AbLEF: Antibody Language Ensemble Fusion for thermodynamically empowered property predictions

Zachary A Rollins
Modeling and Informatics
Merck & Co., Inc.
South San Francisco, CA, USA
zachary.rollins@merck.com

Talal Widatalla
Modeling and Informatics
Merck & Co., Inc.
South San Francisco, CA, USA

Andrew Waight
Discovery Biologics
Merck & Co., Inc.
South San Francisco, CA, USA

Alan C Cheng
Modeling and Informatics
Merck & Co., Inc.
South San Francisco, CA, USA

Essam Metwally
Modeling and Informatics
Merck & Co., Inc.
South San Francisco, CA, USA
essam.metwally@merck.com

Abstract

Pre-trained protein language and/or structural models are often fine-tuned on drug development properties (i.e., developability properties) to accelerate drug discovery initiatives. However, these models generally rely on a single structural conformation and/or a single sequence as a molecular representation. We present a physics-based model whereby structural ensemble representations are fused by a transformer-based architecture and concatenated to a language representation to predict antibody protein properties. AbLEF enables the direct infusion of thermodynamic information into latent space and this enhances property prediction by explicitly infusing dynamic molecular behavior that occurs during experimental measurement. We find that **(1)** ensembles of structures generated from molecular simulation can further improve antibody property prediction for small datasets, **(2)** fine-tuned large protein language models can match smaller antibody-specific language models at predicting antibody properties, **(3)** trained multimodal sequence and structural representations outperform sequence representations alone, **(4)** pre-trained sequence with structure models are competitive with shallow machine learning (ML) methods in the small data regime, and **(5)** predicting measured antibody properties remains difficult for limited high fidelity datasets. AbLEF has been made publicly available at https://github.com/merck/AbLEF.

1 Introduction

Monoclonal antibodies (mAbs) are the fastest growing therapeutic modality accounting for a $210.1 Billion USD market in 2022 [1] which can be attributed to naturally selected intrinsic properties: target specificity, low toxicity, massive mutation space, thermostability, and metabolic stability. In early stage mAb development, several developability properties are measured to assess the likelihood

The NeurIPS 2023 Workshop on New Frontiers of AI for Drug Discovery and Development (AI4D3 2023), New Orleans, LA, USA, 2023.

of success in the clinic including target antigen binding affinity, polyspecific reactivity, aggregation propensity, titer, hydrophobicity, solubility, thermostability, viscosity, serum half-life, and immunogenicity. Although identification of mAb variants with high antigen affinity is possible via antibody display technologies [2], the optimization on only antigen binding affinity often compromises other developability properties [3–6]. For example, high aggregation and low thermostability contribute to the cost of formulation, production, and storage of mAb therapeutics. These compromises increase the cost of goods (COGS) in mAb manufacturing and result in a current COGs range from $95-200/g, an order of magnitude above the global consensus target $10/g to enable global access to mAb therapies [7]. Thus, there is increasing demand to accelerate and reduce the cost of the drug development process by accurately predicting mAb developability properties.

High throughput experimentation is enabling the collection of developability property datasets that can be used to build machine learning (ML) models to predict properties from sequence and/or structure; however, high fidelity datasets are often small ($\sim 10^2$-10^3) relative to sequence space ($\geq 10^{13}$) [8], contain clustered sequences, and/or are non-trivial to merge due to differences in experimental measurement. In the small dataset regime, often shallow ML models are built on sequence and/or structural descriptors to predict developability properties [5, 9–11]. Recently, pre-trained language models (LM) and/or structural models have been fine-tuned on developability properties and achieved comparable performance [10, 12–14]. Moreover, AbPROP [14] reported rank correlation of antibody developability properties and demonstrated that (1) fine-tuned language models improve predictions over zero-shot, (2) trained multimodal sequence and structural models outperform sequence alone, and (3) structural pre-training on the inverse folding problem did not improve property prediction performance. Despite this success, prediction of measured experimental values is more difficult than relative Spearman ranking [15–17] and accurate predictions are likely required to accelerate experimental protein design. Moreover, uncertainty estimation is often difficult or expensive to assess and is sparsely reported in property prediction models [17–19]. In addition, modern property prediction models are predominantly trained on single-point representations (i.e., a single sequence or single structural conformation) and are potentially deprived of rich thermodynamic information contained along the potential energy surface of molecular ensembles [20]. In fact, molecular ensemble representations have been shown to improve property predictions, however, most of these techniques rely on averages (arithmetic or Boltzmann) [21, 22] and there is a lack of robust methodologies to fuse the members of the molecular ensemble. Such a methodology would enable the flexibility to generate molecular ensembles that are representative of a given experimental condition (e.g., antibodies at high temperature, low pH, high concentration, bound/unbound to antigen, etc.) and may be adept at predicting molecular properties that occur in disparate thermodynamic phase spaces. We present a novel deep learning framework, AbLEF, which fuses members of the molecular ensemble and demonstrates improved predictive performance with increasing ensemble length on two property datasets (**Figure 1**). Additionally, we assess the ability to predict measured experimental values and an estimation of uncertainty.

We frame this task as an analog to the computer vision problem of multi-image super-resolution (MISR) where the objective is to combine an ensemble of low resolution images and fuse them into a single high resolution image. A recent breakthrough in MISR, implements a novel transformer-based architecture (TR-MISR) that is insensitive to image order because of the assignment of simultaneous dynamic attention to all encoded feature vectors [23]. Herein, we propose that this architecture is suitable for learning dynamic molecular behavior from members of the molecular ensemble (**Figure 1**) by demonstrating improved predictive performance on a hydrophobic interaction chromatography retention time (HIC-RT) and temperature of aggregation (T_{agg}) dataset. Our contribution is three-fold:

1. We demonstrate the utility of supplying thermodynamic information from molecular ensembles into latent space for property prediction.
2. We apply a novel transformer-based architecture to robustly fuse information from members of the molecular ensemble.
3. We compare this methodology with current methods including shallow ML, protein language models, and graph neural networks.

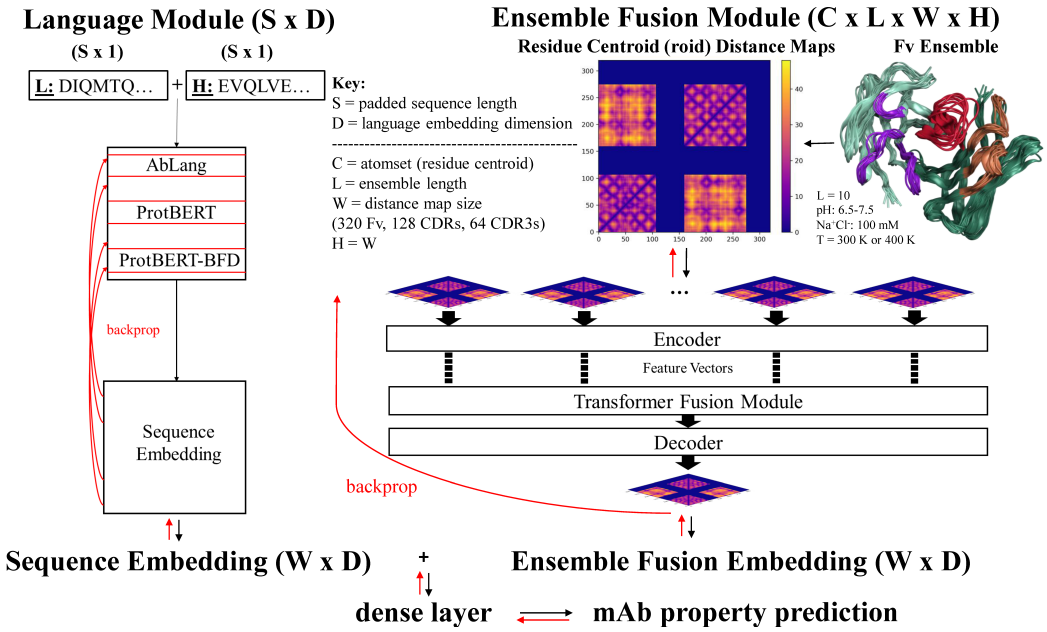

Figure 1: **Graphic of the AbLEF architecture.** This includes the language module (left), and the ensemble fusion module (right). Black arrows indicate the flow of information (sequence or distogram(s)) during training or inference. Red arrows indicate backpropagation into the AbLEF architecture. In the ensemble fusion module, there is a key (left), centroid distograms (center), and an ensemble of aligned Fvs (right) (framework (teal), CDRL1-L3 (purple), CDRH1-H2 (pink), CDR3H (red)).

2 Methods

2.1 Datasets, Fv structure generation, and ensemble sampling

The hydrophobic interaction chromatography retention time (HIC-RT) dataset contains 659 mAbs (597/62) (**Figure A1A**) and the temperature of aggregation (T_{agg}) dataset contains 521 mAbs (438/83) (**Figure A1B**). Detailed description of the experimental methods are provided (**Appendix B.1**). The mAbs are modeled as variable fragments (Fv) using the MOE 2022.01 Antibody Modeler application [24] using homologous template structures and ensemble sampling is performed with LowModeMD [25] in the Stochastic Titration application (**Appendix B.2**). The ensemble of structures are clustered by the backbone atoms' pairwise distance maps using density-based spatial clustering (eps=1.9, min_samples=1) [26] and the core structures from each cluster are used as the structural representations in AbLEF and graph neural network baselines (**Appendix B.2**). For example, the top ten most populated structural clusters are selected for AbLEF and the number of structural clusters used in training is referred to as ensemble length (L). Moreover, the most populated structural cluster is used for the graph neural network (GNN) baselines. Briefly, the molecular representations for AbLEF are paired residue distance matrices or distograms computed from the residue centroids (roids) (**Figure 1**). Details on the AbLEF and GNN molecular representations are provided (**Appendix B**). The test sets are designed to contain low sequence identity from the train/validation set ($< 95\%$) (**Appendix B.3**) while simultaneously containing a representative distribution of the developability property (**Figure A1**). Model performance was assessed based on the coefficient of determination (R^2) and compared with a two-tailed t-test after testing k-fold cross-validated models on the test set (k=10) (**Appendix B.3**).

2.2 AbLEF and baselines

AbLEF. The Antibody Language Ensemble Fusion model (AbLEF) contains two central components: (1) the pre-trained language module (AbLang, ProtBERT, or ProtBERT-BFD) and (2) the transformer-

based ensemble fusion module (**Figure 1**). The language module concatenates a fine-tuned heavy and light chain language model (W x D) where the width (W) is the size of the Fv sequence (320) and the hidden dimension (D) is determined by the pre-trained language model: AbLang (768), ProtBERT (1024), and ProtBERT-BFD (1024). The ensemble fusion module is supplied an ensemble of distograms (C x L x W x H) and outputs a fused distogram embedding (C x 1 x W x D). The fused distogram embedding is concatenated to the language module and supplied to a dense layer to predict the mAb property. Detailed description of the AbLEF architecture is provided (**Appendix B.4**).

Baselines. We compared AbLEF performance to several baselines including structural descriptors with machine learning (i.e., MOE + ML) and fine-tuned language models with graph neural networks. The mAb structural descriptors used in this study are included in MOE 2022.10: surface patch areas (hydrophobic, positive, negative), interaction energy between the heavy and light chain, relative angles of the heavy and light chain [27], potential energy, ASPmax [28], mono/dipole/quadrupole moments [29], isoelectric point [30], mass, etc. (43 total). The features were selected and trained on seven scikit-learn [31] regressors (**Appendix B.5**). The graph neural network baselines (GNNs) included geometric vector perceptrons (GVP) [32] and graph attention networks (GAT) [33]. Similar to AbLEF, the fine-tuned language model embeddings (AbLang, ProtBERT, and ProtBERT-BFD) were concatenated to the GNN node features during training and inference (**Appendix B.5**).

3 Results and Discussion

We investigated the performance of AbLEF by hyperparameter tuning the model architecture on an ensemble length of ten (L=10). The performance was assessed by retraining and testing at varying ensemble length (L=1-10). In addition, we performed several ablation studies. Finally, we compared AbLEF to numerous baselines including: ensemble averaged structural descriptors from MOE with machine learning (MOE + ML), fine-tuned language models (AbLang, ProtBERT, ProtBERT-BFD), and fine-tuned language with graph neural networks (GAT, GVP). We used the coefficient of determination, R^2, to assess the models' ability to predict measured experimental values on the test set. In addition, we quantified uncertainty and compared model performance by training and testing k-fold models (k=10). We find that **(1)** ensembles of structures can improve antibody property prediction, **(2)** pre-trained multimodal models can be competitive with shallow ML methods in the small data regime, **(3)** fine-tuned large protein language models can match smaller antibody-specific language models at predicting antibody properties, **(4)** multimodal molecular representations outperform sequence-only representations, and **(5)** predicting measured antibody properties remains difficult with modern methods/datasets.

3.1 AbLEF predictions

Language-only baselines. The hyperparameter search for the language-only models included the number of layers to backpropagate (AbLang, ProtBERT, ProtBERT-BFD), the learning rates (fine-tuned heavy chain model, fine-tuned light chain model, dense layer), the number of warm-up steps, and the dense layer dropout (**Table A1**). The language models were hyperparameter tuned on the HIC-RT and T_{agg} datasets (**Figure A2**). ProtBERT was the worst performing language model with an R^2 = 0.25 $\pm$ 0.04 (**Figure A2A**) and R^2 = 0.01 $\pm$ 0.01 (**Figure A2D**) on the HIC-RT and T_{agg} test set, respectively. ProtBERT-BFD and AbLang demonstrate comparable performance with an R^2 = 0.40 $\pm$ 0.03 (**Figure A2B**) and R^2 = 0.41 $\pm$ 0.06 (**Figure A2C**) on the HIC-RT test set, respectively. ProtBERT-BFD also matched AbLang (p=0.19) with an R^2 = 0.04 $\pm$ 0.02 (**Figure A2E**) and R^2 = 0.07 $\pm$ 0.01 (**Figure A2F**) on the T_{agg} test set, respectively. These results clearly demonstrate that, with datasets on the order of 10^2-10^3, HIC-RT is an easier property to predict from sequence than T_{agg} and other properties that were considered (e.g., temperature of melt, temperature of melt onset, polyspecific reactivity, etc.) (**not shown**). Importantly, this is consistent with AbPROP which demonstrates that HIC-RT and T_{agg} are the most predictable properties [14]. In addition, ProtBERT performed significantly worse than AbLang (p=0.04 and p=0.0005) and ProtBERT-BFD (p=0.008 and p=0.05) on the HIC-RT and T_{agg} datasets, respectively, demonstrating that generic protein language models can achieve comparable performance to an antibody-specific language model when trained on a large enough dataset ($\sim$2.1 billion protein sequences vs < 15 million antibody sequences). Although the protein language scaling laws are not yet well-defined, significantly increased performance of the fine-tuned ProtBERT-BFD model compared to the ProtBERT model demonstrates the importance of the pre-training task because BFD (Big Fantastic Database) is trained on 10X more protein sequences

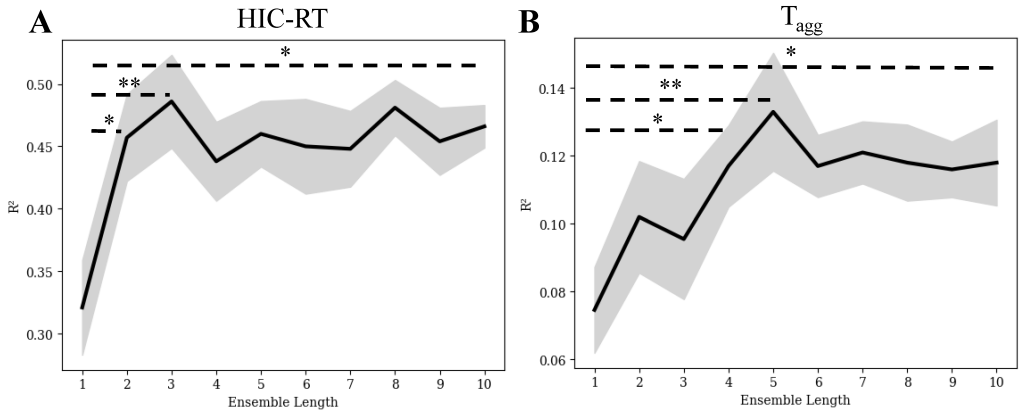

Figure 2: **AbLEF performance with increasing ensemble length.** The coefficient of determination (R^2), y-axis, is plotted as a function of the ensemble length, x-axis. Results represent the mean (dark lines) and standard error (light shading) of the models (k=10) on the test set for **(A)** the HIC-RT and **(B)** T_{agg} datasets. Statistical significance is assessed using a two-tailed t-test (*, **, *** correspond to p values < 0.05, 0.01, 0.001, respectively).

with an identical model architecture (**Appendix B.4**). This indicates that language model fine-tuning performance may continue to scale with additional protein/antibody sequences. Next, we proceeded to utilize the AbLang language model to investigate various ablation studies and the effect of ensemble length on the AbLEF model.

Ensemble length. The AbLEF model performance was evaluated on two datasets, HIC-RT and T_{agg}. The molecular ensembles were generated using MOE stochastic titration [24] at 300 K and 400 K, respectively, which employs LowModeMD [25] to represent the molecular conformations at the measured thermodynamic conditions (**Appendix B**). LowModeMD was primarily chosen for compute efficiency and future work will assess the computation trade-off of more robust ensemble generation methodologies (e.g., molecular dynamics, metadynamics, replica-exchange, etc.). During hyperparameter tuning, the AbLEF model was trained on an ensemble length of 10 (L=10) and re-trained and tested at varying ensemble length (L=1-10). Interestingly, the AbLEF model performance improved with ensemble length for both HIC-RT and T_{agg} datasets (**Figure 2**). For example, the maximum performance was achieved at L=3 for HIC-RT (R^2 = 0.49 $\pm$ 0.04) and L=5 for T_{agg} (R^2 = 0.13 $\pm$ 0.02) on the test set. In addition, we found consistent performance for ProtBERT and ProtBERT-BFD language models on the HIC-RT and T_{agg} datasets **(Figure A3)**. Moreover, HIC-RT performance at L=3 (p=0.0006) and L=2 (p=0.02) was significantly better than performance at L=1 (R^2 = 0.32 $\pm$ 0.06) (**Figure 2A**). Beyond L=3, the AbLEF HIC-RT performance flattened or slightly fluctuated around $R^2\sim$0.45 and remained statistically significant at L=10 (p=0.02). Similarly, T_{agg} performance at L=5 (p=0.01) and L=4 (p=0.03) was significantly better than performance at L=1 (R^2 = 0.07 $\pm$ 0.01) (**Figure 2B**). Beyond L=5, the AbLEF T_{agg} performance flattened or slightly fluctuated around $R^2\sim$0.12 and also remained statistically significant at L=10 (p=0.03). These results demonstrate that the AbLEF architecture can robustly fuse members of the molecular ensemble by significantly improved performance on the test set with ensemble length for the HIC-RT and T_{agg} datasets. The flattening of the AbLEF performance beyond the L^{th} ensemble length (L = 3, 5) may be attributed to reduced structural diversity in the ensemble beyond the L^{th} structural cluster (i.e, a limitation of the ensemble generation method) and/or destructive interference [34]. Moreover, since performance increased with ensemble length for both datasets and remained statistically significant at L=10, we expect a continued performance increase with more advanced ensemble generation methods that will increase the structural diversity within a given ensemble. Additionally, these results demonstrate that the optimal ensemble length for property prediction is dependent on the property as well as the ensemble generation method. Next, we assessed the performance of the AbLEF architecture on the HIC-RT dataset by performing several ablation studies.

Ablations. To unveil the minimum description length, we investigated AbLEF performance with the variable regions sliced out of the Fv distograms(**Figure 3**). The variable region indices were

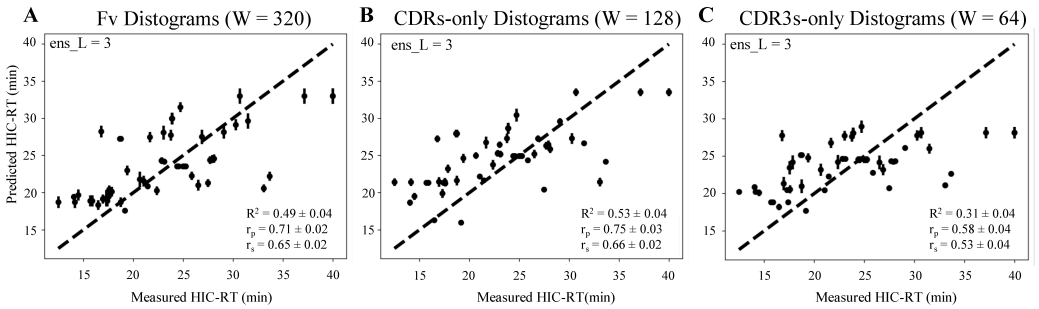

Figure 3: **AbLEF performance with CDRs-only and CDR3s-only.** This includes (**A**) AbLEF with Fv (**B**) AbLEF with CDRs-only and (**C**) AbLEF with CDR3s-only. AbLang is included in the AbLEF with CDRs-only and AbLEF with CDR3s-only hyperparameter search. The loss function was mean squared error and the hyperparameters were scored based on the average mean squared error across the validation sets (k=10). The ensemble length was set to 3 corresponding to the best performing AbLEF with Fv model. The x-axis is the measured HIC-RT and the y-axis is the predicted HIC-RT in minutes. The dashed line is the parity line, y=x. Performance metrics are displayed in the bottom right panel and include the coefficient of determination (R^2), Pearson correlation coefficient (r_p), and the Spearman rank correlation coefficient (r_s). The metric is the performance on the test set and the error corresponds to the standard error across the ten models trained during k-fold cross-validation (k=10).

determined by IMGT numbering of the complementarity determining regions (CDRs) which included the 6 CDRs (CDRs-only) and the 2 CDR3s (CDR3s-only). The CDRs-only and CDR3s-only ablation studies were performed by slicing out the distograms at the CDRs-only and CDR3s-only indices, respectively. Setting the ensemble length to 3 (i.e., maximum performance for the Fv distogram), we redid the hyperparameter search for AbLEF and evaluated performance on the HIC-RT test set. The best performing AbLEF model with CDRs-only had an average R^2 = 0.53 $\pm$ 0.04 on the test set (**Figure 3B**). Moreover, the best performing AbLEF model with CDR3s-only had an average R^2 = 0.31 $\pm$ 0.04 on the test set (**Figure 3C**). Interestingly, the CDRs-only is not significantly different from the full Fv distogram ensemble (p=0.50), but the CDR3s-only is significantly worse than both (Fv: p=0.005, CDRs-only: p=0.001). This suggests that inclusion of the 6 CDRs is crucial in predicting HIC-RT.

Next, we ablated the best performing AbLEF Fv model by separately removing the language and transformer-based ensemble fusion modules. We found that the AbLang module was contributing to most of the performance with an R^2 = 0.41 $\pm$ 0.06 compared to the transformer-based ensemble fusion module with an R^2 = 0.09 $\pm$ 0.02 (**Figure A4**). However, the ensemble fusion module significantly reduces model error across the k models (p=0.02) from k-fold cross-validation (k=10) thus contributing to the overall performance. This is consistent with AbPROP which demonstrates that pre-trained language models contribute a substantial portion of the prediction performance [14]. Finally, the AbLEF ensemble-only model was hyperparameter tuned to investigate the maximum performance that can be expected from the ensemble of structures (**Figure A5**). The hyperparameter search for the AbLEF ensemble-only included the learning rates (encoder/decoder and transformer), the number of warm-up steps, the number of layers in the encoder, the size of the encoded feature vectors, the number of layers in the transformer, the number of attention heads, size of the multilayer perceptron, and dropout (dense layer and transformer) (**Table A1**). The best performing AbLEF ensemble-only model had an average R^2 = 0.14 $\pm$ 0.01 on the test set (**Figure A5**). Interestingly, this is significantly better than the ablated ensemble-only model (p=0.04), however, significantly worse than the AbLang only. Future work will address the optimal structural ensemble pre-training tasks for downstream property prediction.

3.2 Baseline predictions

MOE + ML. We computed numerous ensemble averaged structural descriptors from MOE (43 total), learned the most important features, and trained multiple sci-kit learn regressors on the HIC-RT and

T_{agg} datasets (**Table A2**). For HIC-RT, the best structural descriptors included the CDR patch surface areas (hydrophobic, positive, negative), the hydrophobic moment, and the dipole moment (**Figure A6A**). The CDR patch areas are consistent with a previous generalized linear regression model on a smaller dataset [5] and a recently reported xgboost model [11]. Moreover, the selection of CDR structural descriptors is consistent with the AbLEF ablation study where we demonstrate that the 6 CDRS are crucial in predicting HIC-RT (**Figure 3**).The AbLEF model significantly outperformed this random forest regressor with a R^2 = 0.43 $\pm$ 0.01 on the HIC-RT test set (p=0.0002)(**Table 1**). For T_{agg}, the best structural descriptors included the heavy and light chain vector distance [27], radius of gyration, the 3D isoelectric point [30], and negatively charged surface area (**Figure A6B**). The vector distance measures the length between the heavy and light chain along a principal axis of rotation which may be representative of the molecular stability. Accordingly, the radius of gyration is a relative measure of molecular flexibility. These learned dynamic properties ranked highest in feature importance (**Figure A6B**) which supports the hypothesis that T_{agg} is an intrinsically dynamic feature best captured by molecular ensembles. This random forest regressor achieved a R^2 = 0.17 $\pm$ 0.005 on the T_{agg} test set (**Table 1**) comparable to the best AbLEF model performance (p=0.07). These results demonstrate that the AbLEF deep learning models are at least competitive with the MOE + ML baselines in the small data regime (10^2-10^3) and outperform MOE + ML baseline on the HIC-RT prediction task.

Table 1: **AbLEF and baseline model performance.** Model performance is assessed by the coefficient of determination, R^2, after k-fold cross validation (k=10). The mean and standard error of the k-models are reported on the test set. The columns are the model type, performance on the HIC-RT dataset, and performance on the T_{agg} dataset, respectively. The rows consist of model types separated by class: shallow ML, language-only, GNNs with language, and AbLEF models. The model class is partitioned by a black horizontal line. The best performing model for each class is *italicized* and the best deep learning model is also ***bolded***.

MODEL	**HIC-RT**	$\mathbf{T_{agg}}$
MOE + ML	*0.43 $\pm$ 0.01*	*0.17 $\pm$ 0.005*
ProtBERT (only)	0.25 $\pm$ 0.04	0.01 $\pm$ 0.01
ProtBERT-BFD (only)	0.40 $\pm$ 0.03	0.04 $\pm$ 0.02
AbLang (only)	*0.41 $\pm$ 0.06*	*0.07 $\pm$ 0.01*
GVP (ProtBERT)	0.30 $\pm$ 0.03	0.05 $\pm$ 0.02
GVP (ProtBERT-BFD)	0.42 $\pm$ 0.02	0.11 $\pm$ 0.02
GVP (AbLang)	0.23 $\pm$ 0.05	0.10 $\pm$ 0.02
GAT (ProtBERT)	0.35 $\pm$ 0.03	0.04 $\pm$ 0.01
GAT (ProtBERT-BFD)	*0.47 $\pm$ 0.07*	*0.11 $\pm$ 0.01*
GAT (AbLang)	0.40 $\pm$ 0.04	0.07 $\pm$ 0.02
AbLEF (ProtBERT)	0.27 $\pm$ 0.03[1]	0.03 $\pm$ 0.02[2]
AbLEF (ProtBERT-BFD)	0.45 $\pm$ 0.03[1]	0.05 $\pm$ 0.02[2]
AbLEF (AbLang)	***0.49 $\pm$ 0.04***[1]	***0.13 $\pm$ 0.02***[2]

Graph neural network baselines. We compared the AbLEF model to several graph neural network baselines including geometric vector perceptrons (GVP) (**Figure A7**) and graph attention networks (GAT) (**Figure A8**). The GNNs were trained on the HIC-RT and T_{agg} datasets (**Table A3**). We found that the GNNs with ProtBERT or ProtBERT-BFD outperformed the language-only models in 4/4 cases on HIC-RT and 4/4 cases on T_{agg} (**Table 1**). Interestingly, the GNNs with AbLang achieved similar performance as the language-only models in 1/2 cases on HIC-RT and 2/2 cases on T_{agg}. Overall, the GNNs either achieve equivalent or improved performance in 5/6 cases on HIC-RT and 6/6 cases on T_{agg}. This demonstrates that multimodal sequence and structural models outperform sequence models alone on antibody property datasets, however, this performance boost is predominately achieved with the ProtBERT or ProtBERT-BFD language models. Finally, the GNNs were competitive with the AbLEF model on the HIC-RT and T_{agg} datasets. For example, the best performing GAT model with ProtBERT-BFD achieved a R^2 = 0.47 $\pm$ 0.07 on the HIC-RT

[1] AbLEF model performance (L=3)

[2] AbLEF model performance (L=5)

and R^2 = 0.11 $\pm$ 0.04 on the T_{agg} test set (**Figure A8**). Although AbLEF with AbLang slightly outperformed GAT with ProtBERT-BFD, this was not statistically significant for HIC-RT (p=0.8) or T_{agg} (p=0.4). Despite this, AbLEF was able to significantly outperform GAT with AbLang and achieved comparable results to GAT/GVP with ProtBERT-BFD (**Table 1**). This demonstrates that the AbLEF architecture achieves competitive performance with modern GNNs (**Table 1**). Future work will address the appropriate attention or message passing mechanism to fuse ensembles of graph representations.

4 Conclusion

We present a novel transformer-based fusion method (AbLEF) for predicting antibody developability properties from Fv structural ensembles combined with fine-tuned language models. We demonstrate that AbLEF significantly outperforms language-only baselines. Additionally, we demonstrate that AbLEF is capable of fusing ensembles of structural representations to improve predictions of antibody developability properties. By generating ensembles of structures at their measured thermodynamic conditions, this methodology directly infuses thermodynamic information into latent space and significantly improves property prediction. For example, AbLEF significantly improved when increasing ensemble length from 1 to 3 for HIC-RT and 1 to 5 for T_{agg}. Despite this success, several datasets of similar size (e.g., temperature of melt, temperature of melt onset, polyspecific reactivity, etc.) were not able to achieve a sufficient regression performance $R^2 > 0$ at predicting measured experimental values and will likely require much more data $> 10^3$ or improved algorithm sample efficiency. Moreover, future work is needed to improve sample efficiency because high fidelity datasets for antibody developability properties are likely to remain $< 10^5$ for the foreseeable future. We believe that the AbLEF architecture is a promising method for predicting antibody developability properties from structural ensembles because AbLEF achieved competitive results to modern graph neural networks with language as well as structural descriptors with shallow ML on small datasets $\sim 10^2$. Future research will aim to improve sample efficiency by comparing molecular ensemble generation methods, assessing various ensemble fusion pre-training tasks, and fusing molecular graph representations.

Appendix A Supplementary Information

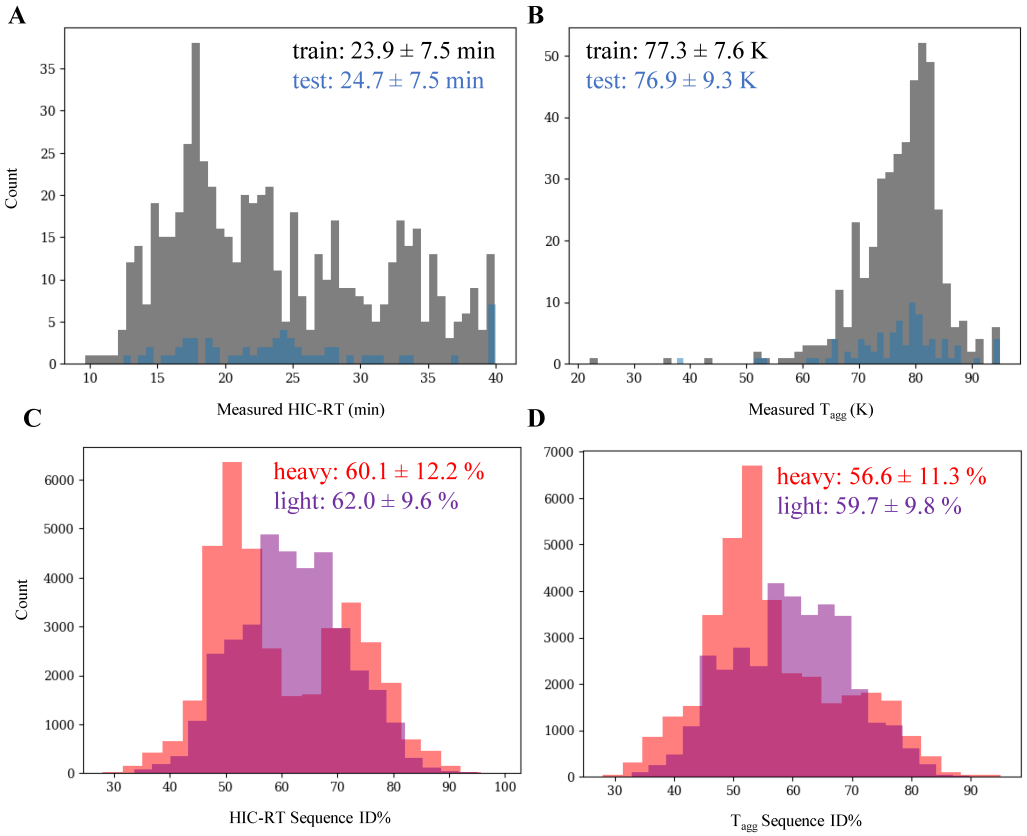

Figure A1: **HIC-RT and T_{agg} dataset distributions.** This includes train and test sets for (**A**) HIC-RT and (**B**) T_{agg}. The y-axis represents the count of datapoints and the x-axis is the property endpoint in (**A**) minutes and (**B**) degrees Kelvin. The train and test sets are shaded in black and blue, respectively. The mean and standard deviation of the train and test sets is displayed in their respective color and units. Pairwise sequence identity is computed between the train/validation and test sets after multiple sequence alignment (MSA) (**C-D**). The x-axis is the pairwise identity percentage and the y-axis is the count of sequence pairs from the heavy (red) or light (purple) chain distributed into 5% bins. This includes all sequence pairs between the train/validation and test sets for the (**C**) HIC-RT and (**D**) T_{agg} datasets. The mean and standard deviation of the sequence identities is displayed in the top right corner of each panel.

Table A1: **The searched AbLEF hyperparameters spaces.** For each model, the respective hyperparameter sets used in 100 sample Bayesian optimization with hyperband (BOHB) runs.

HYPERPARAMETERS	ABLEF	LANGUAGE-ONLY	ABLEF ENSEMBLE-ONLY	ABLEF CDRS-ONLY	ABLEF CDR3S-ONLY
LR CODER	LOGUNIFORM(1E-6, 1E-2)	LOGUNIFORM(1E-6, 1E-2)	LOGUNIFORM(1E-6, 1E-2)	LOGUNIFORM(1E-6, 1E-2)	LOGUNIFORM(1E-6, 1E-2)
LR ENS TRANSFORMER	LOGUNIFORM(1E-6, 1E-2)	N/A	LOGUNIFORM(1E-6, 1E-2)	LOGUNIFORM(1E-6, 1E-2)	LOGUNIFORM(1E-6, 1E-2)
LR LC	LOGUNIFORM(1E-9, 1E-5)	LOGUNIFORM(1E-9, 1E-5)	N/A	LOGUNIFORM(1E-9, 1E-5)	LOGUNIFORM(1E-9, 1E-5)
LR HC	LOGUNIFORM(1E-9, 1E-5)	LOGUNIFORM(1E-9, 1E-5)	N/A	LOGUNIFORM(1E-9, 1E-5)	LOGUNIFORM(1E-9, 1E-5)
LR STEP	RANDINT(1,5)	RANDINT(1,5)	RANDINT(1,5)	RANDINT(1,5)	RANDINT(1,5)
ENCODER N LAYERS	CHOICE([1,2])	N/A	CHOICE([1,2])	CHOICE([1,2])	CHOICE([1,2])
ENCODER N FEATURE VECTORS	CHOICE([16, 32, 64])	N/A	CHOICE([16, 32, 64])	CHOICE([16, 32, 64])	CHOICE([16, 32, 64])
TRANSFORMER DEPTH	CHOICE([1,2])	N/A	CHOICE([1,2])	CHOICE([1,2])	CHOICE([1,2])
TRANSFORMER HEADS	CHOICE([1, 2, 4, 8])	N/A	CHOICE([1, 2, 4, 8])	CHOICE([1, 2, 4, 8])	CHOICE([1, 2, 4, 8])
TRANSFORMER MLP DIM	CHOICE([16, 32, 64, 128])	N/A	CHOICE([16, 32, 64, 128])	CHOICE([16, 32, 64, 128])	CHOICE([16, 32, 64, 128])
TRANSFORMER DROPOUT	UNIFORM(0.1, 0.5)	N/A	UNIFORM(0.1, 0.5)	UNIFORM(0.1, 0.5)	UNIFORM(0.1, 0.5)
LANGUAGE FREEZE LAYER COUNT	RANDINT(8/20,12/30)	RANDINT(8/20,12/30)	N/A	RANDINT(8/20,12/30)	RANDINT(8/20,12/30)
DENSE LAYER DROPOUT	UNIFORM(0.1, 0.5)	UNIFORM(0.1, 0.5)	UNIFORM(0.1, 0.5)	UNIFORM(0.1, 0.5)	UNIFORM(0.1, 0.5)
CDR PATCH	FALSE	FALSE	FALSE	CDRS	CDR3S

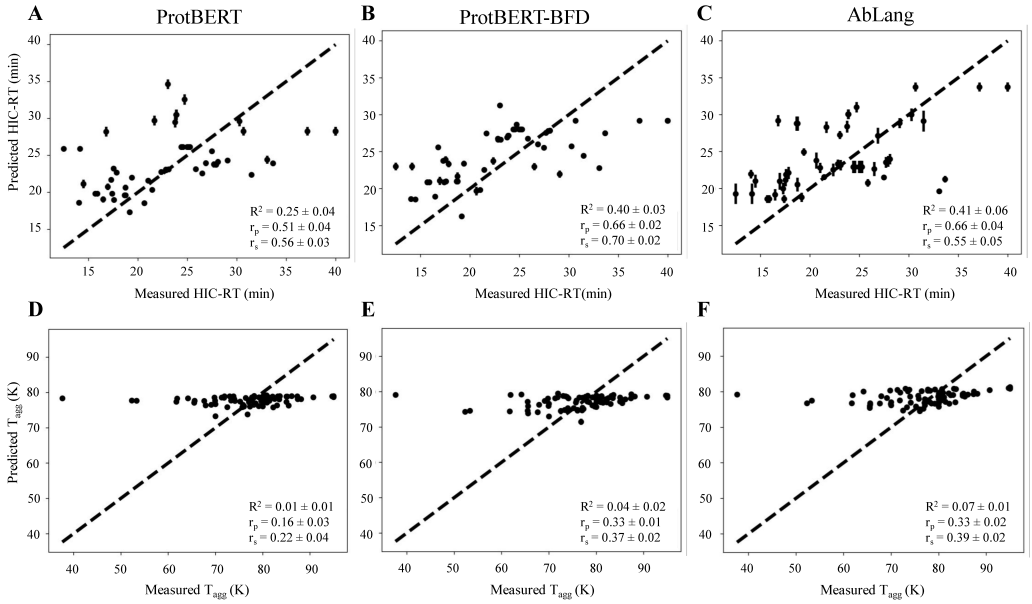

Figure A2: **Language-only baseline performance.** The top and bottom row designate the performance of the hyperparameter-tuned language models on the HIC-RT and T_{agg} datasets, respectively. The columns designate the language model: ProtBERT (**A,D**), ProtBERT-BFD (**B,E**), and AbLang (**C,F**). The x-axes are the measured values and the y-axes are the predicted values in minutes and degrees Kelvin for HIC-RT (**A-C**) and T_{agg} (**D-F**), respectively. The performance is reported as the mean and standard error of the k-trained models on the test set (k=10). The dashed line is the parity line, y=x. Performance metrics are displayed in the bottom right panel and include the coefficient of determination (R^2), Pearson correlation coefficient (r_p), and the Spearman rank correlation coefficient (r_s).

Table A2: **The searched MOE + ML hyperparameters spaces.** For each model, the respective hyperparameter sets were used in 100 sample skopt runs.

Hyperparameters	Linear Regression	Elastic Net	kNN	SVM	Random Forest	AdaBoost	XGBoost
power	uniform(0, 3)	N/A	N/A	N/A	N/A	N/A	N/A
alpha	uniform(1e-6, 1e6)	uniform(1e-6, 1e6)	N/A	N/A	N/A	N/A	N/A
l1 ratio	N/A	uniform(0, 1)	N/A	N/A	N/A	N/A	N/A
n neighbors	N/A	N/A	uniform(1, 50)	N/A	N/A	N/A	N/A
leaf size	N/A	N/A	uniform(1, 50)	N/A	N/A	N/A	N/A
p	N/A	N/A	uniform(1,2)	N/A	N/A	N/A	N/A
C	N/A	N/A	N/A	uniform(1e-1,1e3)	N/A	N/A	N/A
gamma	N/A	N/A	N/A	uniform(1e-4,1e3)	N/A	N/A	N/A
max depth	N/A	N/A	N/A	N/A	N/A	N/A	uniform(1,10)
max features	N/A	N/A	N/A	N/A	choice(['sqrt', 'log2'])	N/A	N/A
max leaf nodes	N/A	N/A	N/A	N/A	N/A	N/A	N/A
n estimators	N/A	N/A	N/A	N/A	uniform(10, 1000)	uniform(10, 1000)	uniform(10, 1000)

Table A3: **The searched GNN hyperparameters spaces.** For each model, the respective hyperparameter sets used in 100 sample Bayesian optimization with hyperband (BOHB) runs.

Hyperparameters	GVP	GAT
LR coder	loguniform(1e-6, 1e-2)	loguniform(1e-6, 1e-2)
LR lc	loguniform(1e-9, 1e-5)	loguniform(1e-9, 1e-5)
LR hc	loguniform(1e-9, 1e-5)	loguniform(1e-9, 1e-5)
LR step	randint(1,5)	randint(1,5)
Language freeze layer count	randint(8/20,12/30)	randint(8/20,12/30)
Dense layer dropout	uniform(0.1, 0.5)	uniform(0.1, 0.5)

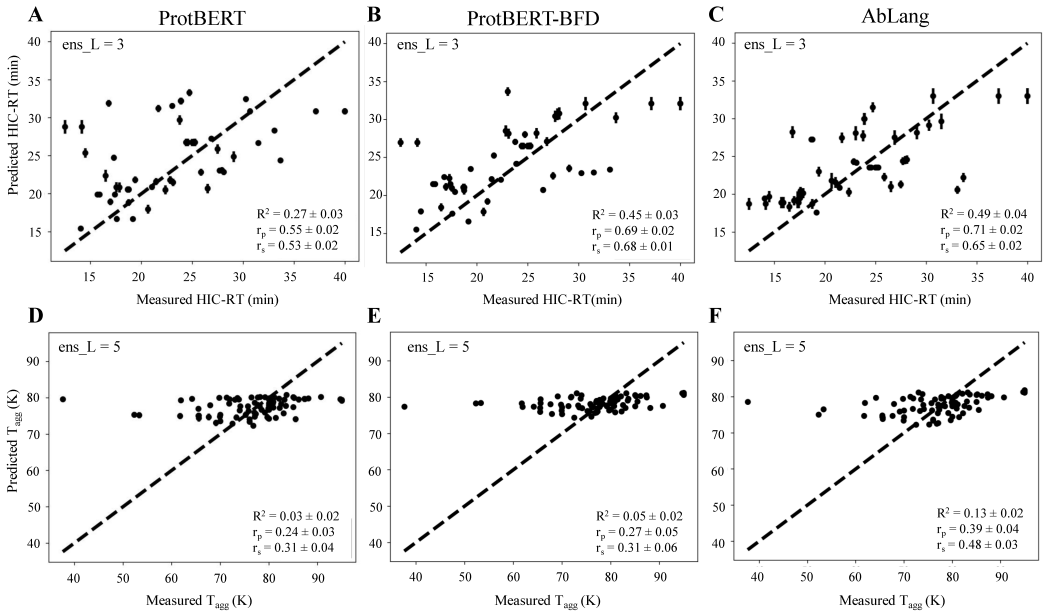

Figure A3: **AbLEF performance with language models.** The top and bottom row designate the performance of the hyperparameter-tuned AbLEF models on the HIC-RT and T_{agg} datasets, respectively. The columns designate the language model: ProtBERT (**A,D**), ProtBERT-BFD (**B,E**), and AbLang (**C,F**). The x-axes are the measured values and the y-axes are the predicted values in minutes and degrees Kelvin for HIC-RT (**A-C**) and T_{agg} (**D-F**), respectively. The performance is reported as the mean and standard error of the k-trained models on the test set (k=10). The dashed line is the parity line, y=x. Performance metrics are displayed in the bottom right panel and include the coefficient of determination (R^2), Pearson correlation coefficient (r_p), and the Spearman rank correlation coefficient (r_s).

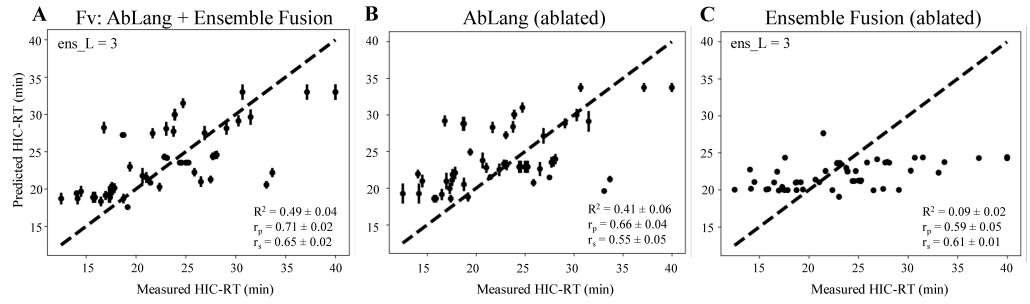

Figure A4: **Ablation of the AbLEF architecture.** The best performing AbLEF model was ablated to determine the relative contribution of the modules. This includes (**A**) the best performing AbLEF model, (**B**) the ablated AbLang module, and (**C**) the ablated ensemble fusion module. The loss function was mean squared error and the best performing model was saved based on the average mean squared error across the validation sets (k=10) during training. The ten models were then assessed on the test set and the panels display performance metrics in the bottom right corner. This includes the mean and standard error of the coefficient of determination (R^2), Pearson correlation coefficient (r_p), and the Spearman rank correlation coefficient (r_s). The ensemble length was set to 3 corresponding to the best performing AbLEF with Fv model. The x-axis is the measured HIC-RT and the y-axis is the predicted HIC-RT in minutes. The dashed line is the parity line, y=x.

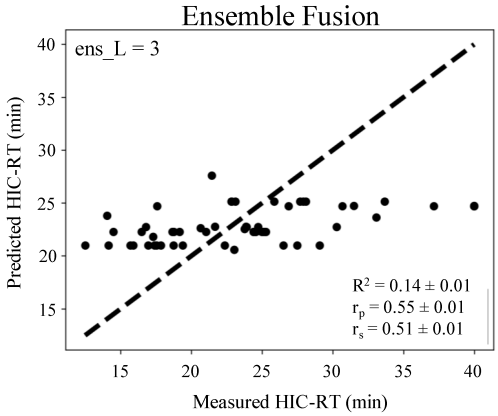

Figure A5: **AbLEF ensemble-only performance.** The ensemble fusion module was hyperparameter tuned on the HIC-RT dataset to evaluate the maximum possible performance with our given dataset. The loss function was mean squared error and the best performing model was saved based on the average mean squared error across the validation sets (k=10) during training. The ten models were then assessed on the test set and the panels display performance metrics in the bottom right corner. This includes the mean and standard error of the coefficient of determination (R^2), Pearson correlation coefficient (r_p), and the Spearman rank correlation coefficient (r_s). The ensemble length was set to 3 corresponding to the best performing AbLEF with Fv model. The x-axis is the measured HIC-RT and the y-axis is the predicted HIC-RT in minutes. The dashed line is the parity line, y=x.

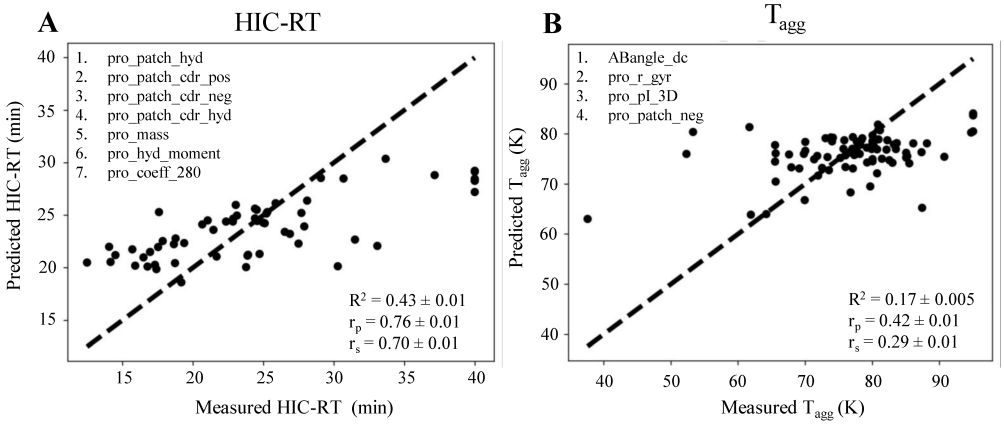

Figure A6: **MOE + ML performance.** Structural descriptors were computed in MOE and used to generate comparative models with shallow machine learning methods. After feature selection and hyperparameter tuning seven scikit-learn regressors, random forest models were the best performing regressors on both the (**A**) HIC-RT and (**B**) T_{agg} test sets. The list of features from exhaustive feature selection are ranked by feature importance and displayed in the upper left of each panel. This includes train and test sets for (**A**) HIC-RT and (**B**) T_{agg}. The x-axis is the measured values the y-axis is the predicted values for HIC-RT and T_{agg}, respectively. The dashed line is the parity line, y=x. Performance metrics are displayed in the bottom right panel and include the coefficient of determination (R^2), Pearson correlation coefficient (r_p), and the Spearman rank correlation coefficient (r_s). The mean and standard error of the models are computed with a jackknifing technique where k-models are trained and tested (k=10).

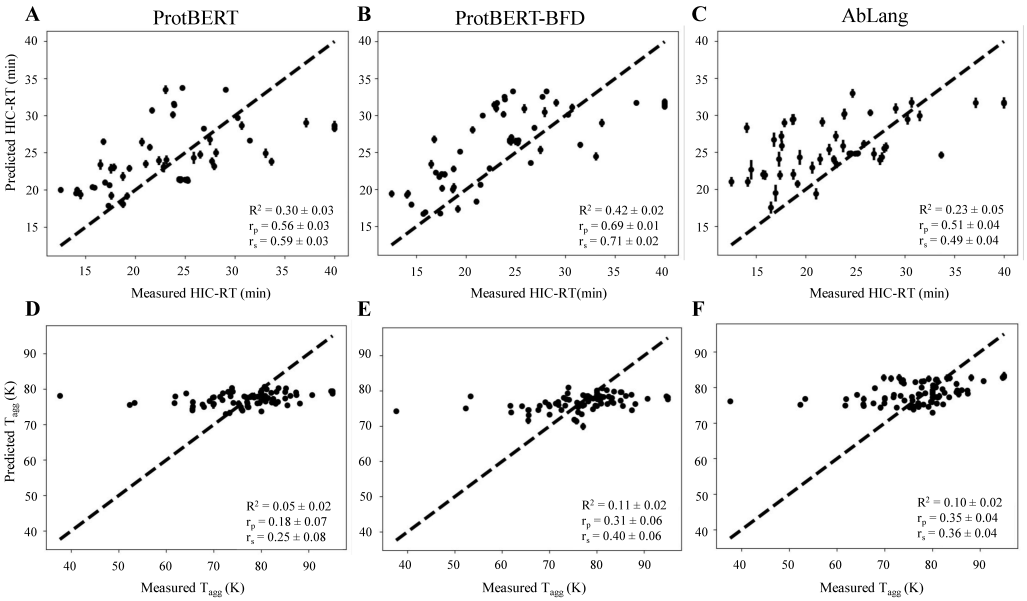

Figure A7: **GVP model performance.** The top and bottom row designate the performance of the hyperparameter-tuned GVP models on the HIC-RT and T_{agg} datasets, respectively. The columns designate the language model: ProtBERT (**A,D**), ProtBERT-BFD (**B,E**), and AbLang (**C,F**). The x-axes are the measured values and the y-axes are the predicted values in minutes and degrees Kelvin for HIC-RT (**A-C**) and T_{agg} (**D-F**), respectively. The performance is reported as the mean and standard error of the k-trained models on the test set (k=10). The dashed line is the parity line, y=x. Performance metrics are displayed in the bottom right panel and include the coefficient of determination (R^2), Pearson correlation coefficient (r_p), and the Spearman rank correlation coefficient (r_s).

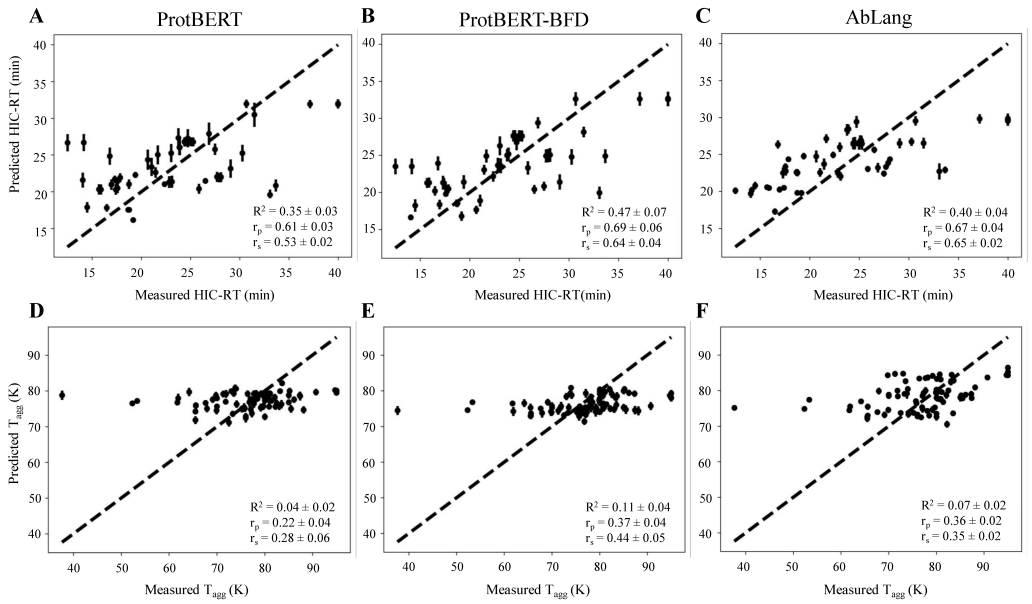

Figure A8: **GAT model performance.** The top and bottom row designate the performance of the hyperparameter-tuned GAT models on the HIC-RT and T_{agg} datasets, respectively. The columns designate the language model: ProtBERT (**A,D**), ProtBERT-BFD (**B,E**), and AbLang (**C,F**). The x-axes are the measured values and the y-axes are the predicted values in minutes and degrees Kelvin for HIC-RT (**A-C**) and T_{agg} (**D-F**), respectively. The performance is reported as the mean and standard error of the k-trained models on the test set (k=10). The dashed line is the parity line, y=x. Performance metrics are displayed in the bottom right panel and include the coefficient of determination (R^2), Pearson correlation coefficient (r_p), and the Spearman rank correlation coefficient (r_s).

Appendix B Extended Methods

B.1 Datasets

Hydrophobic Interaction Chromatography Retention Time (HIC-RT). The HIC-RT dataset used in this study contains 659 mAbs (597/62) (**Figure A1A**) with groups of sequences designed for different antigen targets [11]. Described previously [5], the hydrophobicity of a given mAb was determined by recording the elution time through a hydrophobic interaction chromatography column at 100 mM sodium phosphate, pH 7.0, and T = 300 K.

Temperature of aggregation ($\mathbf{T_{agg}}$). The T_{agg} dataset used in this study contains 521 mAbs (438/83) (**Figure A1B**) with groups of sequences designed for different antigen targets [11]. Described previously [5, 35], Nano-DSF (nano - differential scanning fluorimetry) studies were performed using the Nanotemper Prometheus NT.48 instrument. Briefly, the samples (10 μL at 0.5 mg/mL) were loaded into capillaries and the temperature ramped 1 °C/min from 20 °C to 94.8 °C. The colloidal stability of the sample can be determined by measuring the attenuation of back reflected light intensity passing through the sample and the aggregation temperature (T_{agg}) is defined as the point at which light scattering increases (or back reflected light intensity decreases) due to colloidal instability.

B.2 Fv structure generation and ensemble sampling

The variable fragment (Fv) regions of the IgG1 mAbs were homology modeled using the Antibody Modeler application in MOE 2022.01 [24]. Homology search is performed to identify the most similar template structure in the PDB for the framework region and the six complementarity determining regions (CDRs). The models were selected by the best MOE score (default settings) and minimized with the Amber10:EHT force field using the GB implicit solvent [36]. Ensemble sampling was performed using the stochastic titration application in MOE 2022.01 which employs LowModeMD sampling [25] to generate 50 structures under the following thermodynamic conditions: pH 7.0 $\pm$ 0.5, T = 300 K (HIC-RT) or 400 K (T_{agg}), NaCl = 100 mM, and the GB implicit solvent [36]. The Fv backbone atoms are subject to default harmonic position restraints: 0.5 Å buffer zone and 10 $kcal/mol\cdot$ Å^2 force constant. LowModeMD was selected for ensemble generation because it is a fast and efficient method to sample the potential energy surface. Structural descriptors were calculated for all structures and the arithmetic averages are used in the shallow ML methods. For antibody language ensemble fusion (AbLEF), the 50 structures were clustered using density-based spatial clustering (eps=1.9, min_samples=1) [26] based on the backbone atoms' pairwise distance maps and the top ten clusters were selected. Additionally, the eps core structure was selected from each cluster as the predominate structural representation of each cluster. The residue-residue distance maps or distograms were calculated for each eps core structure using the residue centroid (roid) where the centroid represents the center of mass (excluding hydrogens). For the graph neural network baselines, geometric vector perceptrons and graph attention network (GVP and GAT), the most populated eps core cluster was used as the structural representation.

B.3 Sequence clustering and train/test assessment

The test set was designed to contain low sequence identity ($< 95\%$) from the train/validation set to minimize data leakage and provide a worst-case scenario model fit. Briefly, the sequence analysis can be summarized in 2 steps [11, 37, 38]: (1) sequences are clustered by generating a pairwise mutation matrix (2) sequences are stratified such that sequence identity is $< 95\%$ in the test set, the test set contains approximately 10-20$\%$ of the dataset, and there is a representative distribution of the developability property in both the train/validation and test sets (**Figure A1**) [11]. The maximum sequence identity between training/validation and test sets was 94.1$\%$ and 90.1$\%$ for the HIC-RT (**Figure A1C**) and T_{agg} (**Figure A1D**) datasets, respectively. Sequence identity is computed in MOE [24] after multiple sequence alignment (MSA) and defined as the number of residue matches divided by sequence length. Additional datasets were considered (e.g., temperature of melt, temperature of melt onset, polyspecific reactivity, etc.), but did not achieve a sufficient regression performance $R^2 > 0$ at predicting measured experimental values and will likely require more data $> 10^3$ (**not shown**). To assess the confidence and significance of predictions on these small datasets, k-fold (k=10) cross-validation was employed in the training/validation stage and all 10 models were used

to predict the test set. The significance of difference between means and standard deviations is determined with a two-tailed t-test and two-tailed F-test, respectively.

B.4 AbLEF

This model contains two central components: (1) the pre-trained language module (AbLang, ProtBERT, or ProtBERT-BFD) and (2) the transformer-based ensemble fusion module (**Figure 1**). AbLang [39] is a transformer-based language model with ∼85 million parameters and descendant of BERT [40] that is pre-trained on a database of antibody sequences from the Observed Antibody Space (OAS) [41] (∼14 million heavy chains, ∼0.19 million light chains). The original task of AbLang is to predict masked residues in the antibody heavy or light chain. Similarly, ProtBERT and ProtBERT-BFD [42] were trained on ∼217 million and ∼2.1 billion antibody nonexclusive protein sequences from UniRef100 [43] and BFD (Big Fantastic Database) [44], respectively, and contain ∼420 million parameters. The language module is supplied a heavy and light chain sequence (max length 160 x 1 per chain) and outputs a concatenated sequence embedding from the respective pre-trained language light or heavy chain model (W x D) (**Figure 1**). The hidden dimension (D) is determined by the pre-trained language model: AbLang (768), ProtBERT (1024), and ProtBERT-BFD (1024). The pre-trained language models are fine-tuned on the heavy and light chain sequences by backpropagating into the last 1/3 of language model layers. Therefore, the number of frozen layers for each model was hyperparameter tuned: AbLang (8-12), ProtBERT (20-30), and ProtBERT-BFD (20-30). The width of the tensors (W) is determined by AbLEF mode: Fv (320), CDRs-only (128), CDR3s-only (64). The ensemble fusion module is supplied an ensemble of distograms (C x L x W x H) and outputs a fused distogram embedding (C x 1 x W x D). The language module and ensemble fusion module are concatenated and supplied to a dense layer to predict the mAb property. The distograms are calculated from the residue-residue pair centroids and shown as heat maps in Å (**Figure 1**). IMGT numbering [45] is used to determine the CDR locations in the distograms and zero-padding is consistent with the sequence gap tokens (-) determined by the IMGT canonical alignment.

The transformer-based ensemble fusion module is based on the TR-MISR architecture containing an encoder, transformer, and decoder [23]. The distograms are first encoded into feature vectors (**Figure 1**). The transformer contains an additional learnable embedding vector that fuses these feature vectors. This allows for attention to be applied to all members of the ensemble simultaneously when generating the fused representation. As in TR-MISR, this architecture does not require pre-training because the fusion module does not need to learn the spatial relation between feature vectors (number of trained image fusions, n=593) [23]. In addition, AbPROP [14] previously found no advantage to using pre-trained structural models on inverse folding, however, future work may explore alternative pre-training tasks to increase sample efficiency for molecular property prediction. The fused representation is then decoded, concatenated to the language model, and run through a dense layer to predict the mAb property.

Hyperparameters were selected using the Bayesian optimization with hyperband (BOHB) algorithm [46] implemented by Ray Tune [47] (**Table A1**). This algorithm reduces hyperparameter search wall time up to 50X by combining the sample efficiency of Bayesian optimization and the adaptive sampling/ early stopping advantages of bandit methodologies [46]. The best hyperparameter set was selected based on the lowest average mean squared error across the k-fold validation sets after a 100 sample BOHB run (coefficient of determination (R^2), Pearson correlation coefficient (r_p), and the Spearman rank correlation coefficient (r_s) are also monitored). Final performance is evaluated based on the average and standard deviation of the k models on the test set. Hyperparameter selection process was performed at an ensemble length of 10 and then retrained and tested at ensemble lengths 1-10. For the AbLang-only and ensemble-only baselines, the hyperparameters were re-selected with the same 100 sample BOHB runs (**Table A1**). Likewise, for the CDRs-only and CDR3s-only hyperparameter searches. The learning rate decay strategy was adopted from BERT with linear increase warm-up and inverse square root decay [40]. All runs were performed for 50 epochs, with batch size 16, at 32-bit precision, using L2 loss (MSE), utilizing the Adam optimizer [48], and on 4 A100 GPUs with 80 GB of memory.

B.5 Baselines

Structural descriptors + ML baselines. The mAb structural descriptors used in this study are included in MOE 2022.10: surface patch areas (hydrophobic, positive, negative), interaction energy

between the heavy and light chain, relative angles of the heavy and light chain [27], potential energy, ASPmax [28], mono/dipole/quadrupole moments [29], isoelectric point [30], mass, etc. (43 total). The descriptors were arithmetically averaged over the ensemble, and selected in the following sequential manner before training of the regressors:(1) ranked by random forest feature importance, (2) recursive features selected, and (3) exhaustive features selected from the top 10 ranked descriptors after recursion (i.e., 10 chose 1-10) [49]. Next, the features were trained on seven scikit-learn [31] regressors using k-fold (k=10) cross-validation, hyperparameter searched (skopt), ranked by mean squared error, and refit (**Table A2**). The regressors included were linear regression, elastic net, support vector machine, k nearest neighbors, random forest, adaboost, and xgboost. The best regressor was selected based on the lowest mean squared error on the test set. Uncertainty estimation was determined by a jackknifing technique [50] where the standard deviation in performance is computed by testing the k-fold (k=10) trained models on the test set. We denote these models that are learned from the MOE structural descriptors as MOE + ML throughout the manuscript.

Graph neural network baselines. The graph neural network (GNN) baselines included geometric vector perceptrons (GVP) [32] and graph attention networks (GAT) [33]. First, the most populated eps core cluster structure is selected and converted to a k nearest neighbor (kNN) graph representation G(n,e) with nodes (n) representing C^{α} coordinates and edges (e) representing the k nearest neighbors (k=30). In addition, several node and edge features are computed for GVP message passing including: sine and cosine of the three backbone dihedral angles as scalar node features, six unit vectors describing the orientation of the protein backbone as vector node features, a distance between residues (C^{α}-C^{α}) as a scalar edge feature, and a unit vector describing the orientation of the residues as a vector edge feature [32]. For comparison to AbLEF, the GNNs were not pre-trained and the language model embeddings from the fine-tuned AbLang, ProtBERT, or ProtBERT-BFD models are concatenated to the node scalar features during training and inference [13]. Hyperparameter search was performed analogously to AbLEF with 100 BOHB tuning runs on the hyperparameters (**Table A3**).

Acknowledgments and Disclosure of Funding

The authors acknowledge the contributions of all the members of the Protein Sciences Department within Discovery Biologics at Merck & Co., Inc., South San Francisco, CA, USA and Marc Bailly and Laurence Fayadat-Dilman. The authors also acknowledge the contributions of all the members of the Biologics Process R&D and Sterile Formulation Sciences Departments within Pharmaceutical Sciences at Merck & Co., Inc., Rahway, NJ, USA and members of the Modeling & Informatics group within Discovery Chemistry at Merck & Co., Inc., South San Francisco, CA, USA, especially BoRam Lee, Jingzhou Wang, Tanmoy Pal, Yunsie Chung, and Katherine Delevaux. All authors are/were Merck & Co., Inc. employees when conducting the work. The work here was fully funded by Merck & Co., Inc.

References

[1] Sameer Shah. Monoclonal antibodies market size, share and trends analysis report. Technical report, Grand View Research, Inc., San Francisco, CA, 2022.

[2] Bernhard Valldorf, Steffen C. Hinz, Giulio Russo, Lukas Pekar, Laura Mohr, Janina Klemm, Achim Doerner, Simon Krah, Michael Hust, and Stefan Zielonka. Antibody display technologies: selecting the cream of the crop. *Biological Chemistry*, 403:455–477, April 2022. doi: 10.1515/hsz-2020-0377.

[3] Tushar Jain, Tingwan Sun, Stéphanie Durand, Amy Hall, Nga Rewa Houston, Juergen H. Nett, Beth Sharkey, Beata Bobrowicz, Isabelle Caffry, Yao Yu, Yuan Cao, Heather Lynaugh, Michael Brown, Hemanta Baruah, Laura T. Gray, Eric M. Krauland, Yingda Xu, Maximiliano Vásquez, and K. Dane Wittrup. Biophysical properties of the clinical-stage antibody landscape. *Proceedings of the National Academy of Sciences*, 114:944–949, January 2017. doi: 10.1073/pnas.1616408114.

[4] Laila Shehata, Daniel P. Maurer, Anna Z. Wec, Asparouh Lilov, Elizabeth Champney, Tingwan Sun, Kimberly Archambault, Irina Burnina, Heather Lynaugh, Xiaoyong Zhi, Yingda Xu, and Laura M. Walker. Affinity Maturation Enhances Antibody Specificity but Compromises Conformational Stability. *Cell Reports*, 28:3300–3308.e4, September 2019. doi: 10.1016/j.celrep.2019.08.056.

[5] Marc Bailly, Carl Mieczkowski, Veronica Juan, Essam Metwally, Daniela Tomazela, Jeanne Baker, Makiko Uchida, Ester Kofman, Fahimeh Raoufi, Soha Motlagh, Yao Yu, Jihea Park, Smita Raghava, John Welsh, Michael Rauscher, Gopalan Raghunathan, Mark Hsieh, Yi-Ling Chen, Hang Thu Nguyen, Nhung Nguyen, Dan Cipriano, and Laurence Fayadat-Dilman. Predicting Antibody Developability Profiles Through Early Stage Discovery Screening. *mAbs*, 12, 2020. doi: 10.1080/19420862.2020.1743053.

[6] Monica L. Fernández-Quintero, Anne Ljungars, Franz Waibl, Victor Greiff, Jan Terje Andersen, Torleif T. Gjølberg, Timothy P. Jenkins, Bjørn Gunnar Voldborg, Lise Marie Grav, Sandeep Kumar, Guy Georges, Hubert Kettenberger, Klaus R. Liedl, Peter M. Tessier, John McCafferty, and Andreas H. Laustsen. Assessing developability early in the discovery process for novel biologics. *mAbs*, page 2171248, December 2023. doi: 10.1080/19420862.2023.2171248.

[7] Kevin J Whaley and Larry Zeitlin. Emerging antibody-based products for infectious diseases: Planning for metric ton manufacturing. *Human Vaccines & Immunotherapeutics*, 18:1930847, April 2022. doi: 10.1080/21645515.2021.1930847.

[8] Hedda Wardemann and Christian E. Busse. Novel Approaches to Analyze Immunoglobulin Repertoires. *Trends in Immunology*, 38:471–482, July 2017. doi: 10.1016/j.it.2017.05.003.

[9] Nathaniel L. Miller, Thomas Clark, Rahul Raman, and Ram Sasisekharan. Learned features of antibody-antigen binding affinity. *Frontiers in Molecular Biosciences*, 10, 2023.

[10] Yong Xiao Yang, Pan Wang, and Bao Ting Zhu. Binding affinity prediction for antibody protein antigen complexes: A machine learning analysis based on interface and surface areas. *Journal of Molecular Graphics and Modelling*, 118:108364, January 2023. doi: 10.1016/j.jmgm.2022.108364.

[11] Andrew B. Waight, David Prihoda, Rojan Shrestha, Kevin Metcalf, Marc Bailly, Marco Ancona, Talal Widatalla, Zachary Rollins, Alan C Cheng, Danny A. Bitton, and Laurence Fayadat-Dilman. A machine learning strategy for the identification of key in silico descriptors and prediction models for IgG monoclonal antibody developability properties. *mAbs*, 15:2248671, December 2023. ISSN 1942-0862. doi: 10.1080/19420862.2023.2248671.

[12] Vladimir Gligorijević, P. Douglas Renfrew, Tomasz Kosciolek, Julia Koehler Leman, Daniel Berenberg, Tommi Vatanen, Chris Chandler, Bryn C. Taylor, Ian M. Fisk, Hera Vlamakis, Ramnik J. Xavier, Rob Knight, Kyunghyun Cho, and Richard Bonneau. Structure-based protein function prediction using graph convolutional networks. *Nature Communications*, 12:3168, May 2021. doi: 10.1038/s41467-021-23303-9.

[13] Zichen Wang, Steven A. Combs, Ryan Brand, Miguel Romero Calvo, Panpan Xu, George Price, Nataliya Golovach, Emmanuel O. Salawu, Colby J. Wise, Sri Priya Ponnapalli, and Peter M. Clark. LM-GVP: an extensible sequence and structure informed deep learning framework for protein property prediction. *Scientific Reports*, 12:6832, April 2022. doi: 10.1038/s41598-022-10775-y.

[14] Talal Widatalla, Zachary A Rollins, Ming-Tang Chen, Andrew Waight, and Alan Cheng. AbPROP: Language and Graph Deep Learning for Antibody Property Prediction. *ICML Workshop on Computational Biology*, July 2023.

[15] Davide Chicco, Matthijs J. Warrens, and Giuseppe Jurman. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ. Computer Science*, 7:e623, 2021. ISSN 2376-5992. doi: 10.7717/peerj-cs.623.

[16] D. L. J. Alexander, A. Tropsha, and David A. Winkler. Beware of R2: simple, unambiguous assessment of the prediction accuracy of QSAR and QSPR models. *Journal of chemical information and modeling*, 55:1316–1322, July 2015. ISSN 1549-9596. doi: 10.1021/acs.jcim.5b00206.

[17] KyungPyo Ham, JeongNoh Yoon, and Lee Sael. Towards Accurate and Certain Molecular Properties Prediction. In *2022 13th International Conference on Information and Communication Technology Convergence (ICTC)*, pages 1621–1624, October 2022. doi: 10.1109/ICTC55196.2022.9952716.

[18] AkshatKumar Nigam, Robert Pollice, Matthew F. D. Hurley, Riley J. Hickman, Matteo Aldeghi, Naruki Yoshikawa, Seyone Chithrananda, Vincent A. Voelz, and Alán Aspuru-Guzik. Assigning Confidence to Molecular Property Prediction. *Expert opinion on drug discovery*, 16:1009–1023, September 2021. ISSN 1746-0441. doi: 10.1080/17460441.2021.1925247.

[19] Prudencio Tossou, Cas Wognum, Michael Craig, Hadrien Mary, and Emmanuel Noutahi. Real-World Molecular Out-Of-Distribution: Specification and Investigation, June 2023.

[20] Simon Axelrod and Rafael Gomez-Bombarelli. Molecular machine learning with conformer ensembles, February 2021.

[21] A. J. Hopfinger, Shen Wang, John S. Tokarski, Baiqiang Jin, Magaly Albuquerque, Prakash J. Madhav, and Chaya Duraiswami. Construction of 3D-QSAR Models Using the 4D-QSAR Analysis Formalism. *Journal of the American Chemical Society*, 119:10509–10524, October 1997. doi: 10.1021/ja9718937.

[22] Markus A. Lill. Multi-dimensional QSAR in drug discovery. *Drug Discovery Today*, 12: 1013–1017, December 2007. doi: 10.1016/j.drudis.2007.08.004.

[23] Tai An, Xin Zhang, Chunlei Huo, Bin Xue, Lingfeng Wang, and Chunhong Pan. TR-MISR: Multiimage Super-Resolution Based on Feature Fusion With Transformers. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:1373–1388, 2022. doi: 10.1109/JSTARS.2022.3143532.

[24] Chemical Computing Group ULC. Molecular Operating Environment (MOE), 2022.

[25] Paul Labute. LowModeMD—Implicit Low-Mode Velocity Filtering Applied to Conformational Search of Macrocycles and Protein Loops. *Journal of Chemical Information and Modeling*, 50: 792–800, May 2010. doi: 10.1021/ci900508k.

[26] Martin Ester, Hans-Peter Kriegel, Jorg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231, 1996.

[27] J. Dunbar, A. Fuchs, J. Shi, and C.M. Deane. ABangle: characterising the VHVL orientation in antibodies. *Protein Engineering, Design and Selection*, 26:611–620, October 2013. ISSN 1741-0126. doi: 10.1093/protein/gzt020.

[28] J. Cristian Salgado, Ivan Rapaport, and Juan A. Asenjo. Predicting the behaviour of proteins in hydrophobic interaction chromatography. 2. Using a statistical description of their surface amino acid distribution. *Journal of Chromatography. A*, 1107:120–129, February 2006. ISSN 0021-9673. doi: 10.1016/j.chroma.2005.12.033.

[29] Darrell Velegol, Jason D. Feick, and Lance R. Collins. Electrophoresis of Spherical Particles with a Random Distribution of Zeta Potential or Surface Charge. *Journal of Colloid and Interface Science*, 230:114–121, October 2000. ISSN 0021-9797. doi: 10.1006/jcis.2000.7049.

[30] A. Sillero and J. M. Ribeiro. Isoelectric points of proteins: theoretical determination. *Analytical Biochemistry*, 179:319–325, June 1989. ISSN 0003-2697. doi: 10.1016/0003-2697(89)90136-x.

[31] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[32] Bowen Jing, Stephan Eismann, Patricia Suriana, Raphael J. L. Townshend, and Ron Dror. Learning from Protein Structure with Geometric Vector Perceptrons, May 2021.

[33] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph Attention Networks, February 2018.

[34] Xiangyun Zhao, Haoxiang Li, Xiaohui Shen, Xiaodan Liang, and Ying Wu. A Modulation Module for Multi-task Learning with Applications in Image Retrieval, September 2018.

[35] Soo Hyun Kim, Han Ju Yoo, Eun Ji Park, and Dong Hee Na. Nano Differential Scanning Fluorimetry-Based Thermal Stability Screening and Optimal Buffer Selection for Immunoglobulin G. *Pharmaceuticals*, 15:29, January 2022. ISSN 1424-8247. doi: 10.3390/ph15010029.

[36] Michal Wojciechowski and Bogdan Lesyng. Generalized Born Model: Analysis, Refinement, and Applications to Proteins. *The Journal of Physical Chemistry B*, 108:18368–18376, November 2004. ISSN 1520-6106. doi: 10.1021/jp046748b.

[37] Georgios Pavlopoulos. How to cluster protein sequences: tools, tips and commands. *MOJ Proteomics & Bioinformatics*, Volume 5, June 2017. doi: 10.15406/mojpb.2017.05.00174.

[38] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C. J. Carey, Ilhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17:261–272, March 2020. doi: 10.1038/s41592-019-0686-2.

[39] Tobias H Olsen, Iain H Moal, and Charlotte M Deane. AbLang: an antibody language model for completing antibody sequences. *Bioinformatics Advances*, 2, January 2022. doi: 10.1093/bioadv/vbac046.

[40] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423.

[41] 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: A Publication of the Protein Society*, 31:141–146, January 2022. doi: 10.1002/pro.4205.

[42] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44: 7112–7127, October 2022. ISSN 1939-3539. doi: 10.1109/TPAMI.2021.3095381.

[43] Baris E. Suzek, Yuqi Wang, Hongzhan Huang, Peter B. McGarvey, Cathy H. Wu, and the UniProt Consortium. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31:926–932, March 2015. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu739.

[44] Martin Steinegger and Johannes Söding. Clustering huge protein sequence sets in linear time. *Nature Communications*, 9:2542, June 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-04964-5.

[45] Marie-Paule Lefranc, Véronique Giudicelli, Quentin Kaas, Elodie Duprat, Joumana Jabado-Michaloud, Dominique Scaviner, Chantal Ginestoux, Oliver Clément, Denys Chaume, and Gérard Lefranc. IMGT, the international ImMunoGeneTics information system ®. *Nucleic Acids Research*, 33:D593–D597, January 2005. doi: 10.1093/nar/gki065.

[46] Stefan Falkner, Aaron Klein, and Frank Hutter. BOHB: Robust and Efficient Hyperparameter Optimization at Scale, July 2018.

[47] Richard Liaw, Eric Liang, Robert Nishihara, Philipp Moritz, Joseph E. Gonzalez, and Ion Stoica. Tune: A Research Platform for Distributed Model Selection and Training, July 2018.

[48] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization, January 2017.

[49] Zachary A. Rollins, Jun Huang, Ilias Tagkopoulos, Roland Faller, and Steven C. George. A computational algorithm to assess the physiochemical determinants of T cell receptor dissociation kinetics. *Computational and Structural Biotechnology Journal*, 20:3473–3481, January 2022. ISSN 2001-0370. doi: 10.1016/j.csbj.2022.06.048.

[50] Rupert G. Miller. The Jackknife–A Review. *Biometrika*, 61:1–15, 1974. ISSN 0006-3444. doi: 10.2307/2334280. Publisher: [Oxford University Press, Biometrika Trust].