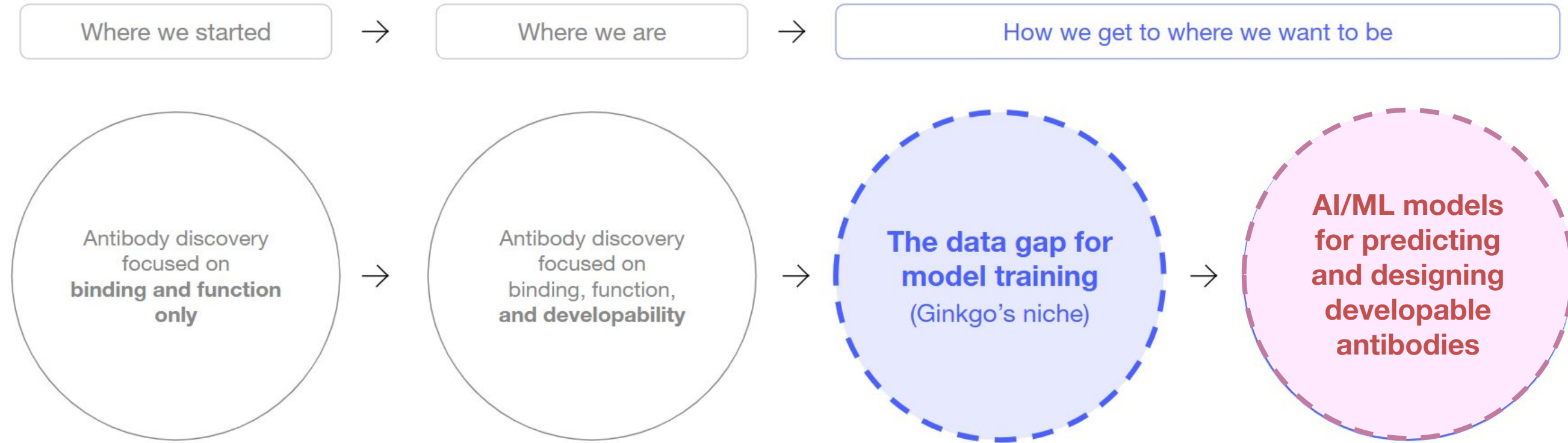


Prediction of antibody developability via Protein Language Models to drive in silico generation

Porfirio Quintero-Cadena, Joshua Moller, Siqi Zhao, Lood van Niekerk, Hao Shen, Seth Ritter
 Ginkgo Bioworks Inc., Boston

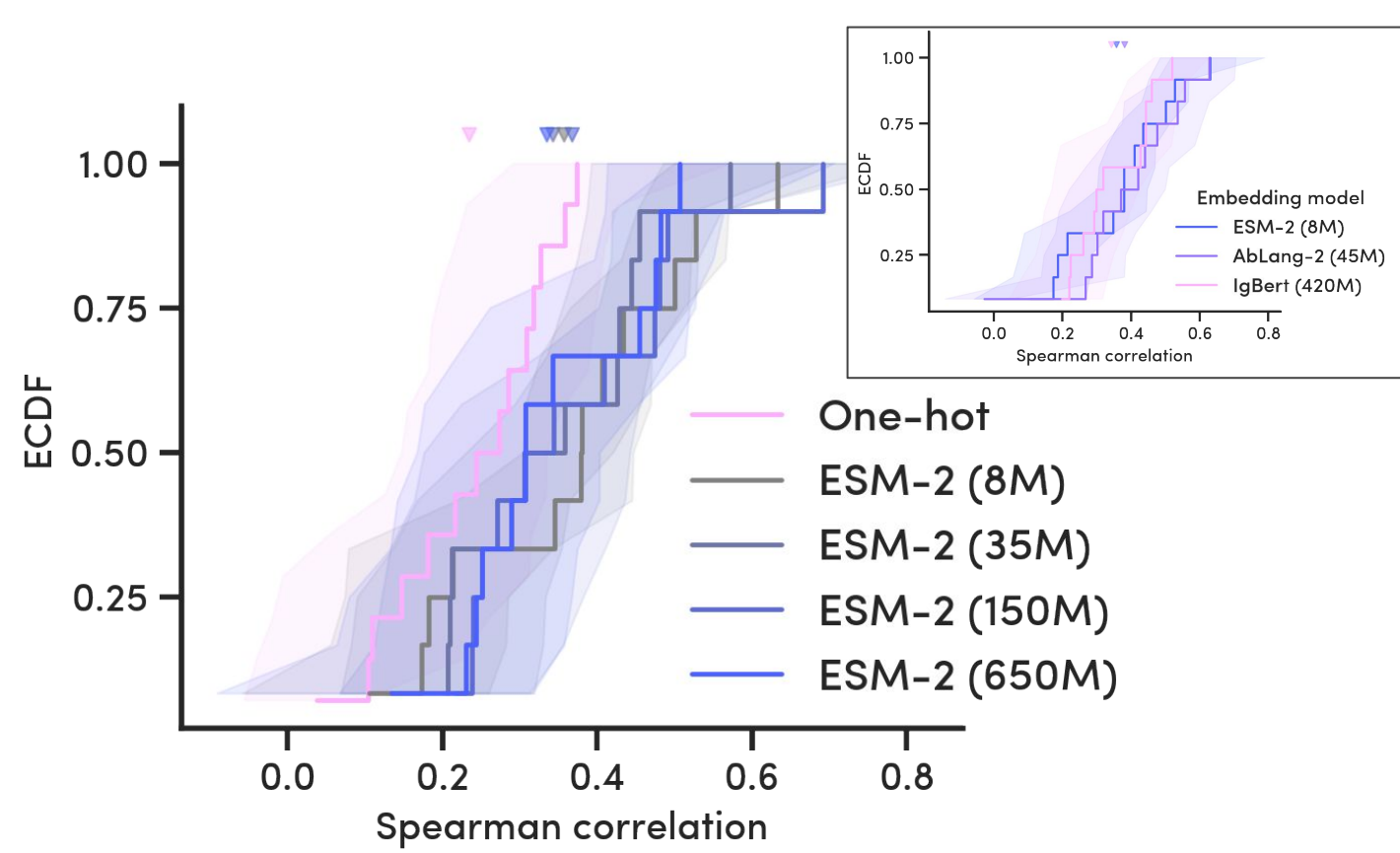
Leveraging High-throughput Antibody Developability Datasets for Machine Learning

Successful antibody therapeutics require not only binding and function but also an acceptable “developability” profile, which encompasses biophysical properties essential for manufacturability, formulation, and in vivo performance. While there have been significant advances in computational modeling and high-throughput screening for binding, the **prediction of developability properties remains a challenge**. This gap is largely due to the scarcity of large-scale, high-quality datasets required to train robust machine learning (ML) models.

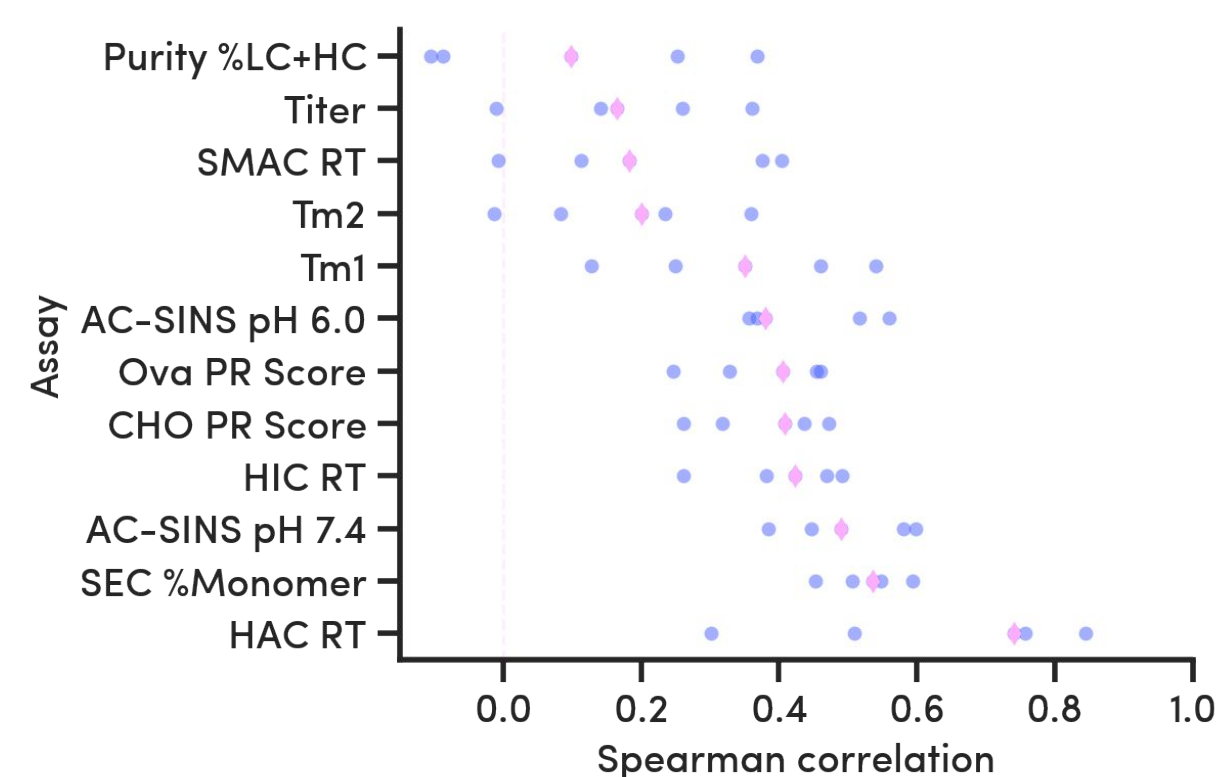


We trained predictive and generative models using an early dataset from our high-throughput (HT) antibody developability platform, showing that we can predict developability from these data, and use these models to guide generative design of developable sequences. Our models can be used as a funnel to select sequences meeting critical developability parameters such as aggregation propensity, thermal stability, hydrophobicity, and self-association, and to improve them along the same axes.

Protein Language Model embeddings yield models predictive of antibody developability assays

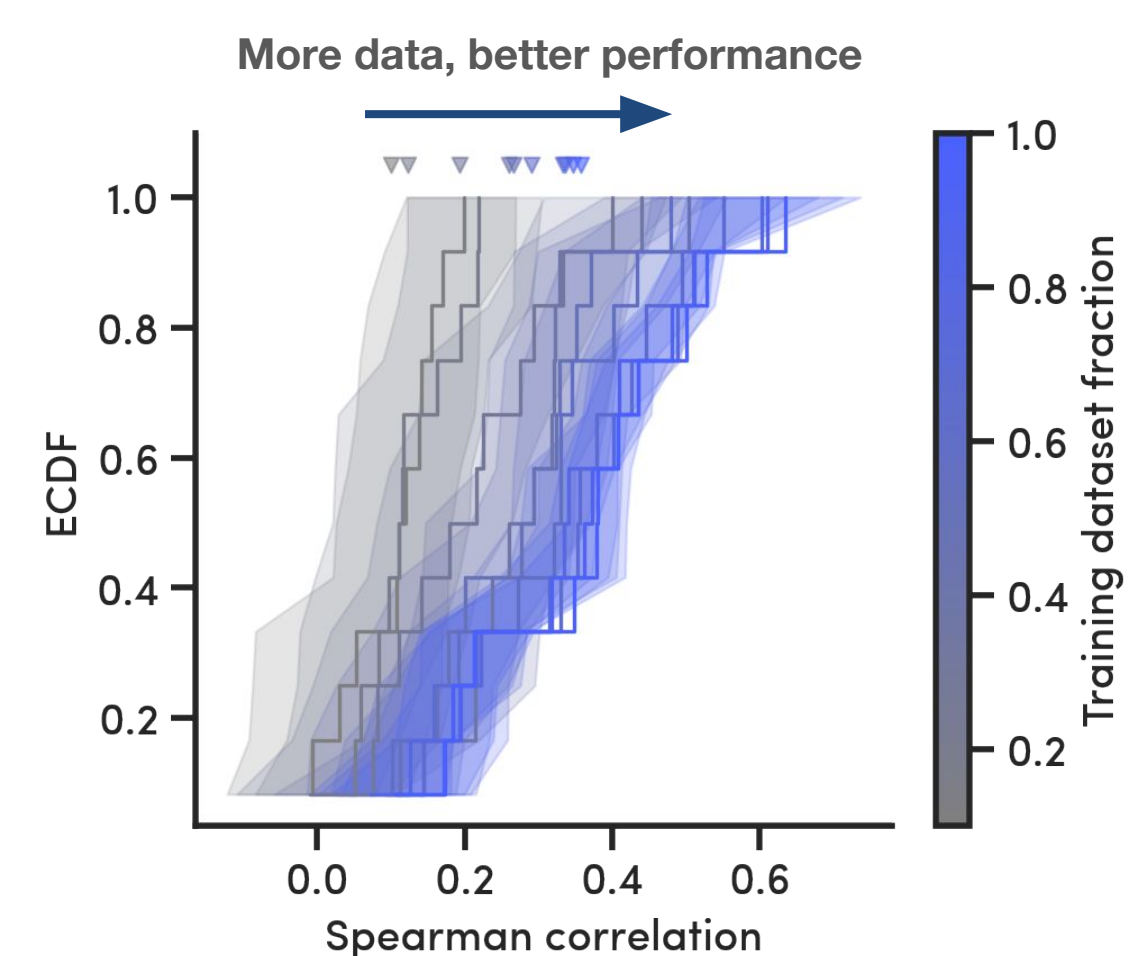


Spearman correlations with ESM-2 (8M), assessed with 5-fold cross validation, are shown below. pLM embeddings provide a positive baseline for the prediction of all developability assays.



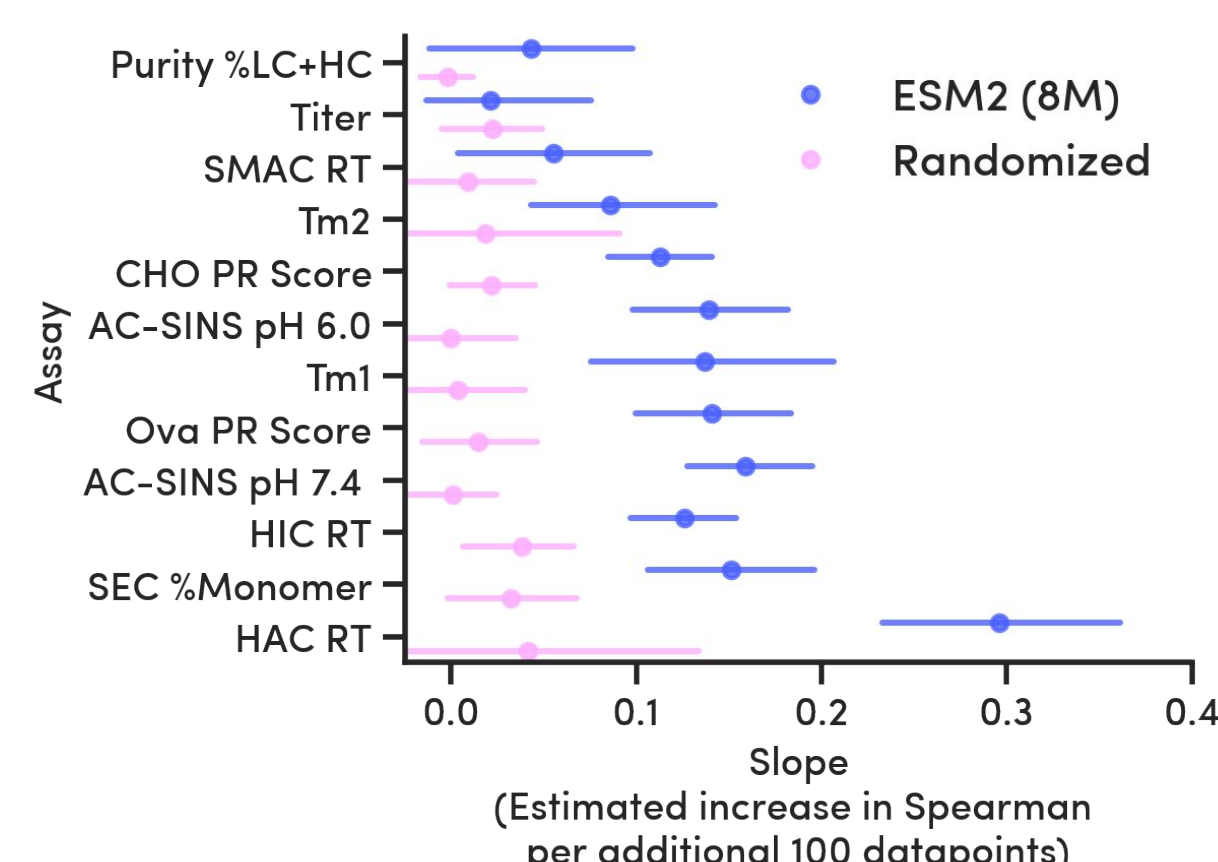
We generated sequence embeddings using ESM-2 (Lin et al. 2023) protein Language Models (pLMs), with 8 million (M) to 650M model parameters. pLM embeddings outperformed the one-hot encoded baseline, but the number of parameters did not meaningfully nor consistently improve performance distributions. Antibody pre-trained pLM (Inset; AbLang-2 from Olsen et al. 2024; IgBert from Kenlay et al. 2024) embeddings did not improve performance either.

Predictive Models Improve with More Training Data

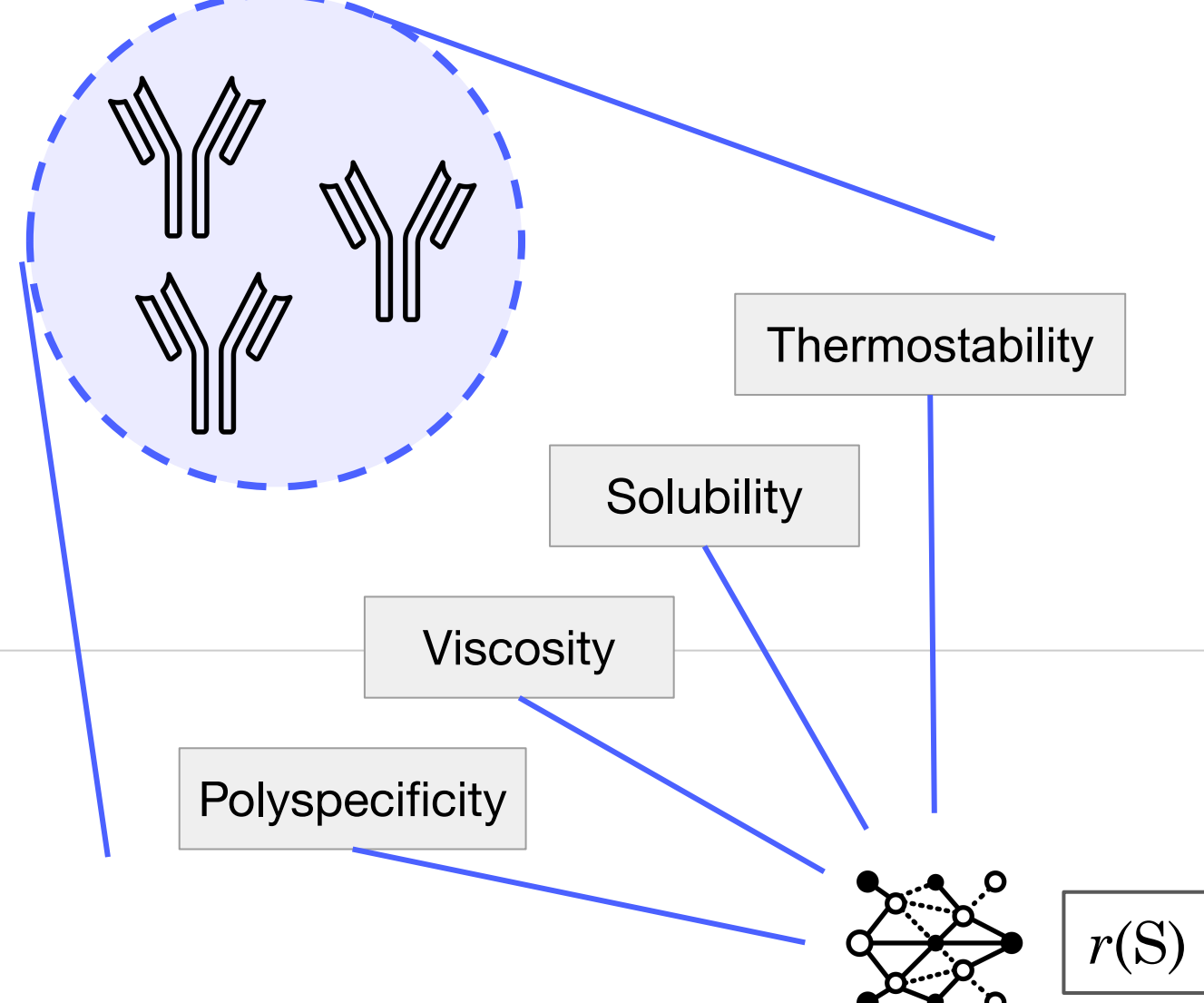


To assess how the amount of training data influences predictive performance, we simulated smaller training datasets by sampling. The distribution of Spearman correlations consistently increased with dataset size

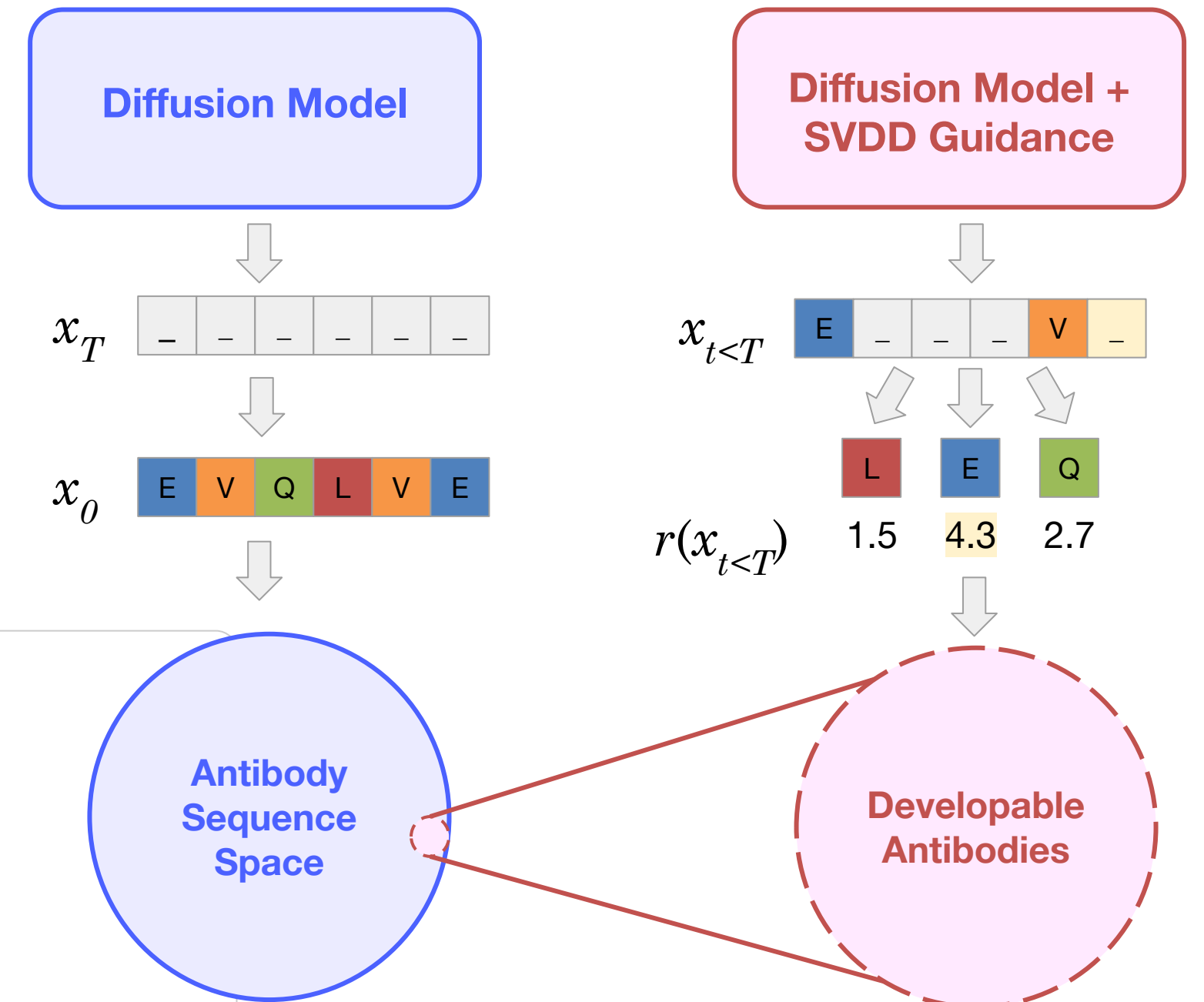
We quantified this increase by fitting linear models to these correlations. For simplicity, we expressed their coefficients as the gain in Spearman correlation per additional 100 data points. These estimates support that additional data, around the order of magnitude generated in this work, leads to meaningful improvements in predictive performance.



Prediction: training model(s) to predict functionality from lab data using sequence and structural information

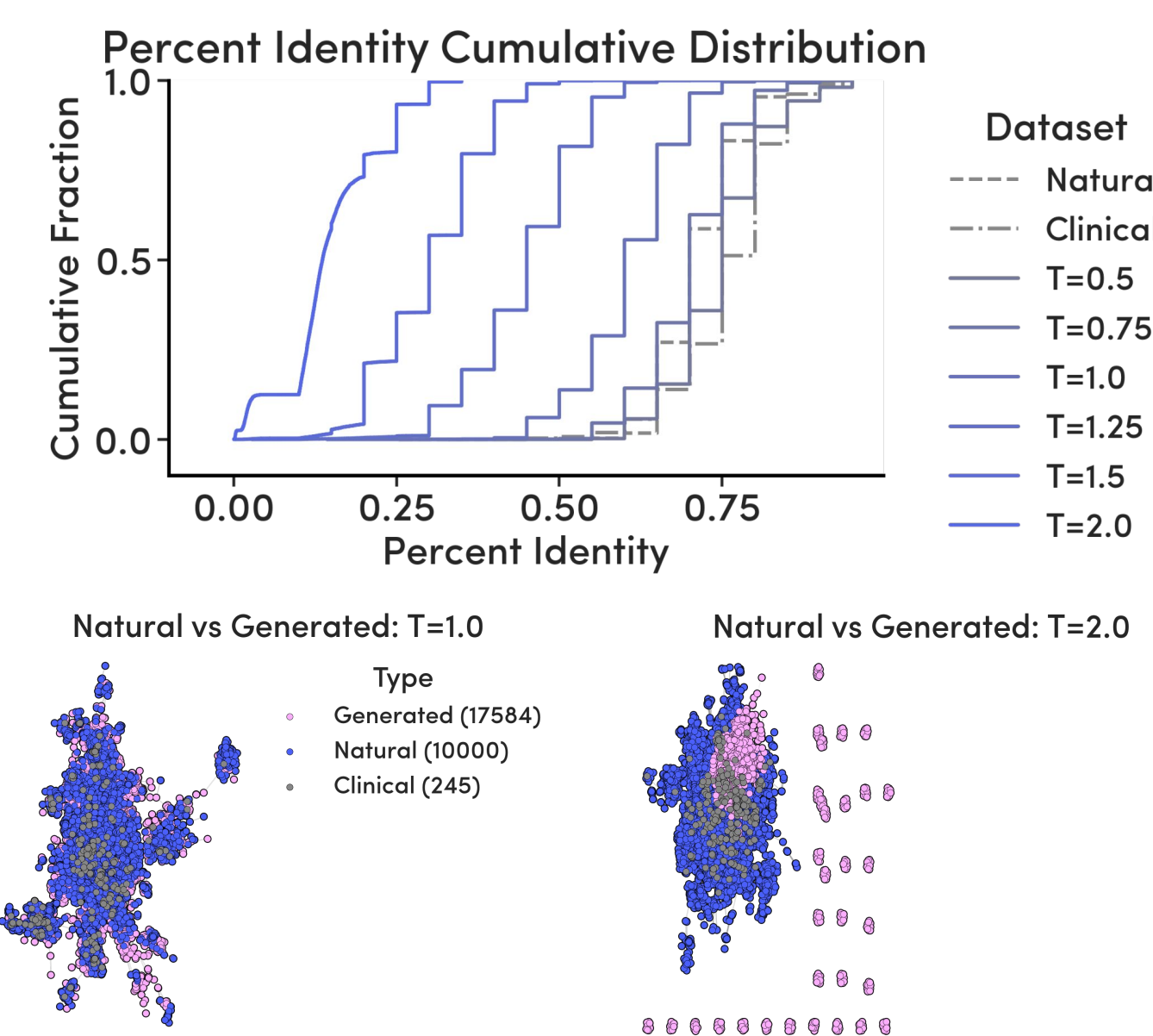


Generation: Discrete Diffusion Models with guidance pushes generation towards developable antibodies

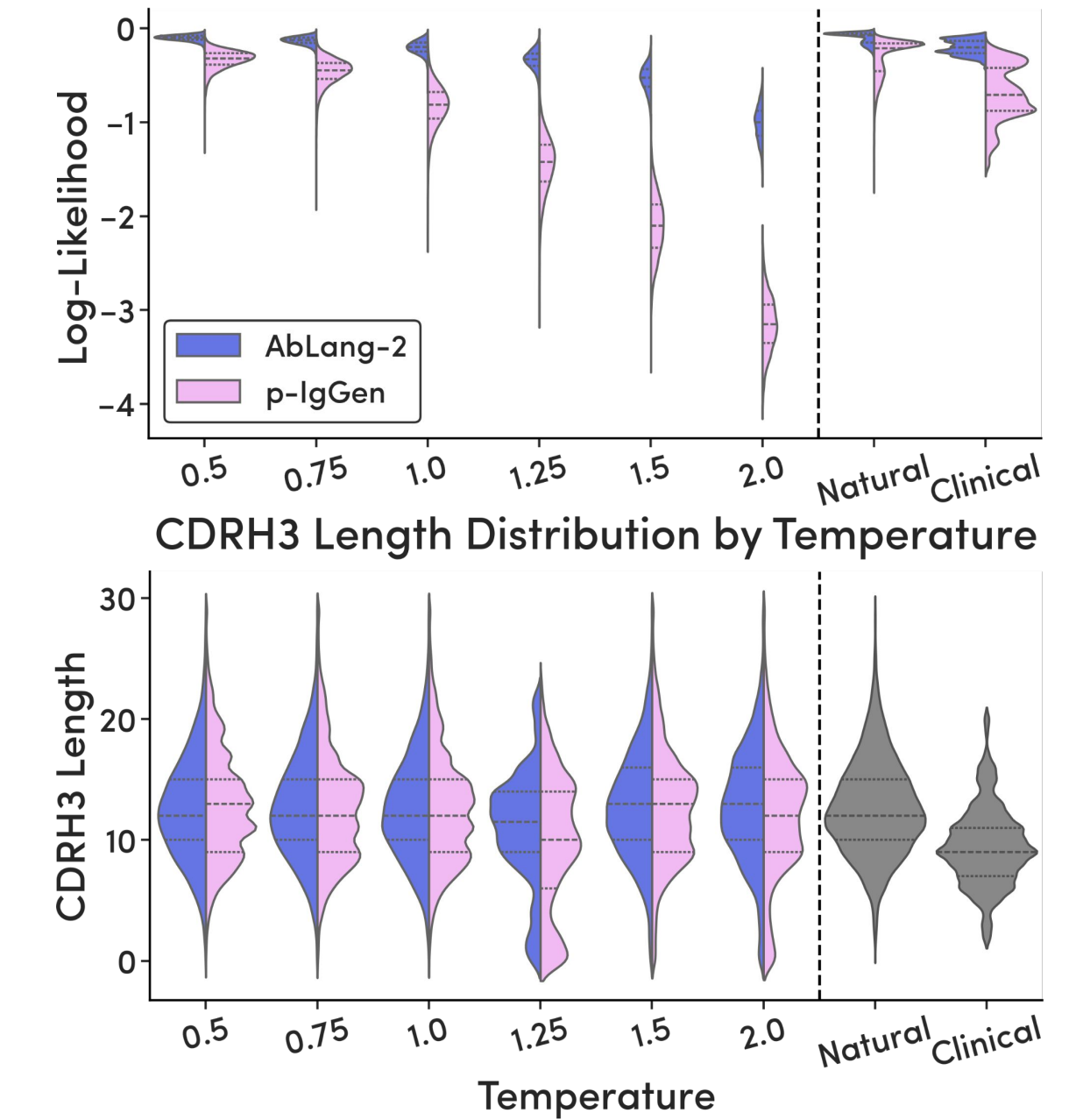


Unconditional generation of sequences yields tunable parameters for generating plausible and diverse sequences

We generated thousands of sequences varying the temperature and the generation strategy from our **in-house paired discrete diffusion generative antibody model**. We evaluate the diversity of the generated sequences by temperature using the percent identity cumulative distribution plots. As temperature increases, sequence similarity within each set decreases showing an increase in diversity. The resultant sequence similarity networks highlight the closeness of these diverse sequences to the naturals.



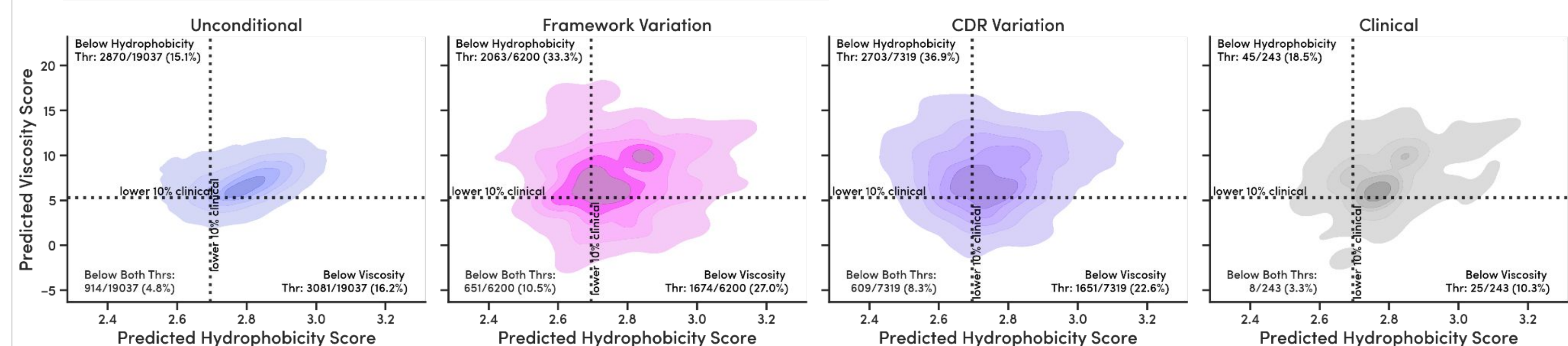
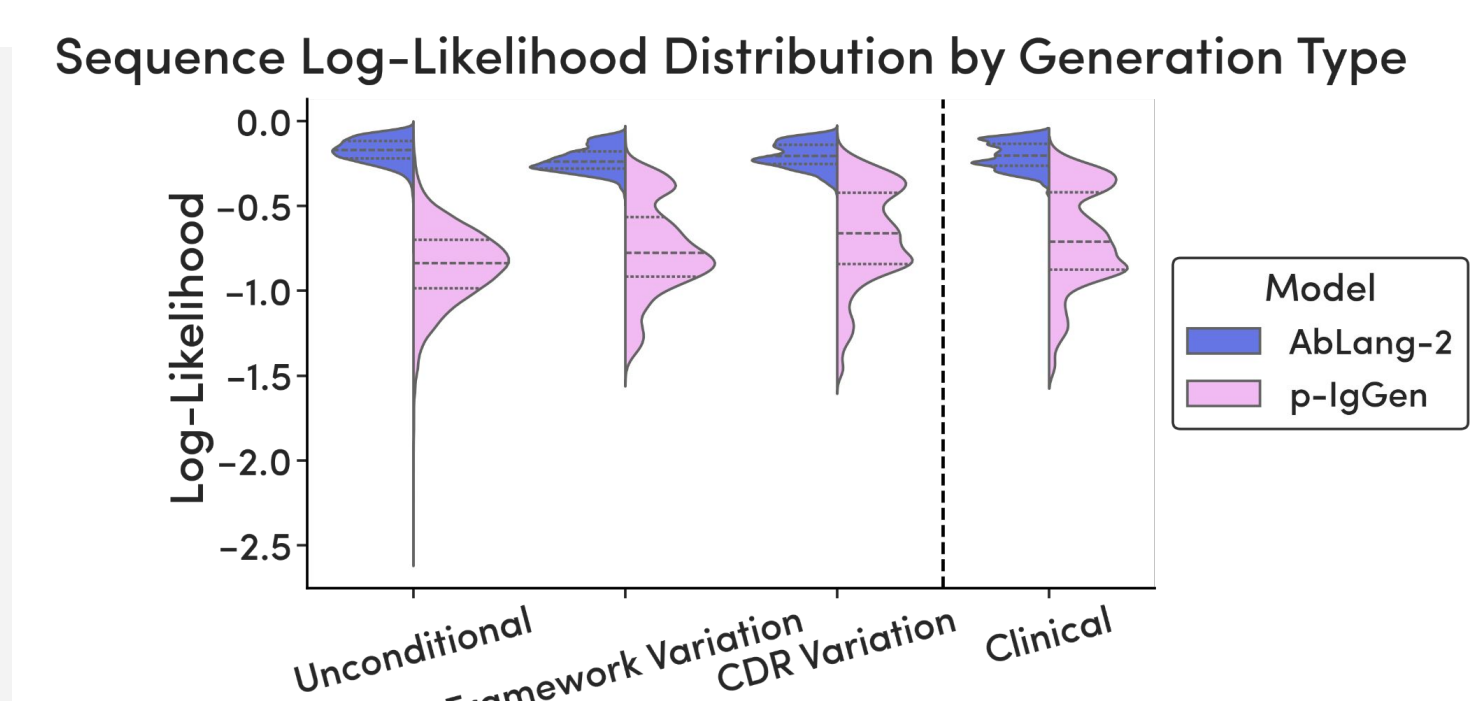
Sequence Log-Likelihood Distribution by Temperature



Increasing temperature of sampling from a discrete diffusion model yields lower log-likelihood probabilities from two external models (AbLang-2 and p-IgGen). We can see that despite these changes, the CDR3 length distributions on the heavy chains stay predominantly the same. The antibodies from our internal dataset show closest similarity to a sampling temperature of 1.0

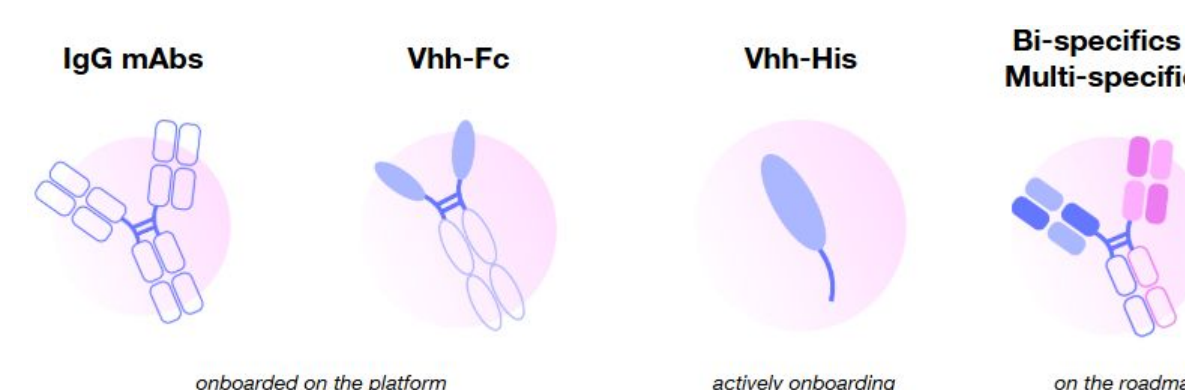
Sequence generations can be guided towards functionality of interest combining our predictive and generative workflows

We generated thousands of sequences with our in-house paired discrete diffusion model (temperature = 1.0), guided by **predictive models trained on internal developability measurements**. Specifically, we steered generation using both hydrophobicity (HIC) and viscosity (AC-SINS pH 7.4) predictors via the Soft Value-based Decoding in Diffusion Models (SVDD) method. By masking the CDR3 region of fixed natural sequences and partially masking the framework, our approach proposed novel sequences that are both natural (right) and possess the desired properties (below).

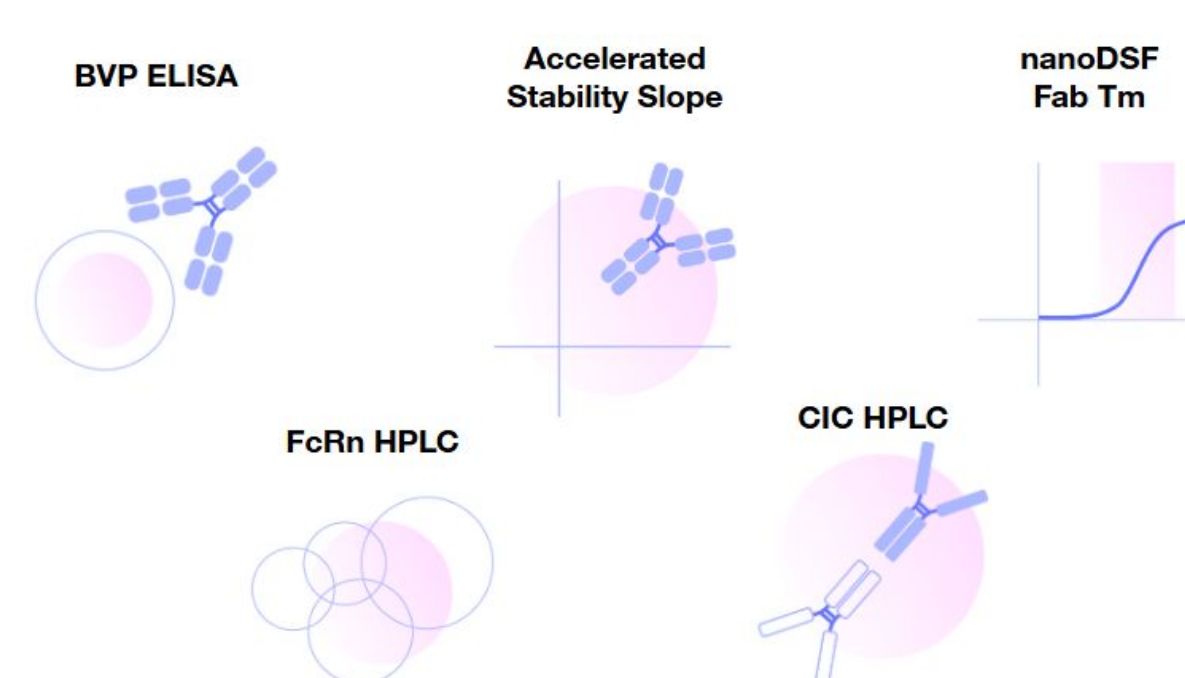


Conclusions and Future Directions

We work with a wide range of antibody formats, and growing



Actively onboarding more developability assays to the platform



In partnership with our high-throughput wet-lab infrastructure, we demonstrate how machine learning (ML) can leverage data from 250 clinical antibodies to extract meaningful insights into antibody developability. This work highlights the transformative potential of ML to not only predict key properties from limited experimental data, but also guide the in silico design of antibodies with optimized biophysical profiles. As data availability grows, these models will become increasingly powerful tools for accelerating discovery, derisking development, and shaping the future of antibody engineering.

- Our platform is rapidly expanding to support a broader range of antibody formats and next-generation developability assays.
- Our expert AI/ML team offers end-to-end support for model development, tailored to your specific pipeline needs.
- As we scale our platform and datasets, our ML models become even more predictive—unlocking faster, smarter, and more confident antibody design decisions.

For technical inquiries, please contact:
Ammar Arsiwala
 Ginkgo Bioworks, Inc.
 Email: aarsiwala@ginkgobioworks.com

Generate Data for Your ML Models



Antibody Developability Modeling for Sequence Selection and Iterative Design

Ginkgo's antibody developability offerings provides a streamlined, fee-for-service model where customers can send antibody sequences for synthesis, expression, purification, and testing across a comprehensive panel of biophysical assays with a readout report format that is AI/ML compatible.

We can use these data to build predictive models to select the most developable sequences, and pair these oracles with generative models to improve the developability of your top candidates.