MABS
2026, VOL. 18, NO. 1, 2634216
https://doi.org/10.1080/19420862.2026.2634216

Taylor & Francis
Taylor & Francis Group

Ginkgo Datapoints Antibody Developability Competition outcomes: limited model performance and a call for data standardization

Lood van Niekerk[a*], Joshua Moller[a*], Seth Ritter[a*], Porfirio Quintero-Cadena[a], Rich Cohen[a], Georgia Channing[b], Michael Chungyuon[c], Laura Rand[a], Alexander Smith[a], Aanal Bhatt[a], Yolaine Pierre[a], Blake Harris[a], Xiang Ao[a], Lucia Grippo[a], Maximilian Schwenk[a], Adam Rosenbaum[a], Olga Allen[a], Nimra Asi[d], Jiang Zhu[e], Aviral Singh[f], Daksh Sammi[f], Rushikesh Jadhav[f], Antonín Dušek[g], Shyam Chandra[h], Valentin Badea[h], Nels Thorsteinson[i], Nathaniel Blalock[j], Jeonghyeon Kim[j], Oliver M. Turnbull[k], Ameya Kulkarni[j], Vivek Kohar[l], Netsanet Gebremedhin[m], Charlotte M. Deane[k], Peter M. Tessier[n], and Ammar Arsiwala[a]

[a]Ginkgo Bioworks Inc, Boston, MA, USA; [b]Hugging Face, New York City, NY, USA; [c]Whiting School of Engineering, Johns Hopkins University, Baltimore, MD, USA; [d]Boehringer Ingelheim, Ridgefield, CT, USA; [e]X UNFOLD LLC, Boston, MA, USA; [f]Microcrispr Pvt. Ltd, Chala, India; [g]Department of Informatics and Chemistry, UCT-Prague, Prague, Czechia; [h]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA; [i]Chemical Computing Group, Montreal, QC, Canada; [j]Biomedical Engineering, Duke University, Durham, NC, USA; [k]Department of Statistics, University of Oxford, Oxford, UK; [l]Unibio Intelligence, Boston, MA, USA; [m]TensorBio Dynamics, Boston, MA, USA; [n]Chemical Engineering, University of Michigan, Ann Arbor, MI, USA

ABSTRACT

The Ginkgo Datapoints Antibody Developability (AbDev) Competition, a blinded benchmark for developability prediction characterized entirely on a single, industrial-scale experimental platform, was conducted from September 8 to November 18, 2025. We benchmarked predictors across five biophysical properties – hydrophobicity, thermostability, self-association, expression titer, and polyreactivity – using a public training set of 246 clinical antibodies and a blinded, held-out test set of 80 antibodies. We received submissions from 113 teams spanning 25 countries, 38 companies, and 39 universities. Winning submissions differed by assay. Top Spearman's ρ values on the test set reached 0.708 (hydrophobicity), 0.392 (thermostability), 0.356 (polyreactivity), 0.337 (self-association), and 0.310 (titer). Cross-validation scores from the public training set consistently exceeded held-out test performance, indicating overfitting and limited out-of-distribution generalization. Together, these results provide a standardized snapshot of current antibody developability modeling capabilities and highlight a key bottleneck: available datasets are too small and heterogeneous to support robust, assay-spanning prediction. Meaningful progress will require larger, standardized, and diverse experimental datasets – with harmonized protocols and rich metadata – to train and validate models that generalize reliably for future antibody discovery campaigns.

ARTICLE HISTORY
Received 3 December 2025
Revised 12 February 2026
Accepted 13 February 2026

KEYWORDS
Antibody; competition; developability

Introduction

Monoclonal antibodies (mAbs) are among the most successful therapeutic modalities, with over 200 marketed products and nearly 1400 candidates in clinical development.[1] While high-affinity binding remains essential for therapeutic efficacy, late-stage clinical and commercial success is strongly governed by "antibody developability." Antibody developability here means the potential for efficient manufacturing and delivery of antibodies at scale. The properties that reflect antibody developability are numerous and costly to fully characterize. In the context of this article, developability is assessed via proxy attributes such as solubility, aggregation propensity, self-association, thermal stability, expression yield, and nonspecific binding.[2] Deficits in these drug-like properties often lead to formulation, manufacturing, and safety challenges, resulting in costly late-stage attrition. Consequently, early, standardized assessment and

CONTACT Joshua Moller jmoller@ginkgobioworks.com; Ammar Arsiwala aarsiwala@ginkgobioworks.com Ginkgo Bioworks Inc, 27 Drydock Ave 8th Floor, Boston, MA 02210, USA
*Equal Authorship.

Supplemental data for this article can be accessed online at https://doi.org/10.1080/19420862.2026.2634216

predictive modeling of developability are now essential complements to traditional binding and function optimization strategies.[2]

Despite the necessity of early developability assessment, the required process changes strain wet-lab resources.[3] Therefore, to streamline the discovery pipeline and reduce overall costs, reliable and generalizable predictive antibody developability models are clearly needed. These models are essential during candidate selection, where a diverse set of sequence clusters must be evaluated. While significant progress has been made with local models that describe mutational effects, global models that rank or directly predict properties of interest still lag.[4] Predictive capability is not uniform; prior reports indicate that properties like hydrophobicity, which arise from relatively low-resolution surface features, are the most readily predicted. From this baseline, complexity increases, requiring progressively larger datasets to support the prediction of properties governed by more intricate underlying mechanisms.[5]

A central challenge in building generalizable, predictive models for developability lies in the limited availability of suitable, high-quality data. Public developability datasets remain constrained by size, assay standardization, and metadata completeness, which ultimately hinders the fidelity and external validation of predictive models. Faster, more meaningful progress requires the release of larger, standardized panels that include raw continuous readouts, comprehensive metadata including assay conditions and construct context, and explicit negatives.[6] Furthermore, blinded, held-out datasets that vary in sequence or germline families are essential to stress-test out-of-distribution robustness – a critical capability for models to add value to real-world drug programs.

Related work

Antibody developability prediction has seen accelerated progress through the use of community benchmarks, which offer a transparent yardstick for comparing baseline models. In the broader protein-design space, resources like TAPE,[7] FLIP,[8] FLOP,[9] ProteinGym,[10] and VenusMutHub [11] have coordinated functional prediction across diverse properties. Recent community challenges, including AIntibody [12] and the Adaptyv Protein Design Competitions,[13] have primarily focused on affinity maturation and binding. While AIntibody characterized developability properties of submitted antibody candidates – including hydrophobicity via hydrophobicity interaction column retention time (HIC RT), polyspecificity via baculovirus particle enzyme-linked immunosorbent assay (BVP ELISA), self-association via affinity-capture self-interaction nanoparticle spectroscopy (AC-SINS), thermal stability (Tm), and thermal aggregation (Tagg) – these metrics were treated as auxiliary properties of interest rather than the primary modeling objective. The field still lacks a benchmark built on a dataset with unified experimental characterization designed to decouple algorithmic performance from inter-laboratory experimental noise.

Some benchmarks for antibody developability exist, such as FLAb,[5,14] NbBench,[15] and DOTAD.[16] However, these often test models across heterogeneous datasets aggregated from varied sources.[2,17–19] This approach, while valuable, can lack the rigor and standardization necessary for principled comparison, making it difficult to delineate model performance from assay variability. A robust, industry-relevant benchmark must therefore address three critical requirements: 1) data splitting reflective of the diverse genotypic and phenotypic shifts relevant to discovery campaigns; 2) emphasis on decision-relevant metrics; and 3) high-quality labeled training data characterized under strictly standardized conditions to ensure that model performance reflects predictive power rather than assay heterogeneity.

To address this critical gap, we introduced the Ginkgo Datapoints Antibody Developability (AbDev) Competition in 2025. Building upon our previous work, the competition utilizes the GDPa1 dataset of 246 unique antibodies,[6] alongside a new, blinded, held-out test set of 80 diverse sequences assayed at Ginkgo Datapoints using our PROPHET-Ab platform (GDPa3). Our core goals are to: 1) release a strengthened set of open, reproducible baseline models; 2) provide a principled evaluation framework tailored to industrial developability prediction; and 3) establish clear data standards to enable the community to generate results that are comparable, interpretable, and practically actionable. Together, these steps position our competition and associated leaderboard as both a snapshot of current performance and a mechanism for advancing data standards in the field of antibody discovery and engineering.

Table 1. The developability assays evaluated in the competition dataset (adapted from [6])

Biophysical or Biological Property Measured	Developability Assay	Readout	Developability Relevance
Expression yield	Titer	mg/L	Composite measure of antibody secretion, folding and correct chain association
Temperature induced domain unfolding	nanoDSF, DSF-SYPRO	Tm2	Indicative of real-time and accelerated storage stability
Surface hydrophobicity	HIC	Retention Time (RT)	High RT indicative of increased risk of aggregation and nonspecific binding
Self-association	AC-SINS	Δλmax	Propensity for aggregation, poor colloidal stability and increased viscosity
Binding to CHO lysate	Polyreactivity	PR Score	Propensity for nonspecific off-target binding to non-antigen species. Potential impact to pharmacokinetics (PK)

Methods

Competition framework

Our competition, in collaboration with Hugging Face, provides a standardized, open evaluation of models that predict antibody developability from sequence.[20] Using the public GDPa1 dataset of 246 antibodies [6] (see reference for experimental methods), competition participants submitted predictions for five prespecified developability properties: hydrophobicity (HIC), polyreactivity (PR-CHO), self-association (AC-SINS at pH 7.4), thermostability (second unfolding transition, Tm2), and expression (Titer). Prior to holdout set predictions, participants could evaluate their model(s) using a predefined 5-fold cross-validation split on the public 246 antibody dataset. This cross-validation procedure served as the primary evaluation framework during the competition's development phase. Each fold was constructed to minimize information leakage by grouping sequences that were 90% identical or higher using hierarchical clustering and separating clusters into different folds while maintaining isotype balance (IgG1/IgG2/IgG4).

Model performance was computed on each held-out fold after training on the remaining four, with the final cross-validation score reported as the mean Spearman's rank correlation coefficient (□) across all five folds. The submissions were scored on a heldout panel of 80 antibodies, which were selected from the paired OAS,[21] sampling for diversity in terms of sequence identity, germline variant, and heavy CDR3 lengths. With the conclusion of the competition, we release the previously held-out test set as GDPa3. We note here that the size of the dataset is small for predictive performance, but still a useful test of standardization practices. Both datasets are freely available to download at https://datapoints.ginkgo.bio/dataset-access. Model performance was evaluated using Spearman's ρ between the models' predicted developability scores and experimentally measured assay values (by Ginkgo Datapoints, Table 1). Each property was evaluated independently with its own leaderboard and award, contributing to a total prize pool of up to $60,000. We also included a dedicated open-source prize for the best fully reproducible entry trained solely on public data. The competition ran for 10 weeks from September 8 to November 18, 2025. A representation of the competition is shown in Figure 1.

Data preparation

To characterize the competition as an evaluation on out-of-distribution genotypic and phenotypic data, we analyzed the sequence and label diversity within the two datasets. The results of this analysis are shown in Figure 2. The pairwise identity distributions (PID) between the datasets are shown in Figure 2(a–d). These percent identities are calculated based on AHO alignment of the antibody sequences using ANARCI.[22] We note the GDPa1 dataset had a full sequence intra-cluster median percent identity of 68.3% (Figure 2(a)) and inter-cluster median of 64.5% (Figure 2(c)). Conversely, the 80 sequence holdout dataset had a median full sequence identity of 59.9% (Figure 2(b)), suggesting slightly higher sequence diversity within the holdout set compared to the provided training data. The median full sequence percent identity between the two datasets was 60.8% (Figure 2(d)), demonstrating genotypic difference between the two datasets of 8 more mutations than within the cross-validation splits.

Additionally, we evaluated the V-gene and J-gene pairs within the two datasets to test if they consist of diverse, unique sets of V-J pairs. This information is represented in Figure 2e,f. There is a substantial coverage of V and J genes for both the heavy and light chains within the two datasets. Genes that are shared between the two

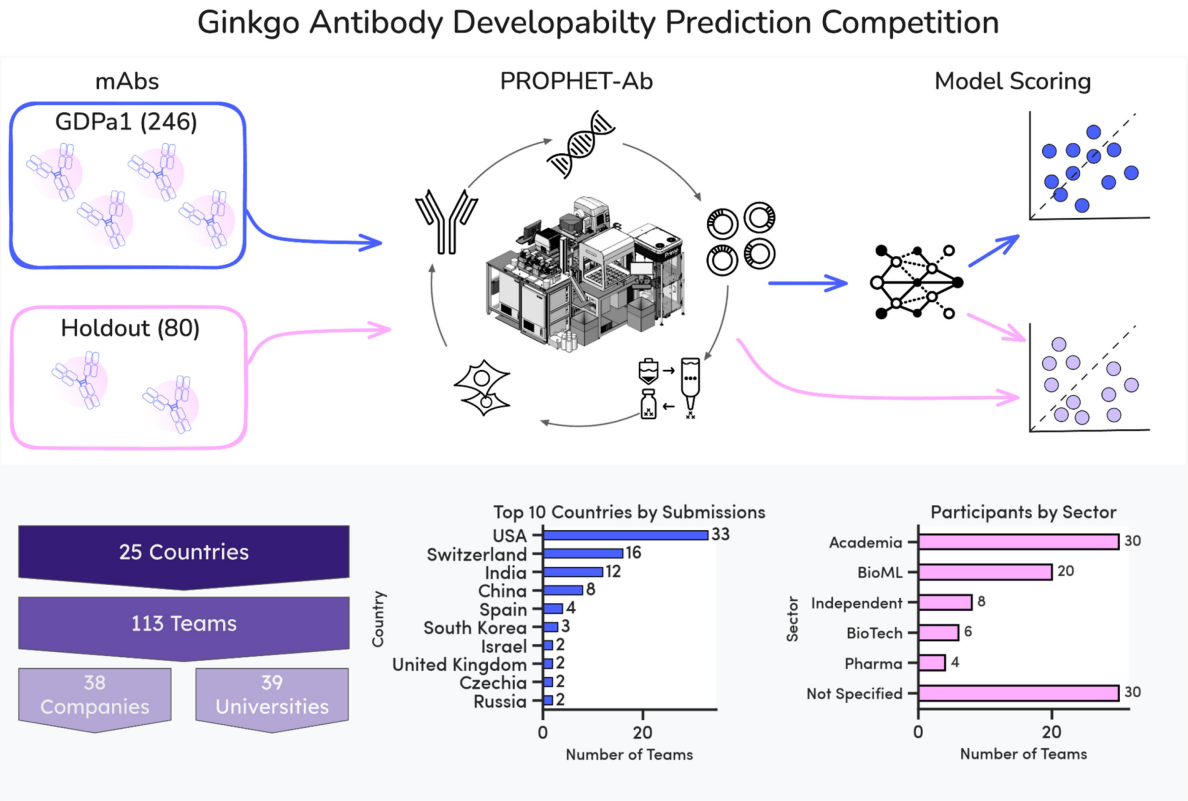

Figure 1. Schematic of the competition setup and participation. Participants were provided with data on 246 antibodies assayed using PROPHET-Ab [6] and submitted predictions evaluated on 80 held-out sequences. Submissions came from 113 teams across 25 countries, 38 companies, and 39 universities. Bar charts show counts for the top 10 countries and sectors as reported by the teams.

datasets as represented by the gray dots. There are more unique permutations as the pairs are different within chains, suggesting the discrepancy in the PIDs. Taken together, these analyses demonstrate the diversity of the datasets.

Although we primarily selected the holdout dataset for sequence diversity from the GDPa1 dataset, it does not guarantee phenotypic diversity. In addition to the sequence diversity, we examine the label diversity in Figure 2 (g). We first evaluated label coverage across the five properties using a pair plot. This plot shows the cumulative distribution function on the diagonal and joint probability density plots for each permutation of properties off diagonal. We note that there is similar coverage across all of these plots for both datasets. To quantify how label similarity and sequence similarity are recapitulated, we provide an inset calculating the centered kernel alignment (CKA). This method calculates the relatedness between two distance matrices, here represented as sequence percent identity and label Euclidean distance. A high relatedness between these two would have a value of 1. We calculated these values between heavy, light, and full sequence chains. For comparison, we calculated a null hypothesis via random permutation of features and labels of full sequences for each dataset. We highlight the 95% confidence interval for these random values in gray. Overall, we find the CKA scores to be low between the labels and sequences for both datasets with a full sequence value of 0.18 for GDPa1 and 0.22 for the GDPa3 holdout set. The GDPa1 dataset is above the random score threshold, but still low, while the GDPa3 holdout set is not.

Despite the separation between sequence and label identity, it is worth mentioning that both training datasets are small for the task at hand, increasing the complexity of this challenge. Additionally, both datasets are imbalanced with few "poor performers," as can be seen in Figure 2(g). Nevertheless, we assumed that the feature set dictating the GDPa1 cross-validation strategy should be commensurate with the difference between GDPa1 and GDPa3.

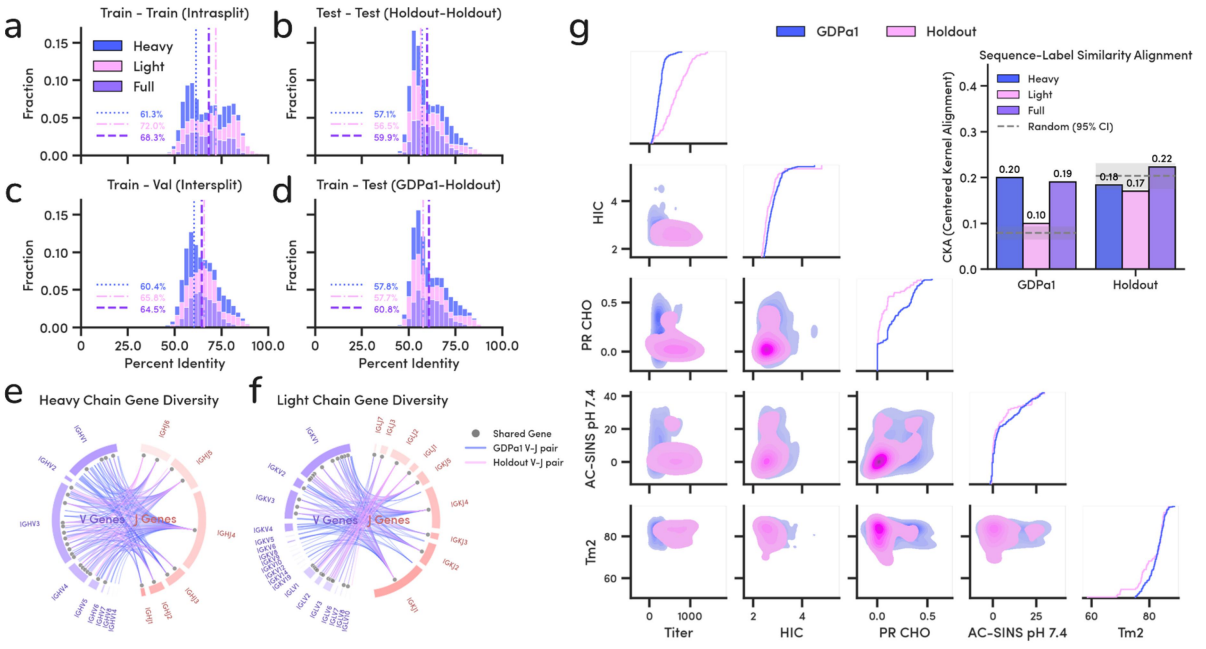

Figure 2. Sequence and label diversity. Pairwise identity distributions show the similarity between the 246 sequences a) within clusters, b) within the holdout test set c) between cluster splits, and d) between the GDPa1 and holdout test sets. Medians of the light, heavy, and full sequences are highlighted by dashed lines. V-gene and J-gene pair chord diagrams between the two datasets. Pairs are shown for both the e) heavy chains and f) light chains. A connection represents the V-J pairs for the specific chain colored by dataset. g) pairwise plot of label distributions between the five assays of interest for the competition. Cumulative distribution functions are shown on the diagonal and joint distributions are shown off diagonal. The inset shows the centered kernel alignment scores between sequence similarity and label similarity. A random permutation of labels is shown as random with a 95% confidence interval for both the GDPa1 and holdout datasets.

Models

Developability predictive models rely on structure- and/or sequence-based features. Sequence-based approaches consist of general or antibody-specific protein language models (PLMs) that generate embeddings or empirically-derived features capturing charge and hydrophobicity distribution within the Fv region. On the other hand, structure-based methods consist of a mixture of inverse-folding models, such as AntiFold [23] or AbMPNN,[24] that encode structural information or compute empirical spatial features, such as surface patches and charge distribution, from predicted structures using, for example, AlphaFold,[25] ABodyBuilder,[26] Boltz,[27] or Chai.[28] A list of pertinent methods and their associated modalities are listed in Table 2.

To contextualize participant performance, we establish several baselines, including three sequence-based models: p-IgGen [31] (Ridge regression on p-IgGen embeddings), ESM2 [29] (Ridge regression on general ESM2 embeddings), and SaProt.[35] We note that performance is highly assay-specific. We evaluate these baselines

Table 2. List of models available as part of associated baselines and their associated modalities.

Model Name	Modality	Reference
ABodyBuilder3	Structure	26
ESM2	Sequence	29
AbLang2	Sequence	30
p-IgGen	Sequence	31
Aggrescan3D	Structure	32
DeepViscosity	Structure/Dynamics	33
DeepSP	Structure/Dynamics	34
AntiFold	Structure	23
SaProt	Sequence/Structure	35
MOE	Structure	36
TAP	Sequence/Structure	37

using the competition training and testing structure, to ensure proper standardization practices. These baselines and more resources are publicly available in our GitHub repository: https://github.com/ginkgobioworks/abdev-benchmark.

Results and discussion

Performance discrepancies and the generalization challenge

As shown in Figure 3, the competition results reveal a significant gap between model performance observed during cross-validation (CV) on the GDPa1 dataset (Figure 3(a)) and performance on the GDPa3 held-out test set (Figure 3(b)) for most properties. This discrepancy highlights a fundamental challenge in antibody developability modeling: generalization to unseen antibody sequences. The gap between performance on GDPa1 and GDPa3 refutes our assumption that the feature set for the cross-validation strategy would apply to the out-of-distribution approach. This suggests that the cross-validation strategy did not appropriately emulate the differences between GDPa1 and GDPa3. For example, the top-performing model for self-association (AC-SINS pH 7.4) achieved a ρ

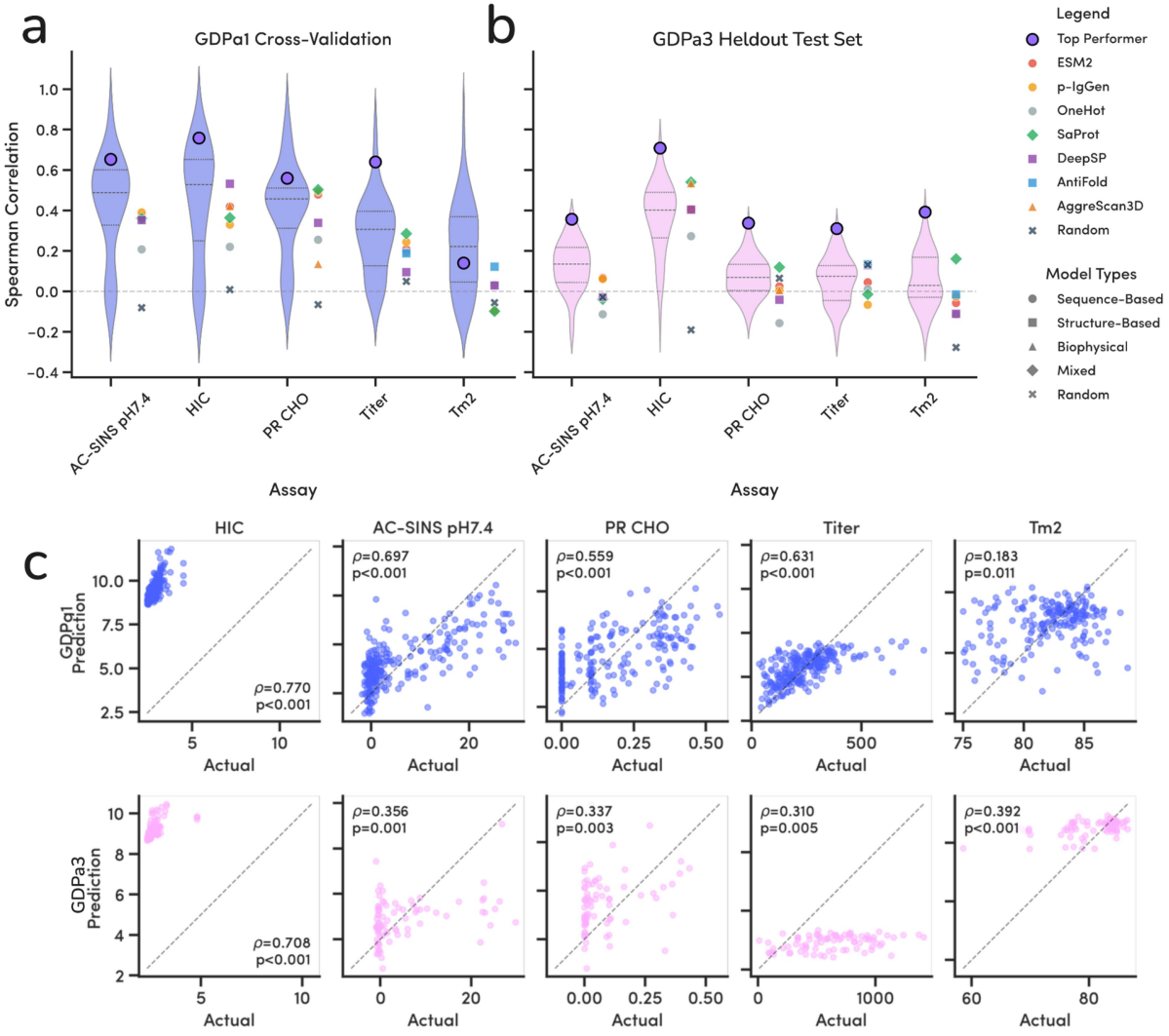

Figure 3. Participant performance in the 2025 Ginkgo datapoints antibody developability prediction competition across five developability properties. a) Blue violin plots show the distribution of all cross-validation submissions; adjacent markers denote out-of-the-box baseline models; top participants per assay are highlighted (purple circle). b) Pink markers show average performance for the hold-out test set. c) Predictions vs. targets for the winning models in each assay: cross-validation (top, blue) and hold-out test (bottom, pink).

Table 3. List of top performers for each developability assay.

Developability Property	Team	Model Name
Titer	Antonín Dušek as a hobbyist from UCT-Prague	meta2
Thermostability	Aviral Singh, Daksh Sammi, and Rushikesh Jadhav from Microcrispr Pvt. Ltd.	microcrisprtm2
HIC	Anonymous	Anonymous
AC-SINS pH = 7.4	Jiang Zhu from X UNFOLD	AbDevelop
PR CHO	Nimra Asi from Boehringer Ingelheim through Data Science Talent LLC	PR_CHO_NOV

of 0.653 during CV but dropped dramatically to a ρ of 0.356 on the held-out test set (Figure 3(c)). Similarly, large performance gaps were observed for polyreactivity (PR CHO: top CV ρ of 0.559 to test ρ of 0.337) and titer (top CV score ρ of 0.640 to test ρ of 0.31). These severe CV-to-test set discrepancies suggest substantial overfitting to training data and underscore that predicting these complex, interaction-dependent properties across a diverse panel of antibodies remains a difficult, out-of-distribution generalization problem with current data limitations and model architectures. Identifying information related to the top competition performers are outlined in Table 3.

In contrast, models for hydrophobicity (HIC) and thermostability (Tm2) exhibited more robust generalization, suggesting that predictions for these properties are more reliable. For hydrophobicity (HIC), the best model exhibited the lowest discrepancy between training and testing, achieving ρ of 0.758 during CV and maintaining a high ρ (0.708) on the held-out test set (Figure 3(c)). For thermostability (Tm2), model performance also showed better correlation on the test set (ρ of 0.392) compared to the training set (ρ of 0.140), which may indicate that the winning model's architecture and training data was inherently better suited to the germline distribution of the held-out panel. These results support the interpretation that hydrophobicity and melting temperature primarily reflect physicochemical features such as local and global folding energetics, side-chain composition, and packing. These features can be reasonably approximated from the structure of a single folded molecule, with solvent and assay context playing a less dominant role and are more amenable to accurate modeling with current data and architectures.

A call for standardization and rigorous validation

The primary lesson of the competition is that developing models for developability properties is less a problem of architectural novelty and more a problem of data and validation rigor. The wide variance across all submissions (Figure 3(b)) and the relatively large differences in training and test performance demonstrate that reliance on cross-validation alone is insufficient and could be misleading. Our findings reinforce the urgent need for rigorous evaluation benchmarks based on carefully curated, standardized data splits that mimic real-world drug discovery, where accurate prediction is necessary on unseen sequences. For antibody development, this highlights the risk of deploying models without validation on a separate, out-of-distribution test set, as the true predictive power of models submitted for the competition remains insufficient for reliable, decision-making fidelity.

Conclusions and future perspectives

The 2025 Ginkgo Datapoints Antibody Developability Competition establishes, to our knowledge, the first multi-property, blinded benchmark for developability prediction built entirely on a single, industrial-scale experimental platform. By anchoring both training and evaluation to PROPHET-Ab, the study isolates model performance from assay heterogeneity and provides an industry-relevant view of how current methods generalize to out-of-distribution antibodies.

The leaderboard (https://huggingface.co/spaces/ginkgo-datapoints/abdev-leaderboard) highlights a clear stratification of model capability. In decreasing order, the Spearman's ρ values started at roughly 0.7 for hydrophobicity and 0.4 for thermostability, and then decreased to 0.3 for self-association, expression titer, and polyreactivity. In particular, the consistent gap between training and held-out test performance

underscores substantial overfitting and the limits imposed by current data scale, diversity, and modeling approaches. Additionally, the hierarchical sequence clustering we used for cross-validation does not capture the feature set differences between the training set and hold-out test set. These observations reinforce a central lesson: meaningful advances in developability prediction will depend on both incremental modeling and featurization changes, but more importantly on standardized, high-throughput data generation.

Looking ahead, the community must move toward integrated affinity and developability predictions, as well as expanding this analysis to multispecific antibodies. The inherent complexity of these next-generation scaffolds requires first understanding predictions for the individual building blocks (e.g., Fabs, VHHs, scFvs). Applying the lessons learned here to this challenging area is paramount for advancing drug discovery and delivering safer therapeutics, faster.

List of abbreviation

OAS	Observed Antibody Space database
HIC	Hydrophobicity interaction chromatography
AC-SINS	Affinity-capture self-interaction nanoparticle spectroscopy
GDPa1	Ginkgo Datapoints antibody dataset 1
GDPa3	Ginkgo Datapoints antibody dataset 3
PID	Pairwise identity distribution
CKA	Centered kernel-alignment
PROPHET-Ab	Platform for Reliable Outcome Prediction in High-throughput Evaluation of Therapeutic Antibodies
BVP ELISA	Baculovirus particle enzyme-linked immunosorbent assay
Tm	Melting temperature
Tagg	Aggregation temperature
CV	Cross-validation
PLM	Protein language model

Acknowledgments

The authors would like to thank all participants in the competition, especially Simon Crouzet and Harshit Singh for their contributions to the writing process. We also thank John Androsavich for the initial discussions on the competition's conception and creating the competition website. Additionally, we would like to thank Holly Lynch for her marketing support for maximum reach. We would also like to thank Julian Englert for providing additional help on extending the reach of the competition.

Author Contributions

L.v.N and J.M. contributed equally to this work. L.v.N set up and ran the competition. J.M. analyzed the results and wrote this manuscript. L.R., A.S., A.B., Y.P., B.H., X.A., L.G., M.S., A.R., and O.A. ran the assays for the test set sequences. N.A., J.Z., A.S., D.S., R.J., and A.D were the winners of the competition. M.C., S.C., V.B., N.B., and N. T. contributed to the public github repo. N.B., J.K., O.M.T., A.K., V.K., N.G, C.M.D., P.M.T., and A.A provided substantial support in writing the manuscript.

Disclosure Statement

P.M.T. and C.M.D are scientific consultants for Ginkgo Bioworks, Inc. All other authors are present or past employees of Ginkgo who funded this work, winners of the competition who received financial compensation, or significant contributors to the public Github repository.

Funding

Ginkgo Bioworks, Inc., funded the work. P.M.T. and C.M.D are scientific consultants for Ginkgo Bioworks, Inc.

ORCID

Porfirio Quintero-Cadena http://orcid.org/0000-0003-0067-5844
Georgia Channing http://orcid.org/0009-0001-6354-7527
Alexander Smith http://orcid.org/0009-0009-3298-4431
Aanal Bhatt http://orcid.org/0000-0002-0534-9274
Xiang Ao http://orcid.org/0009-0002-3565-9473
Lucia Grippo http://orcid.org/0000-0002-0058-9279
Adam Rosenbaum http://orcid.org/0009-0003-2316-477X
Aviral Singh http://orcid.org/0009-0001-4654-1182
Daksh Sammi http://orcid.org/0009-0007-7780-5674
Rushikesh Jadhav http://orcid.org/0009-0007-8553-2311
Antonín Dušek http://orcid.org/0009-0007-5461-7734
Shyam Chandra http://orcid.org/0009-0008-9986-3303
Valentin Badea http://orcid.org/0009-0001-3943-4637
Nels Thorsteinson http://orcid.org/0000-0001-8333-0045
Nathaniel Blalock http://orcid.org/0000-0001-7993-7936
Jeonghyeon Kim http://orcid.org/0009-0001-4468-5035
Oliver M. Turnbull http://orcid.org/0000-0002-0239-1207
Vivek Kohar http://orcid.org/0000-0003-1813-1597
Charlotte M. Deane http://orcid.org/0000-0003-1388-2252
Peter M. Tessier http://orcid.org/0000-0002-3220-007X
Ammar Arsiwala http://orcid.org/0000-0002-5121-9163

References

1. Crescioli S, Kaplon H, Wang L, Visweswaraiah J, Kapoor V, Reichert JM. Antibodies to watch in 2025. mAbs. 2025;17(1). doi: 10.1080/19420862.2024.2443538.
2. Jain T Boland T, Vasquez M. Identifying developability risks for clinical progression of antibodies using high-throughput in vitro and in silico approaches. mAbs. 2023 Apr;18;15(1). doi: 10.1080/19420862.2023.2200540.
3. Zhang W, Wang H, Feng N, Li Y, Gu J, Wang Z. Developability assessment at early-stage discovery to enable development of antibody-derived therapeutics. Antibody Ther. 2023 Jan. 6(1):13–29. doi: 10.1093/abt/tbac029.
4. Michalewicz K, Barahona M, Bravi B. Machine learning approaches for interpretable antibody property prediction using structural data. arXiv. 2025. doi: 10.48550/arXiv.2510.23975.
5. Chungyoun M, Gray J. Fitness landscape for antibodies 2: Benchmarking reveals that protein AI models cannot yet consistently predict developability properties. bioRxiv. 2025 Dec 27. doi: 10.64898/2025.12.27.696706.
6. Arsiwala A, R, van Niekerk L, Quintero-Cadena P, Ao X, Rosenbaum A, Bhatt A, Smith A, Yang Y, Anderson KC, et al. A high-throughput platform for biophysical antibody developability assessment to enable AI/ML model training. mAbs. 2025;17(1). doi: 10.1080/19420862.2025.2593055.
7. Rao R, Bhattacharya N, Thomas N, Duan Y, Chen P, Canny J, Abbeel P, Song Y. Evaluating protein transfer learning with TAPE. Adv Neural Inf Process Syst. 2019;32:9689–9701.
8. Dallago C, Mou J, Johnston KE, Wittmann BJ, Bhattacharya N, Goldman S, Madani A, Yang KK. Flip: Benchmark tasks in fitness landscape inference for proteins. bioRxiv. 2022 Jan 19. doi: 10.1101/2021.11.09.467890.
9. Groth PM, Michael R, Salomon J, Tian P, Boomsma W. Flop: tasks for fitness landscapes of protein wildtypes. bioRxiv. 2023 June 21. doi: 10.1101/2023.06.21.545880.
10. Notin P, Kollasch A, Ritter D, Van Niekerk L, Paul S, Spinner H, Rollins N, Shaw A, Orenbuch R, Weitzman R, et al. Protein gym: large-scale benchmarks for protein design and fitness prediction. bioRxiv. 2023 Dec 8. doi: 10.1101/2023.12.07.570727.
11. Yu J, Li G. VenusMutHub—a benchmark for protein mutation effect prediction. Acta Pharmaceutica Sin B. 2025 May 14. 15(5):2805–2807. doi: 10.1016/j.apsb.2025.05.001.
12. Martin ACR. Aintibody: an experimentally validated in silico antibody discovery design challenge. Nat Biotechnol. 2024;42(11). doi: 10.1038/s41587-024-02469-9.
13. Cotet T-S, Krawczuk I, Pacesa M, Nickel L, Correia BE, Haas N, Qamar A, Challacombe CA, Kidger P, Ferragu C, et al. Crowdsourced protein design: lessons from the Adaptyv EGFR binder competition. 2025 Apr 24. doi: 10.1101/2025.04.17.648362.
14. Chungyoun M, Ruffolo J, Gray J. Flab: benchmarking deep learning methods for antibody fitness prediction. bioRxiv. 2024 Jan 15. doi: 10.1101/2024.01.13.575504.
15. Zhang Y, Tsuda K. NbBench: Benchmarking language models for comprehensive nanobody tasks. arXiv. 2025 May 4. doi: 10.48550/arXiv.2505.02022.

16. Li W, Lin H, Huang Z, Xie S, Zhou Y, Gong R, Jiang Q, Xiang C, Huang J. Dotad: a database of therapeutic antibody developability. Interdiscip Sci. 2024;16(3):623–634. doi: 10.1007/s12539-024-00613-2.
17. Jain T, Sun T, Durand S, Hall A, Houston NR, Nett JH, Sharkey B, Bobrowicz B, Caffry I, Yu Y. Biophysical properties of the clinical-stage antibody landscape. Proc Natl Acad Sci USA. 2017;114(5):944–949. doi: 10.1073/pnas.1616408114.
18. Makowski EK, Wang T, Zupancic JM, Huang J, Wu L, Schardt JS, De Groot AS, Elkins SL, Martin WD, Tessier PM. Optimization of therapeutic antibodies for reduced self-association and non-specific binding via interpretable machine learning. Nat Biomed Eng. 2024 Jan. 8(1):45–56. doi: 10.1038/s41551-023-01074-6.
19. Shehata L, Maurer DP, Wec AZ, Lilov A, Champney E, Sun T, Archambault K, Burnina I, Lynaugh H, Zhi X, et al. Affinity maturation enhances antibody specificity but compromises conformational stability. Cell Rep. 2019;28(13):3300–3308.e4. doi: 10.1016/j.celrep.2019.08.056.
20. Channing G. Ginkgo Datapoints. Announcing the antibody developability prediction competition. Hugging Face Blog. 2025 Sep 8. https://huggingface.co/blog/ginkgo-datapoints/2025-abdev-competition
21. Olsen TH, Boyles F, Deane CM. Observed antibody space: a diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. Protein Sci. 2022;31(1):141–146. doi: 10.1002/pro.4205.
22. Dunbar J, Deane CM. Anarci: antigen receptor numbering and receptor classification. Bioinformatics. 2016;32(2):298–300. doi: 10.1093/bioinformatics/btv552.
23. Høie MH, Hummer AM, Olsen TH, Aguilar-Sanjuan B, Nielsen M, Deane CM. Antifold: improved structure-based antibody design using inverse folding. Bioinf Adv. 2025;5(1). article vbae202. doi: 10.1093/bioadv/vbae202.
24. Dreyer FA, Cutting D, Schneider C, Kenlay H, Deane CM. Inverse folding for antibody sequence design using deep learning. 2023 Oct 30. doi: 10.48550/arXiv.2310.19513.
25. Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, Ronneberger O, Willmore L, Ballard AJ, Bambrick J, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature. 2024 June 13;630(8016):493–500. doi: 10.1038/s41586-024-07487-w.
26. Kenlay H, Dreyer FA, Cutting D, Nissley D, Deane CM. Abodybuilder3: improved and scalable antibody structure predictions. Bioinformatics. 2024 Oct. 40(10). doi: 10.1093/bioinformatics/btae576.
27. Passaro S, Corso G, Wohlwend J, Reveiz M, Thaler S, Somnath VR, Getz N, Portnoi T, Roy J, Stark H, et al. Boltz-2: Towards accurate and efficient binding affinity prediction. bioRxiv. 2025 June 18. doi: 10.1101/2025.06.14.659707.
28. Boitreaud J, Dent J, Geisz D, McPartlon M, Meier J, Qiao Z, Rogozhnikov A, Rollins N, Wollenhaupt P, Wu K. Zero-shot antibody design in a 24-well plate. bioRxiv. 2025 June 30. doi: 10.1101/2025.07.05.663018.
29. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, Smetanin N, Verkuil R, Kabeli O, Shmueli Y, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science. 2023 Mar 17. 379(6637):1123–1130. doi: 10.1126/science.ade2574.
30. Olsen TH, Moal IH, Deane CM. Addressing the antibody germline bias and its effect on language models for improved antibody design. Bioinformatics. 2024 Nov. 40(11):btae618. doi: 10.1093/bioinformatics/btae618.
31. Turnbull OM, Oglic D, Croasdale-Wood R, Deane CM. P-IgGen: a paired antibody generative language model. Bioinformatics. 2024 Nov. 40(11):btae659. doi: 10.1093/bioinformatics/btae659.
32. Kuriata A, Iglesias V, Pujols J, Kurcinski M, Kmiecik S, Ventura S. Aggrescan3D (A3D) 2.0: prediction and engineering of protein solubility. Nucleic Acids Res. 2019 Jul 2. 47(W1):W300–W307. doi: 10.1093/nar/gkz321.
33. Kalejaye L, Chu J-M, Wu I-E, Amofah B, Lee A, Hutchinson M, Chakiath C, Dippel A, Kaplan G, Damschroder M, et al. Accelerating high-concentration monoclonal antibody development with large-scale viscosity data and ensemble deep learning. Mabs. 2025;17(1). Article 2483944. doi: 10.1080/19420862.2025.2483944.
34. Kalejaye L, Wu I-E, Terry T, Lai P-K. DeepSP: deep learning-based spatial properties to predict monoclonal antibody stability. Comput Struct Biotechnol J. 2024;23:2220–2229. doi: 10.1016/j.csbj.2024.05.029.
35. Su J, Han C, Zhou Y, Shan J, Zhou X, Yuan F. SaProt: Protein language modeling with structure-aware vocabulary. The Twelfth International Conference on Learning Representations (ICLR 2024); Vienna, Austria. 2024. https://openreview.net/forum?id=6MRm3G4NiU.
36. Chemical Computing Group ULC. Montreal (QC): Molecular Operating Environment (MOE); 2019. https://www.chemcomp.com/en/index.htm
37. Raybould MIJ, Turnbull OM, Suter A, Guloglu B, Deane CM. Contextualising the developability risk of antibodies with lambda light chains using enhanced therapeutic antibody profiling. Commun Biol. 2024 Jan 8. 7(1). doi: 10.1038/s42003-023-05744-8.