Biophysical Journal
Article

Biophysical Society

AbMelt: Learning antibody thermostability from molecular dynamics

Zachary A. Rollins,[1] Talal Widatalla,[1] Alan C. Cheng,[1] and Essam Metwally[1,*]
[1]Modeling and Informatics, Merck & Co., Inc., South San Francisco, California

ABSTRACT Antibody thermostability is challenging to predict from sequence and/or structure. This difficulty is likely due to the absence of direct entropic information. Herein, we present AbMelt where we model the inherent flexibility of homologous antibody structures using molecular dynamics simulations at three temperatures and learn the relevant descriptors to predict the temperatures of aggregation (T_{agg}), melt onset ($T_{m,on}$), and melt (T_m). We observed that the radius of gyration deviation of the complementarity determining regions at 400 K is the highest Pearson correlated descriptor with aggregation temperature ($r_p = -0.68 \pm 0.23$) and the deviation of internal molecular contacts at 350 K is the highest correlated descriptor with both $T_{m,on}$ ($r_p = -0.74 \pm 0.04$) as well as T_m ($r_p = -0.69 \pm 0.03$). Moreover, after descriptor selection and machine learning regression, we predict on a held-out test set containing both internal and public data and achieve robust performance for all endpoints compared with baseline models (T_{agg} $R^2 = 0.57 \pm 0.11$, $T_{m,on}$ $R^2 = 0.56 \pm 0.01$, and T_m $R^2 = 0.60 \pm 0.06$). In addition, the robustness of the AbMelt molecular dynamics methodology is demonstrated by only training on <5% of the data and outperforming more traditional machine learning models trained on the entire data set of more than 500 internal antibodies. Users can predict thermostability measurements for antibody variable fragments by collecting descriptors and using AbMelt, which has been made available.

SIGNIFICANCE Antibody thermostability properties have critical downstream effects; for example, low thermostability can require higher patient dosages as well as increase the cost of goods in manufacturing. We present a novel method, AbMelt, which combines multitemperature molecular dynamics and machine learning to predict experimental antibody thermostability measurements. AbMelt outperforms modern methods that utilize sequence and/or structure to predict antibody thermostability.

INTRODUCTION

The thermostability of a protein impacts not only biological function but storage stability and bioprocess complexity. Monoclonal antibodies (mAbs) exist isothermally in vivo; however, thermostability impacts protein design and downstream formulation, and contributes to the overall cost of goods in mAb manufacturing (1). Moreover, given the estimated US$2.6 billion cost to bring a biotherapeutic to market (including failed candidates) (2–4), improved prediction of mAb thermostability is justified to eliminate even a fraction of poor candidate molecules. Thermostability predictions rely on sufficient compute and model formulation as well as on experimental data sets. The collection of mAb thermostability data sets is itself capital intensive and the experimental endpoints, such as melting and aggregation temperature, are influenced by multidimensional experimental conditions such as the Ig framework selection, buffer constituents, salt concentrations, pH, protein concentration, temperature ramp, and instrumentation, to name a few (5–10). This dependence results in the inability to cleanly combine data sets and poses a serious restriction on modern data set sizes ($\sim 10^2$–10^3) (8–11).

Previous predictive models have explored the dependence of experimental conditions on mAb thermostability from sequence and structural descriptors (11–15). These attempts include one-hot sequence encoding (12), sequence-based language model embeddings (11,12), and sequence/structure ddG scores based on computed potential energies (13). One-hot sequence encodings represent a protein sequence as a binary tensor and are often used as a baseline model (12). Protein language models are pretrained to predict masked residues in a protein sequence and are thought to contain information-rich embeddings that can be

Submitted November 17, 2023, and accepted for publication June 4, 2024.
*Correspondence: essam.metwally@merck.com
Editor: Erik Lindahl.
https://doi.org/10.1016/j.bpj.2024.06.003

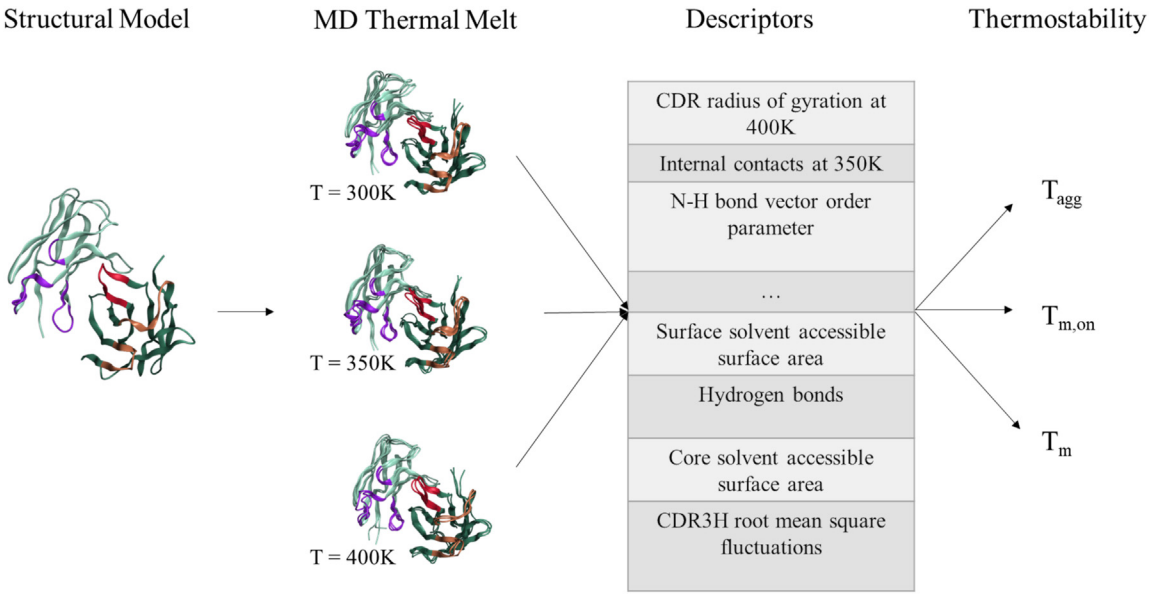

FIGURE 1 Process flow diagram to learn antibody thermostability from molecular dynamics. This includes the structural model of the antibody (*left*), the high-temperature molecular dynamics simulations (*center left*), the computation of descriptors (*center right*), and the prediction of thermostability endpoints (*right*). Refer to methods for details on the descriptor calculations. The color scheme: framework of the antibody variable fragment (Fv) (*teal*), CDRL1-3 (*purple*), CDRH1-2 (*salmon*), and CDRH3 (*red*).

immediately used for prediction or fine-tuned on downstream tasks (11,12). AbLIFT combined deep mutational scanning, Rosetta scoring, and clustering to optimize the variable fragment (Fv) heavy-light chain interface for thermostability (14). Similarly, structural descriptors, such as protein surface patches, have been demonstrated to be useful in predicting mAb developability properties (10,16). Despite a clear need for computational tools to predict mAb thermostability, predictions based on sequence and structure remain a challenge because of limited data sets and the lack of direct conformational entropy embedded in static sequence or structure representations. Recent work utilizing high-temperature molecular dynamics (MD) quantified conformational entropy by measuring internal contacts at 400 K for a small set of single-domain mAbs ($n = 7$) and obtained a high Pearson correlation coefficient ($r_p = 0.79$) to melting temperature (T_m) (15). Similar work for enzymes found a high T_m correlation by measuring the N-H bond vector order parameter in high temperature MD (17). This previous work provided impetus to investigate the ability to learn mAb thermostability directly from molecular simulation data.

Herein, we perform high temperature MD (300, 350, and 400 K) on a set of 25 mAb Fv regions, compute descriptors (described in methods) from the MD trajectories, and use machine learning (ML)-guided feature selection to predict aggregation temperature (T_{agg}), melting temperature onset ($T_{m,on}$), and T_m (Fig. 1). We further demonstrate the robustness of this AbMelt high-temperature MD method by outperforming current baseline models trained on greater than 500 internal Fvs (i.e., with ~20× the training data).

METHODS

Thermostability data set measured by nano-DSF (T_{agg}, $T_{m,on}$, T_m)

The thermostability data set used in this study contains 25 IgG1 datapoints. Nano-DSF (nano-differential scanning fluorimetry) studies were performed as described previously (10) using the Nanotemper Prometheus NT.48 instrument to measure protein stability. In brief, samples (~10 μL at 0.5–1 mg/mL) were loaded into capillaries and the temperature ramped at 1°C/min from 20 to 94.8°C.

The melting point temperatures ($T_{m,on}$ and T_m) indicate the structural stability of the samples, and the unfolding curves (or thermograms) were generated by plotting the ratio of the fluorescence intensities (F350 nm/F330 nm) as a function of temperature, with each intensity tracking the level of folded or unfolded protein. The melting point temperatures were defined by the onset ($T_{m,on}$) and inflection point (T_m) of the thermogram. The colloidal stability of the sample can be simultaneously determined by measuring the attenuation of back-reflected light intensity passing through the sample and the aggregation temperature (T_{agg}) was defined as the point at which light scattering increases (or back-reflected light intensity decreases) due to colloidal instability. For one datapoint, DAB011918, the aggregation temperature was not detected at 94.8°C and thus not included. This likely indicates that, after melt, the protein is in an intrinsically disordered state.

Antibody homology modeling

The Fv regions of the mAbs were homology modeled using the Antibody Modeler application in MOE 2022.02 (18). The Fv region of the mAb was selected to simplify the problem statement and to reduce the computational cost of performing the MD simulations. All collected thermostability measurements were performed on the IgG1 framework under identical experimental conditions. 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

TABLE 1 Descriptor set definitions measured from molecular dynamics

mAb Substructure	Temperature, T (K)	MD descriptor	Description	Units
Fv, H, L, CDRs, CDR1L, CDR2L, CDR3L, CDR1H, CDR2H, CDR3H	T = 300, 350, 400, all	R_g	radius of gyration of the mAb substructure at T	nanometers (nm)
Fv, H, L, CDRs, CDR1L, CDR2L, CDR3L, CDR1H, CDR2H, CDR3H	T = 300, 350, 400, all	RMSF	root mean-square fluctuations of the mAb substructure at T	nanometers (nm)
Fv, H, L, CDRs, CDR1L, CDR2L, CDR3L, CDR1H, CDR2H, CDR3H	T = 300, 350, 400, all	internal contacts	internal atom pairs within 3.5 Å of the mAb substructure at T	count/number (num)
Fv, H, L, CDRs, CDR1L, CDR2L, CDR3L, CDR1H, CDR2H, CDR3H	T = 300, 350, 400, all	hydrogen bonds	internal atom pairs within 3.5 Å and an angle less than 30° of the mAb substructure at T	count/number (num)
Fv	T = 300, 350, 400	S^2	The N-H bond vector order parameter of the Fv substructure at T	magnitude ranges between 0 and 1 ()
Fv	T = all	Λ	dimensionless number, Λ, related to heat capacity and quantifies the temperature dependence of S^2	dimensionless
Fv	T = all	r-Λ	coefficient of determination of the Λ linear fit	dimensionless
Fv	T = 300, 350, 400, all	core k-SASA	residue-level solvent accessible surface area of the core k residues in the Fv substructure at T	nanometers squared (nm^2)
Fv	T = 300, 350, 400, all	surface k-SASA	residue-level solvent accessible surface area of the surface k residues in the Fv substructure at T	nanometers squared (nm^2)
Fv, H, L, CDRs, CDR1L, CDR2L, CDR3L, CDR1H, CDR2H, CDR3H	T = 300, 350, 400, all	SASA	solvent accessible surface area of the mAb substructure at T	nanometers squared (nm^2)

This includes all combinations of the descriptors calculated such as mAb substructure (left) and temperature (middle left). In addition, the descriptor abbreviations (middle), descriptor descriptions (middle right), and descriptor units (right) are provided.

selected by the best MOE score (default settings) and minimized in vacuo with the Amber10:EHT force field (19).

MD

Residue protonation states were determined by calculating pKa values using propka 3.1 (20,21) and residues considered deprotonated if pK_a was below physiological pH 7.4. The systems were solvated in water using the TIP3P water model (22) in rectangular water boxes large enough to satisfy the minimum image convention. Na^+ and Cl^- ions were added to neutralize charge and reach physiologic salt concentration ~150 mM. All simulations were performed utilizing GROMACS 5.4 (23) using the CHARM22 plus CMAP force field for proteins (sometimes referred to as CHARMM27) (24) and the orthorhombic periodic boundary conditions. All simulations were performed in full atomistic detail. The 25 Fvs are simulated for 100 ns at each temperature (300, 350, and 400 K) for a combined 7.5 μs of simulation time. The 100 ns simulation time is in concordance with other molecular simulation studies of Fvs (15,25,26) and agrees with experimental NMR timescale of Fv molecular tumbling time and loop dynamics (27,28). However, large molecular motions such as molecular unfolding and/or domain dissociation are likely only observed at the millisecond timescale (29).

MD simulations were performed in four steps for each Fv structure ($n = 25$) at each temperature (300, 350, and 400 K): 1) steepest descent energy minimization to ensure correct geometry and the absence of steric clashes, 2) 100 ps simulation in the constant particle, volume, and temperature ensemble (NVT) to bring atoms to correct kinetic energies, while maintaining temperature by coupling all protein and nonprotein atoms to separate baths using a velocity rescale thermostat (30) with a 0.1 ps time constant, 3) 100 ps simulation in the constant particle, pressure, and temperature ensemble (NPT) using Berendsen pressure coupling (30) and 2.0 ps time constant to maintain isotropic pressure at 1.0 bar, and 4) production MD simulations conducted for 100 ns with no restraints. To ensure true NPT ensemble sampling during 100 ns production simulations, the Nose-Hoover thermostat (31) and Parrinello-Rahman barostat (32) were used to maintain temperature and pressure, respectively. Time constants were 2.0 and 1.0 ps for pressure and temperature coupling, respectively, utilizing the isothermal compressibility of water 4.5e−5 bar^{-1}. Box size for equilibration simulations was approximately 5.8 × 7.5 × 7.3 nm^3 with ~9000 water molecules, ~60 ions, and ~30,000 total atoms. All simulations used the particle mesh Ewald algorithm (33,34) for long-range electrostatic calculations with cubic interpolation and 0.12 nm maximum grid spacing. Short-range nonbonded interactions were cut off at 1.2 nm using the Verlet cutoff scheme and all bond lengths were constrained using the LINCS algorithm (35) except water constrained using the SHAKE algorithm (36,37). The leap-frog algorithm was used for integrating the equations of motion with a 2-fs time-step. After the production runs, the descriptors were calculated from the trajectories after 20 ns. This equilibration time corresponded to the flattening of the root mean-square deviation (Fig. S1) and was greater than the average equilibration time across all simulations (7.6 ± 2.5 ns) determined by the Chodera algorithm (38), which assesses the variance-bias trade-off. Moreover, the equilibration detection found the following in equilibration times: 7.0 ± 2.8 ns at 300 K, 7.1 ± 2.5 ns at 350 K, and 8.7 ± 1.8 ns at 400 K. Thus, we utilized the 20–100 ns sampling window to provide sufficient equilibration time and to effectively capture the system dynamics of the measured descriptors.

Descriptor calculations

Several descriptors (Table 1) including solvent accessible surface area (SASA), number of hydrogen bonds, number of Lennard-Jones internal contacts, radius of gyration (R_g), and root mean-square fluctuations (RMSF), were evaluated by defining Gromacs index groups (gmx make_ndx) and using Gromacs-suite analysis tools (23) (i.e., gmx hbond, gmx rms, gmx rmsf, gmx sasa, gmx gyrate). The specified index group include the heavy (CDRH) and light (CDRL) chain of each Fv structure as well as the three CDRs on each chain as defined by canonical IMGT residue numbering (i.e., CDRH1, CDRH2, CDRH3, CDRL1, CDRL2, and CDRL3) (39,40). Internal contacts and hydrogen bonds are geometrically defined as donor-acceptor pair distances within 3.5 Å or a distance within

3.5 Å and an angle less than 30°, respectively. SASA is computed using the double cubic lattice method (41). Furthermore, we defined a k-SASA metric for the core and surface Fv residues by computing the residue-level SASA with the Shrake-Rupley algorithm (42). The core and surface residues were selected from the homolog structures and defined by the k residues with the least and most SASA, respectively (where k = 10, 15, 20, 25, …, 90, 95, 100). The k value is selected during feature selection, which allows one to independently quantify effects on the relevant Fv core and surface residues. The backbone N-H bond vector order parameters were calculated from

$$S^2 = \left(3 * \sum_{i=1}^{3}\sum_{j=1}^{3}\langle u_i u_j\rangle^2 - 1\right) \Big/ 2$$

where the i and j indices refer to the x, y, and z components of the bond vector scaled to unit magnitude (17,43,44). Order parameters are scaled by $\xi = (1.02/1.04)^6 \approx 0.89$ to account for zero-point vibrational motions (45,46). Angular brackets indicate averaging over a simulation block size. Prolines do not have a N-H bond vector and were excluded. The block size should reflect global tumbling time (∼10 ns) (47–49); however, inconsistency between NMR spectroscopy measured and MD measured tumbling times indicate that block size should be used as a fitting parameter because current explicit solvent models do not completely recapitulate this phenomenon (50,51). Thus, we elected to compute the order parameter at numerous block sizes (b = 2.5, 5, 7.5, 10, …, 50 ns) and the effective block size was determined during feature selection. The temperature dependence of the order parameter S^2 can be described by a dimensionless number Λ, which relates to the molecular heat capacity (52–54).

$$\Lambda = \frac{dln\,(1 - S)}{dln\,T}$$

The Λ values are determined from the slope of the ln(1−S) vs. ln T plots by linear regression. The quality of the fit is assessed by the coefficient of determination and included as an additional descriptor r-Λ. Experimental and simulated values of Λ can differ by up to a factor ∼2 even with highly correlated order parameters (52); however, this does not impact the approach presented if comparisons are made between Fvs under identical protocols. All descriptors' mean and standard deviation were computed after 20 ns of equilibration at 10 ps intervals. Data analysis was performed by standard python packages for data handling and visualization (i.e., numpy (55), pandas (56), matplotlib (57), scipy (58), Biopython (59), Anarci (40), MDAnalysis (60), MDTraj (61), and custom python scripts).

Regression metrics

The descriptors were evaluated with Pearson correlation coefficients, r_p. The Pearson correlation coefficient is a measure of linear correlation between two data sets that ranges [–1, 1].

$$r_p = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

The descriptors were filtered by a cutoff, $|r_p| > 0.45$, determined by a two-tailed t-test with t = 2.093, $\alpha = 0.05$, and $n = 19$ (assuming bivariate normal distributions):

$$r_p > \frac{t}{\sqrt{n - 2 + t^2}}$$

The r_p cutoff was first determined based on the training set ($n = 19$) and serves as a filtering step. The remaining correlated descriptors were then dropped if they had high cross correlation $r_p > 0.95$. Importantly, the r_p significance was recomputed and displayed in the manuscript to include the test set ($n = 25$, t = 2.06, and $|r_p| > 0.39$). This step prevents data leakage by not including descriptor information from the test set when filtering descriptors by Pearson correlation. To assess the statistical significance and interpretability of the descriptors' Pearson correlation, 95% confidence intervals are computed by jackknife resampling (62).

After cross correlation consideration, the remaining descriptors are then sequentially sent to recursive and exhaustive feature selection. Overall, the feature selection methodology aims to improve interpretability by compressing the descriptors sets into a minimum description length. Moreover, the descriptor filters reduce multicollinearity in linear models (e.g., linear regression, elastic net regression) (63) and reduce descriptor ranking instability in ensemble models (e.g., random forest, adaboost, xgboost) (64–66). After feature selection, the regressors were evaluated utilizing repeated leave-one-out cross-validation (rLOOCV) (n_repeats = 3), which is often used for small data sets because it provides maximum cross-validation of each model (n–1 validation sets). The hyperparameters of the regressors were assessed based on the mean absolute error (MAE).

After the validation stage, the regressors' performance was assessed on the test set by computing the squared Pearson correlation coefficient, r_p^2, and the coefficient of determination, R^2. The r_p^2 ranges from [0,1] and is indicative of the linear relationship between the predictions and the measurements.

$$r_p^2 = \left(\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}\right)^2$$

where x_i and y_i are the measured value and predicted value, respectively. Similarly, the average value is represented as $\bar{x} = \frac{1}{n}\sum_{i=1}^{N} x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{N} y_i$. In general, linear correlation between model predictions and measurements is favorable; however, r_p^2 is invariant to scale and shift. Thus, r_p^2 is not an ideal goodness-of-fit metric in the case that predictions are shifted or scaled from parity (67–69). Thus, we also include R^2, which measures the proportion of variance that can be explained by the model.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(x_i - y_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

The R^2 ranges $(-\infty, 1]$ and assesses the predictive power of the model by computing the deviation from parity where $R^2 \leq 0$ represents a model that predicts the average value of the distribution ($R^2 = 0$) or arbitrarily worse ($R^2 < 0$).

The regressors' 95% confidence interval was generated using a jackknifing resampling technique (62). For small data sets ($n = 18/19$), the regressors were trained on all leave-one-out combinations with the Bayesian optimized hyperparameters. Then, test set performance was evaluated by R^2 and r_p^2 metrics for all leave-one-out regressors. The confidence interval was determined by computing the standard error of the regression metric and multiplying by the critical t value for a 95% confidence level in a two-tailed t-distribution (t = 1.96). Similarly, for the larger data sets ($n = 514/522/538$), the data was k-folded (k = 19), and the 95% confidence intervals were analogously generated.

Descriptor selection and ML

The feature selection process can be summarized as a four-step filter mechanism similar to previous work (70): 1) only the features with correlation above the cutoff ($|r_p| > 0.45$, assume bivariate normal distributions) to the endpoint (T_{agg}, $T_{m,on}$, or T_m) were retained, 2) features with high cross correlation ($r_p > 0.95$) were presumed redundant and dropped, 3) the remaining features were recursively selected using grid search random forest with

cross-validation (cv_splits = 3, cv_repeats = 3) and ranked using random forest feature importance, and 4) the top 10 ranked features are exhaustively searched (i.e., 10 chose 1–5) for the best combination of features. We ensured the numbers of features selected ($\leq$5) was less than the number of datapoints in the test set ($n = 6$) to minimize overfitting and to obtain a minimum description length. Importantly, for this initial data set, some Pearson correlations are relatively sensitive to datapoint removal which may cause the ranking of descriptors by Pearson correlation to change when scaling to larger data sets. Therefore, the descriptor correlations are only considered statistically significant and interpretable if the correlations are above the significance cutoff ($n = 25$, $|r_p| > 0.39$) with jackknife resampled 95% confidence interval. Although the linear descriptor ranking and sensitivity of data may change with larger data sets, we believe the overall selection methodology described is robust to spurious linear correlations because it is based on performance from exhaustive search and random forest feature importance, which considers nonlinear effects.

Next, the selected features were trained on eight scikit-learn (71) regressors using rLOOCV, hyperparameter searched (skopt) (72), ranked (by MAE), and refit. The regressors included were linear regression, elastic net, support vector machine, k nearest neighbors, decision tree, random forest, adaboost, and xgboost (73). The best hyperparameter set for each regressor was selected based on the MAE from rLOOCV (n_repeats = 3). Bayesian optimization was performed for each regressor (100 iterations with 4 points per iteration) over a defined hyperparameter space utilizing the scikit-optimize package (72). Exhaustive feature selection was performed using the mlxtend package (74). Before regression, all thermostability endpoints were normalized by the mean and standard deviation to improve the invariance of the regressors to disparate instrumentation and experimental conditions. The final regressor was selected based on the highest coefficient of determination (R^2) on the test set. Models trained on MD descriptors are denoted AbMelt (MD + ML) throughout the manuscript.

Baseline models

Two sequence baseline models were assessed: one-hot encoding and AbLang embeddings. First a multiple sequence alignment (MSA) for both the heavy and light chains were independently generated using IMGT numbering with MOE software (75). One-hot encoding is derived by encoding each amino acid as a 21-length vector where the i^{th} position corresponds to each of the 20 amino acids, with the final position in each vector corresponding to a gap. These vectors were then concatenated to form a single L $\times$ 21 length feature tensor to represent each sequence, where L is the length of the MSA. AbLang (76) is a transformer-based language model and descendant of BERT (77) that is pretrained on a database of antibody sequences, the Observed Antibody Space (78). The sequence embeddings from AbLang are generated from the pretrained model by inputting the sequence and extracting the last hidden layer. This sequence representation has dimension L $\times$ 768, where L is the length of the MSA. These inputs were used as features for a random forest regressor.

The Fv 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 (75), length of CDRs, potential energy, ASPmax (79), mono/dipole/quadrupole moments (80), isoelectric point (81), mass, etc. (140 total). Descriptors were computed after homology modeling and energy minimization. The same descriptor selection process was performed. Models trained on MOE descriptors are denoted MOE + ML throughout the manuscript.

Data set split

The 25 Fvs selected in this study (Fig. S2, *A–C*) were from a larger internal experimental thermostability data set ($n \sim 500$) where T_{agg} (77.2 $\pm$ 7.91°C), $T_{m,on}$ (60.9 $\pm$ 3.95°C), and T_m (68.1 $\pm$ 3.78°C) represent the distributions of measured experimental values (Fig. S2, *D–F*). The training set size, n, was selected to approximate these distributions ($n = 18/19$): T_{agg} (75.4 $\pm$ 12.5°C, $n = 18$), $T_{m,on}$ (56.0 $\pm$ 9.55°C, $n = 19$), and T_m (64.9 $\pm$ 9.02°C, $n = 19$) (Fig. S2). The sample distributions were identical to the larger distributions by the two-sample Kolmogorov-Smirnov test ($p < 0.01$) (Fig. S2, *A–C*). One temperature of aggregation datapoint was not measured (described previously in methods). The internal data set also contained a small collection of reproduced mAbs from Jain et al. (8), who reported T_m values for ~130 clinical mAbs. Thus, to further validate our model, we selected a held-out test set by optimizing a cost function that simultaneously selected mAbs that are publicly available in Jain et al. (8) and are representative of the mean and standard deviation of the property endpoint. This resulted in a final test set of six datapoints that include four publicly available Fvs with reported T_m values (8) that were experimentally reproduced with a high correlation in measured T_m ($r_p = 0.94$) (Fig. S3). Before regression, all thermostability endpoints were normalized by the mean and standard deviation (Fig. S2). The train and test sets for T_{agg}, $T_{m,on}$, and T_m were identical for all methods trained on the small training set ($n \sim 20$). For the larger training set ($n \sim 500$), the test set remained the same.

RESULTS

The Fvs of 25 mAbs were homology modeled using the MOE Antibody Modeler application and energy minimized (18). MD was performed using GROMACS (23) at three temperatures (300, 350, and 400 K) for 100 ns on each Fv homology model to generate structural ensembles that represent distinct stages along the temperature ramp performed during experimental thermostability measurement. Descriptors from the MD trajectories were computed, and the mean and standard deviation were assessed for each structure and temperature after 20 ns of equilibration. This equilibration point corresponded to the flattening of root mean-square deviation (Fig. S1) and was greater than the point determined by the Chodera algorithm (7.6 $\pm$ 2.5 ns) (38). This equilibration detection algorithm maximizes the number of time-uncorrelated samples in the sampling window, removes spurious samples at the beginning of the production simulations, and automates detection of an approximate descriptor sampling window across many simulations. By taking a 20 ns equilibration point, we minimize spurious sampling and maintain a time-consistent descriptor sampling window across all simulations.

Descriptor correlations and selection

The descriptors from the MD trajectories included RMSF, R_g, SASA, number of the hydrogen bonds, number of internal contacts, and the N-H bond vector order parameter (S^2). In addition, the mean, standard deviation, and temperature derivative were included as separate descriptors. These descriptors were also measured for both the Fv structure as well as Fv substructures: heavy chain (H), light chain (L), and six CDRs (i.e., CDR H1, H2, H3, L1, L2, and L3) (Table 1). We found that the standard deviation of the CDR R_g at 400 K had the highest correlation with T_{agg} ($r_p = -0.68 \pm 0.23$) (Fig. 2 *A*). Interestingly, the standard deviation of internal contacts at 350 K

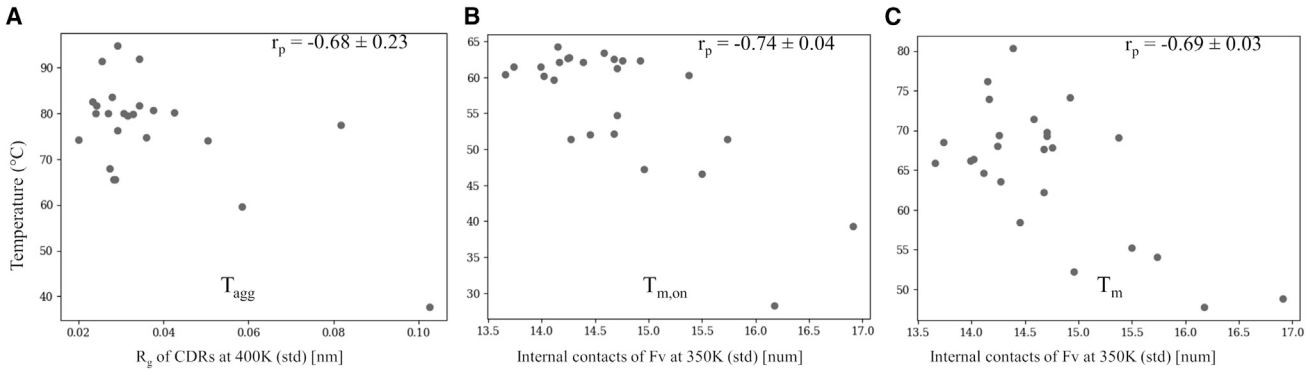

FIGURE 2 Descriptor correlations measured from molecular dynamics. This includes the most highly correlated feature for (*A*) aggregation temperature, (*B*) melting temperature onset, and (*C*) melting temperature. The *x* axis is the descriptors, and the *y* axis is the thermostability endpoints in °C. The Pearson correlation coefficient (r_p) is displayed in the top right of each panel. The Pearson correlation 95% confidence intervals are computed by jackknife resampling.

had the highest correlation with $T_{m,on}$ and T_m ($r_p = -0.74 \pm 0.04$ and $r_p = -0.69 \pm 0.03$), respectively (Fig. 2, *B* and *C*).

A Pearson correlation coefficient cutoff between the descriptor and the thermostability endpoint (i.e., $|r_p| > 0.45$) was determined with a two-tailed significance test ($n = 19$, $t = 2.093$, $\alpha = 0.05$, assume bivariate normal distributions) (described in methods). In addition, the descriptors with high cross correlation ($r_p > 0.95$) (Fig. S4) were eliminated. These Pearson correlation cutoffs serve as filtering steps and resulted in a total descriptor set of 11, 33, and 10 for T_{agg}, $T_{m,on}$, and T_m, respectively. The correlation cutoff of descriptors was first determined based on the training set ($n = 19$) then recomputed with the test set to determine the significance of the descriptor correlations ($n = 25$, $t = 2.06$, and $|r_p| > 0.39$). This step prevents data leakage by not including descriptor information from the test set when filtering descriptors by Pearson correlation. The descriptors were only considered significant and interpretable if the correlations were above the $|r_p| > 0.39$ cutoff with jackknife resampled 95% confidence interval.

After cross correlation filtering, the descriptors were reduced by performing random forest recursive feature elimination with cross-validation (n_folds = 3, n_repeats = 3) and ranked by feature importance. The top 10 ranked descriptors are then selected and exhaustively searched using a random forest regressor for the best combination of descriptors (10 choose 1–5) (74). We found that the best combination of descriptors for T_{agg} included the highest correlated descriptor, standard deviation of the CDR R_g at 400 K ($r_p = -0.68 \pm 0.23$), as well as the mean RMSF of the CDR atoms ($r_p = -0.62 \pm 0.07$) and the temperature derivative of the N-H bond vector order parameter, Λ ($r_p = -0.46 \pm 0.18$) (Fig. 3 *A*). Similarly, the standard deviation of internal contacts at 350 K was selected in the best descriptor set for $T_{m,on}$ ($r_p = -0.74 \pm 0.04$) as well as the fitness metric for the temperature derivative of the N-H bond vector order parameter, *r*-Λ ($r_p = 0.63 \pm 0.11$). In addition, the temperature derivative of Fv core residue SASA, mean ($r_p = 0.63 \pm 0.10$), and standard deviation ($r_p = 0.58 \pm 0.09$), were selected in the $T_{m,on}$ descriptor set (Fig. 3 *B*). Finally, the standard deviation of internal contacts at 350 K ($r_p = -0.69 \pm 0.03$), the standard deviation in the CDR R_g at 350 K ($r_p = 0.44 \pm 0.07$), and the standard deviation in the RMSF of the CDR L1 atoms at 350 K ($r_p = 0.39 \pm 0.06$) were selected as the best descriptor set for T_m (Fig. 3 *C*).

Training, validation, and test set performance

To assess the performance of the selected descriptor sets, we trained 8 scikit learn (71,82) regressors using rLOOCV and optimized each regressors' hyperparameters using the scikit-optimize package (72). The performance of the models was scored based on MAE from rLOOCV and used to predict on the test set. We found that a k-nearest neighbor regressor performed best on the T_{agg} training set ($n = 18$) with a MAE 0.72 ± 0.78 (Fig. S5 *A*). In addition, we found an elastic net regressor performed best on the $T_{m,on}$ training set ($n = 19$) with a MAE 0.61 ± 0.47 (Fig. S5 *B*). Finally, a random forest regressor performed best on the T_m training set ($n = 19$) with a MAE 0.71 ± 0.56 (Fig. S5 *C*). After rLOOCV, the regressors were refit and all achieved a $R^2 > 0.7$ and $r_p^2 > 0.64$ on the training sets (Fig. S5, *A*–*C*).

After training, we assessed the performance of the best regressors on the test set. In addition, the 95% confidence intervals were computed using a jackknifing resampling technique (62). Remarkably, with only 3 descriptors and 18 training datapoints, the T_{agg} regressor maintains predictive power on the test set with $R^2 = 0.57 \pm 0.11$ and $r_p^2 = 0.71 \pm 0.09$ (Fig. 3 *A*). Correspondingly, the $T_{m,on}$ regressor maintains predictive power on the test set with $R^2 = 0.56 \pm 0.01$ and $r_p^2 = 0.61 \pm 0.0003$ (Fig. 3 *B*). In addition, the T_m regressor maintains predictive power on the test set with $R^2 = 0.60 \pm 0.06$ and $r_p^2 = 0.64 \pm 0.04$ (Fig. 3 *C*). To assess the AbMelt MD plus ML models (MD + ML), we compared the performance with several baseline models.

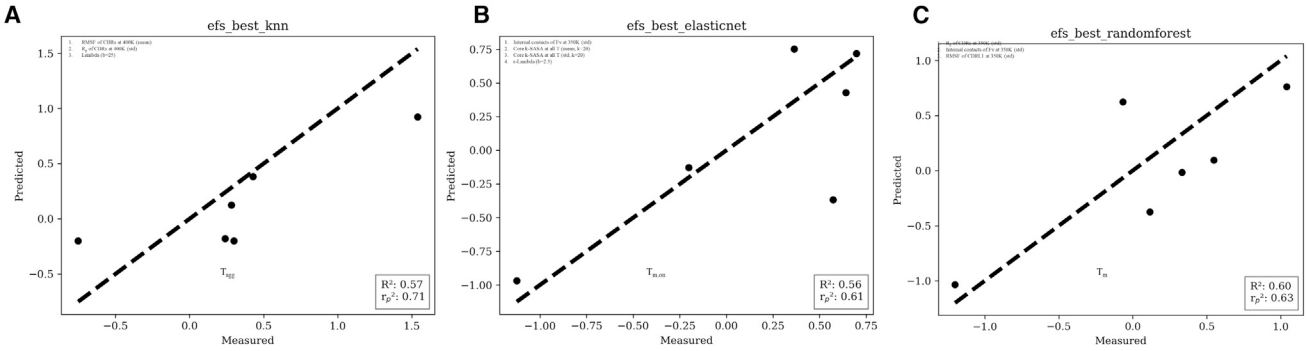

FIGURE 3 Prediction on the test set of the models regressed with molecular dynamic descriptors. This includes the most highly correlated feature for (*A*) aggregation temperature, (*B*) melting temperature onset, and (*C*) melting temperature. The exhaustively selected features are displayed in the top left of each panel ranked by random forest feature importance. The *x* axis is the normalized measured values, and the *y* axis is the normalized predicted value of each thermostability endpoint. The dashed line is the parity line ($y = x$). The coefficient of determination (R^2) and squared Pearson correlation coefficient (r_p^2) are displayed in the bottom right of the panels. The title of each panel includes the best regressor selected from exhaustive feature selection.

Baseline model performance

The performance of the AbMelt (MD + ML) regressors were evaluated against several sequence and structure baselines. The sequence baseline models included random forest regressors trained on one-hot sequence encodings and AbLang sequence embeddings (76). AbLang is a protein language model pretrained on a database of antibody sequences from the Observed Antibody Space (78). In addition, we included a set of 140 sequence/structure descriptors calculated in MOE and executed the same descriptor selection methodology (MOE + ML). When trained on the same training set, we found that both one-hot and AbLang have a very low coefficient of determination on the test set ($R^2 \leq 0.14$) (Fig. 4). Similarly, with MOE + ML we achieved limited performance ($R^2 \sim 0.20$) (Fig. 4). For example, the best feature set for T_{agg} included the heavy chain bend angle (75), the 3D-based isoelectric point (81), and the CDR hydrophobic patch area (Fig. S6 *A*). This decision tree regressor achieved limited predictive power on the T_{agg} test set with $R^2 = 0.21 \pm 0.10$ and $r_p^2 = 0.53 \pm 0.09$. For $T_{m,on}$, the potential energy, zeta quadrupole moment (80), and total hydrophobic SASA were selected in the descriptor set and, again, the regressor achieved limited predictive power on the $T_{m,on}$ test set with $R^2 = 0.23 \pm 0.06$ and $r_p^2 = 0.34 \pm 0.09$ (Fig. S6 *B*). Correspondingly, the Fv mass and heavy/light chain torsion angle (75) were selected in the T_m descriptor set and the random forest regressor achieved a performance of $R^2 = 0.19 \pm 0.03$ and $r_p^2 = 0.23 \pm 0.02$ (Fig. S6 *C*).

In addition, we assessed the performance of AbMelt (MD + ML) ($n \sim 20$) against baseline models trained on a much larger internal data set of the thermostability endpoints ($n \sim 500$). Remarkably, we found that, measured by R^2, AbMelt (MD + ML) outperforms all the baseline models despite being trained on <5% of the data (Fig. 4). For example, one-hot ($n = 514$) is the best performing T_{agg} baseline with $R^2 = 0.35 \pm 0.05$ and $r_p^2 = 0.61 \pm 0.06$ (Fig. S7 *A*). Moreover, MOE + ML ($n = 522$) is the best performing $T_{m,on}$ baseline with $R^2 = 0.18 \pm 0.04$ and $r_p^2 = 0.41 \pm 0.04$ (Fig. S7 *B*). For the T_m baseline, one-hot ($R^2 = 0.45 \pm 0.02$ and $r_p^2 = 0.90 \pm 0.02$) and AbLang ($R^2 = 0.45 \pm 0.02$ and $r_p^2 = 0.86 \pm 0.01$) (Fig. S7 *C*) achieved similar performance. While baseline sequence models achieved high r_p^2 on the larger data set, their corresponding R^2 performed worse.

DISCUSSION

We found that, despite the high computational cost of performing high temperature MD, there is rich and robust dynamical information that can be learned directly from molecular simulations to predict experimental thermostability measurements. Moreover, despite training on a small set of Fvs, AbMelt (MD + ML) outperformed all baseline models. This includes one-hot encoding, AbLang sequence embeddings, and MOE descriptors + ML, which are descriptors with low computational cost. Remarkably, this performance is maintained even when low computational cost descriptors are trained on a 20-fold larger data set. These results underscore the importance of quantifying intrinsic mAb flexibility when learning to predict thermostability.

Descriptors

The MD descriptors were computed after 20 ns of equilibration time at each temperature (300, 350, and 400 K). This cutoff is reasonable given the root mean-square deviation plots (Fig. S1) and the automated equilibration detection results (38); however, these simulations are unlikely to be in thermodynamic equilibrium because unfolding timescales are on the order of microseconds to milliseconds (83,84). Moreover, complete Fv unfolding in an unperturbed simulation is not tractable on the nanosecond timescale and was not observed. The aim of this work was not to observe

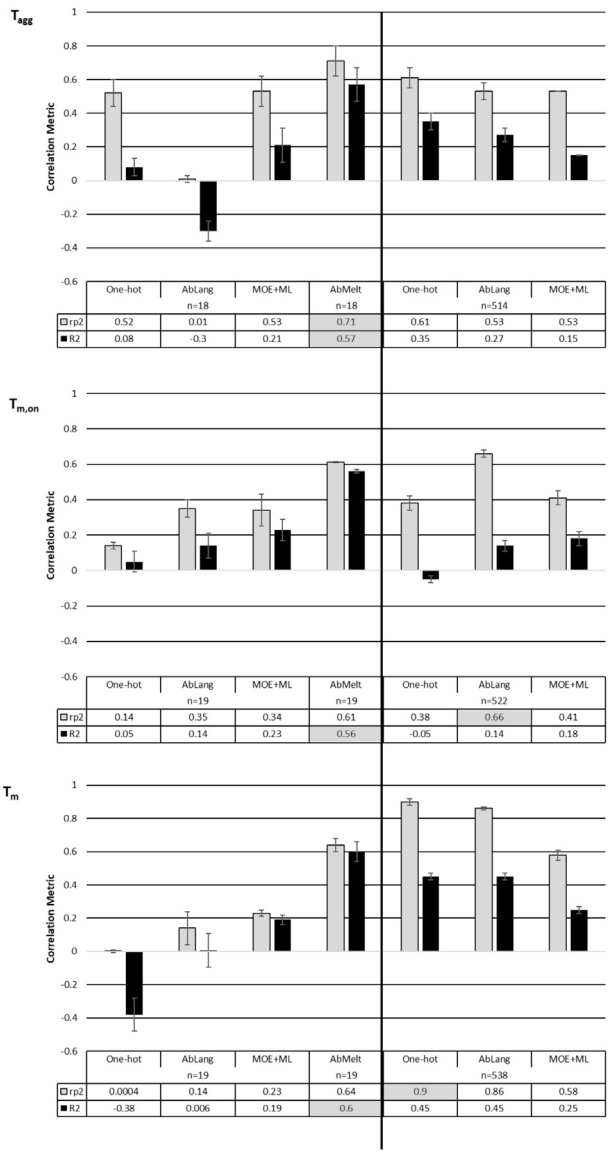

FIGURE 4 Performance of regressors and baseline models on the test set. Graphs are organized by thermostability endpoints (T_{agg}, $T_{m,on}$, or T_m). Performance metrics are rendered in black for coefficient of determination (R^2) and gray for the squared Pearson correlation coefficient (r_p^2). Bars are clustered by model and number of datapoints in the training set (n). The regressors trained on the small data set ($n \sim 18$) are on the left side of the charts, while the regressors trained on the larger data set ($n \sim 520$) are on the right. AbMelt is presented in the center of the charts ($n \sim 18$). The highest-performing metric trained is displayed in tables with a gray background. The error represents the 95% confidence interval after jackknife resampling. Note that AbMelt, despite having a small training set, out-performed all models with respect to R^2. Also, while one-hot and AbLang with large training sets performed better with respect to r_p^2 both fail when R^2 is considered.

unfolding, but rather to measure relative disorder in Fv structure across the sequence and temperature dimensions, and compute descriptors that can be compared, correlated, and regressed to their respective thermostability measurements. Indeed, we found numerous descriptors that significantly correlate with T_{agg}, $T_{m,on}$, and T_m. Moreover, the descriptor correlations are only considered significant and interpretable if the correlations are above the $|r_p| > 0.39$ cutoff with the jackknife resampled 95% confidence interval.

This included the standard deviation of the CDR R_g at 400 K, which had the highest correlation with T_{agg} ($r_p = -0.68 \pm 0.23$). This descriptor correlation indicates that Fv structures that have highly flexible CDRs at 400 K, have a lower T_{agg}. CDR flexibility likely increases the binding pose manifold, thus increasing the probability of self-interactions. This is further supported by the mean RMSF of the CDR atoms ($r_p = -0.62 \pm 0.07$) being added to the descriptor set after exhaustive feature selection. Finally, the temperature derivative of the N-H bond vector order parameter, Λ ($r_p = -0.46 \pm 0.18$), was selected in exhaustive feature selection to predict T_{agg}.

The $T_{m,on}$ represents the red shift onset in the F350 nm/F330 nm spectra caused by tryptophan being unburied from the protein core during the melting process. Unsurprisingly, the standard deviation in internal contacts at 350 K had the highest correlation with $T_{m,on}$ ($r_p = -0.74 \pm 0.04$). This indicates that the $T_{m,on}$ is strongly governed by the intrinsic ability of the Fv structure to maintain time invariant contacts at high temperature. Similarly, the temperature dependence of Fv core residues' SASA ($k = 20$) was selected during exhaustive feature selection. This descriptor is indicative of Fv core flexibility where the mean ($r_p = 0.63 \pm 0.10$) and standard deviation ($r_p = 0.58 \pm 0.09$) are correlated with $T_{m,on}$. This is consistent with the deviation in internal contacts. In addition, the fitness metric of the temperature derivative of the N-H bond vector order parameter, $r\text{-}\Lambda$ ($r_p = 0.63 \pm 0.11$), was selected, which similarly indicates that, with stronger Fv order temperature dependence, the Fv is more likely to begin melting at a lower temperature.

T_m represents the midpoint of unfolding. Similarly, the standard deviation in internal contacts at 350 K had the highest correlation with T_m ($r_p = -0.69 \pm 0.03$). This descriptor is also consistent with $T_{m,on}$. Interestingly, the standard deviation of the CDR R_g at 350 K ($r_p = 0.44 \pm 0.07$) and the standard deviation in RMSF of CDRL1 atoms at 350 K ($r_p = 0.39 \pm 0.06$) were selected during exhaustive feature selection. This reduced CDR flexibility occurs at the same time as an increase of overall internal contact fluctuations for Fvs with lower T_m (Fig. 2 *C*). Because these descriptors were selected based on performance in random forest exhaustive feature selection, the combinatoric effects of these descriptors cannot be easily interpreted linearly. Interestingly, the descriptors selected for T_{agg}, $T_{m,on}$, and T_m corresponded to their measured temperature ranges. For example, observed $T_{m,on}$ 334 ± 4 K (60.9 ± 3.95°C) and T_m 342 ± 4 K (68.1 ± 3.78°C) measurements had the highest correlation with 350 K descriptors. Similarly, observed T_{agg} measurements were 350 ± 8 K (77.2 ± 7.91°C) and correlated best with 400 K

descriptors. These results indicate that more signal is extracted from MD descriptors measured at temperatures at least two standard deviations higher than the observed average. This is likely because measuring descriptors at the average temperature will exclude measurable effects on the upper half of the distribution. Future work may elucidate more precise temperatures that will maximize the correlation between MD descriptors and measured thermostability endpoints.

To understand the effect of the initial structure on MD descriptor calculation, we compared the results of the epratuzumab crystal structure (PDB: 5VKK) (85), the MOE homology model, and the ImmuneBuilder model (86). We found that the initially generated structures (Fig. S8 *A*) deviated 0.88 and 0.55 Å from the crystal structure for MOE and ImmuneBuilder, respectively. Next, we computed the descriptor sets from the MD trajectories after 100 ns runs at 300, 350, and 400 K. We found that, although the predicted structures deviated from the crystal structure (<1 nm), this had no effect on the descriptor calculations (Fig. S8, *C* and *D*). For example, we found no significant difference in the CDR RMSFs at 400 K between the crystal, MOE, or ImmuneBuilder initial structures (Fig. S8 *C*). In addition, we found no significant difference in the number of internal contacts at 350 K between any initial structure (Fig. S8 *D*). These results demonstrate that AbMelt is robust to the initial starting structure if the structure is experimentally determined, or the structure prediction method is well validated (i.e., <1–2 Å root mean-square deviation over the Fv C^{α} coordinates).

Performance

Once the descriptor sets were selected, we trained eight regressors using rLOOCV, and the best regressor was selected by MAE (Fig. S5). We assessed the final performance of all models on a test which contains four publicly available clinical Fvs that were internally reproduced (Fig. S3). Interestingly, the T_{agg} regressor maintained a $R^2 = 0.57 \pm 0.11$ (Fig. 3 *A*) on the test set. This outperformed the one-hot $R^2 = 0.08 \pm 0.05$, AbLang $R^2 = -0.30 \pm 0.06$, and MOE + ML $R^2 = 0.21 \pm 0.10$ baseline models that were trained on the same set of Fvs (Fig. 4). This observation remains consistent for both the $T_{m,on}$ and T_m regressors. In addition, all AbMelt (MD + ML) regressors outperformed their best correlated descriptor in terms of r_p, demonstrating the predictive value in combining multiple features (Figs. 1 and 4). Overall, when trained on the same data set, the AbMelt (MD + ML) regressors outperformed all baseline models on 6/6 (2/2, 2/2, and 2/2) regression metrics for T_{agg}, $T_{m,on}$, and T_m, respectively. This indicates that inherent Fv flexibility can be measured in high-temperature MD and that the rich information from molecular simulation data is useful in predicting experimental thermostability measurements.

Due to the computational expense of performing three temperature MD, we elected to assess the limitations of the high-cost descriptors by comparing the performance with baseline models trained on much larger data sets. Strikingly, although some baseline models improved, AbMelt (MD + ML) still outperforms all baseline models (Fig. 4). For example, one-hot improved from an $R^2 = 0.08 \pm 0.05$ to an $R^2 = 0.35 \pm 0.05$ on the test set when increasing the data set from $n = 18$ to $n = 514$ for T_{agg} (Fig. S7 *A*). Likewise, AbLang improved from $R^2 = -0.30 \pm 0.06$ to $R^2 = 0.27 \pm 0.04$. Despite the improvement, the AbMelt (MD + ML) T_{agg} regressor outperformed these baseline models ($R^2 = 0.57 \pm 0.11$) with <5% of the training data (Fig. 4). In contrast, $T_{m,on}$ achieved no improvement when increasing the data set size. For example, one-hot decreased from an $R^2 = 0.05 \pm 0.06$ to an $R^2 = -0.05 \pm 0.02$ on the test set when increasing the data set from $n = 19$ to $n = 522$. Similarly, AbLang and MOE + ML (Fig. S7 *B*) stayed the same or slightly decreased with an increased $T_{m,on}$ data set (Fig. 4). For T_m, all baseline models improved when increasing the data set from $n = 19$ to $n = 538$ with one-hot and AbLang achieving an $R^2 = 0.45 \pm 0.02$ and 0.45 ± 0.02 (Fig. S7 *C*), respectively. However, this increased performance still did not outperform the AbMelt (MD + ML) T_m regressor ($R^2 = 0.60 \pm 0.06$) trained on a ~20-fold smaller data set (Fig. 4). While baseline models were able to achieve high r_p^2 for T_m, the Pearson correlation metric is scale/shift invariant and thus not an ideal goodness-of-fit metric for predicting the measured T_m values. This is demonstrated by a smaller R^2 on the test set compared with AbMelt (MD + ML) (Fig. 4). Moreover, this is depicted by a large deviation from parity where the predicted values are in a relatively flat line for the regressors trained on the larger data set ($n \sim 500$): one-hot T_{agg} (Fig. S7 *A*), MOE + ML $T_{m,on}$ (Fig. S7 *B*), and AbLang T_m (Fig. S7 *C*).

The sequence-based models generally increased performance with data set size, suggesting that there is meaningful information in the sequence representations. However, the ability to extract this implicit information from the sequence representations requires at least a 20-fold larger data set compared with structural or dynamical descriptors (Fig. 4). In addition, we did not observe a significant performance improvement of AbLang sequence embeddings compared with one-hot sequence encodings at either data set size. These results suggest that the AbLang sequence embeddings do not contain any additional information beyond the sequence itself at predicting thermostability measurements. Overall, despite being trained on <5% of the data, AbMelt (MD + ML) outperformed all baseline models on 4 of 6 (2 of 2, 1 of 2, and 1 of 2) regression metrics for T_{agg}, $T_{m,on}$, and T_m, respectively. This includes robust AbMelt (MD + ML) model performance (3 of 3) measured by the coefficient of determination, R^2, which is used to assess the predictive power of the model to predict the measured T_{agg}, $T_{m,on}$, and T_m values.

Importantly, we have made the trained T_{agg}, $T_{m,on}$, and T_m regressors available to users to infer thermostability endpoints. To our knowledge, only T_m data have been reported for antibodies (8, 9) and these data contain clinical stage antibodies with ideal thermostability properties. For example, the Jain et al. data set does not contain any mAbs with $T_m < 60°C$ (Fig. S9 *A*). Comparatively, the Shehata et al. data contain 6 mAbs with $T_m < 60°C$. To determine if a lack of low thermostability antibodies effects the ability to train an accurate T_m regressor, we computed MD descriptors for an additional 7 mAbs from the Jain et al. data set and 6 mAbs from Shehata et al. with low thermostability (i.e., $T_m < 60°C$). We used these 13 additional mAbs to train T_m regressors with the same MD descriptors (Fig. 3 *C*) and tested performance on the same test set after normalization. Interestingly, we found that including antibodies with low thermostability is crucial in training accurate T_m regressors (Fig. S9 *B*). For example, the random forest regressor trained only on Jain et al. ($n = 7$) achieved a $R^2 = -0.87$ and $r_p^2 = 0.44$ on the test set; however, this performance was greatly improved when including the Shehata et al. data, reaching $R^2 = 0.48$ and $r_p^2 = 0.55$ (Fig. S9 *B*). This indicates that the ability to learn T_m from MD descriptors is dependent on having a wide coverage of the thermostability property distribution including molecules with poor thermostability properties.

In this study, the mAb models only included the Fv region. The Fv region of the mAb was selected because our data set only contains IgG1 frameworks and sequence diversity is primarily in the Fv region. Moreover, this selection reduces the computational cost by reducing the number of atoms required in the simulation box: Fv ∼30,000 atoms, Fab ∼60,000 atoms, and Ig ∼300,000 atoms (N log N scaling where N is the number of atoms). However, future work may benefit from including Fab or Ig residues when warranted by shifts in the Fc that effect thermostability. In addition, tethered dynamics between the Fv and Fc may uncover performance improvement by learning sequence/structural/dynamical information contained in the additional Fab and Ig residues. Future research directions also include combining additional descriptors from separate methodologies, using only experimental starting structures, and extending simulations into the microsecond regime. Importantly, extending to the microsecond regime may boost performance by regressing additional descriptors only observed at longer timescales such as large-scale loop rearrangements and VH-VL domain dissociation (29,87,88). We further found that the top correlated descriptors converged in Pearson correlation coefficient after 80 ns of simulation time (Fig. S10). This may indicate that AbMelt equivalent performance may be achieved with less than 100 ns simulations. Overall, these results demonstrate robust information in molecular simulation data that is not easily extracted from static sequence/structure representations. The ability to collect dynamical descriptors, however, comes at a computational cost. Given the estimated cloud compute vendor prices we can estimate this cost (∼$3/h on an A100 GPU and ∼$0.03/core-hour on an AMD EPYC 64-core CPU). MD performance with 30,000 atoms can be approximated (A100 GPU ∼1000 and ∼130 ns/day on AMD EPYC 64-core CPU) and the total simulation time for 500 Fvs is ∼150,000 ns (500 Fvs × 3 temperatures × 100 ns per Fv per temperature). Therefore, on modern hardware, the cost to perform three-temperature MD on 500 Fvs costs ∼$11,000 on A100 GPUs and ∼$50,000 on AMD EPYC 64-core CPUs. This cost may be further reduced by bandit or opportunistic provisioning approaches (89–91). Of course, there is a trade-off in balancing the capital expense required to compute high-cost dynamical descriptors at scale and the ability to implicitly extract this information from larger data sets. Ultimately, this trade-off will depend on the relative expense of compute and experiment, but this work clearly demonstrates that molecular simulation data are useful in predicting antibody thermostability measurements. Moreover, given modern thermostability data set sizes ($\sim 10^2$–10^3), this computational cost may well be worth the capital investment to eliminate poor candidate molecules and reduce the ∼$2.6 billion cost to bring a therapeutic to market.

CONCLUSION

We performed high-temperature MD simulations to quantify the conformational flexibility of antibodies and show that descriptors measured from simulations correlate with thermostability endpoints and are useful for predicting T_{agg}, $T_{m,on}$, and T_m. The predictive capability of the high-cost MD descriptors compared with several low-cost descriptor baseline models demonstrates that the MD descriptors exhibit robust performance and the additional compute cost is justifiable. Moreover, AbMelt models remained highly predictive on held-out, publicly available sequences and their measured T_m. This performance-edge was maintained despite making available 20-fold larger training data sets to the traditional ML models. This work establishes the utility in directly quantifying Fv flexibility from molecular simulation data to predict thermostability measurements. The potential savings by eliminating even a few poor molecules in the early development stage justifies the computational cost of simulating the MD of the molecules. To that end, AbMelt regressors can be refined by producing the dynamical descriptor data sets at scale and inferring mAb thermostability measurements on prospective candidate molecules. We acknowledge the limitation that the reliability of performance metrics on small train and test sets is not ideal and that performance assessment on a larger data set is a clear next step. In addition, while our initial work has proven successful considering only the Fv, work is underway to examine the impact on inclusion of the complete

Fab and/or Ig as well as additional molecular descriptors and any impact they may have on regressor performance.

DATA AND CODE AVAILABILITY

The descriptor data sets for AbMelt (MD + ML) and MOE + ML are made available at Zenodo: https://doi.org/10.5281/zenodo.10815667. We have released a package to predict thermostability measurements for Fvs by providing the code, trained models, and descriptor data sets. The mAb sequences and structures are proprietary data; however, the provided descriptor sets and trained regressors ensure complete reproducibility. In addition, we have made available all starting structures generated with public sequences including all the structures used to train a T_m regressor trained with public T_m data.

MOE may, optionally, be licensed from Chemical Computing Group https://www.chemcomp.com, although it is not required for the protocol.

SUPPORTING MATERIAL

Supporting material can be found online at https://doi.org/10.1016/j.bpj.2024.06.003.

AUTHOR CONTRIBUTIONS

Z.A.R. performed simulations, analyzed and interpreted the simulation data, and wrote the manuscript. T.W. analyzed and interpreted simulation data and contributed to the manuscript. A.C.C. designed experiments, analyzed and interpreted simulation data, and wrote the manuscript. E.M. designed experiments, analyzed and interpreted simulation data, and contributed to the manuscript.

ACKNOWLEDGMENTS

We acknowledge the contributions of members of the Protein Sciences Department within Discovery Biologics at Merck & Co., Inc., South San Francisco, CA, and especially Drew Waight, Marc Bailly, and Laurence Fayadat-Dilman. We also acknowledge the contributions of members of the Biologics Process R&D and Sterile Formulation Sciences Departments within Pharmaceutical Sciences at Merck & Co., Inc., Rahway, NJ, and members of the Modeling & Informatics group within Discovery Chemistry at Merck & Co., Inc., South San Francisco, CA, especially BoRam Lee, Jingzhou Wang, Tanmoy Pal, and Katherine Delevaux.

DECLARATION OF INTERESTS

The authors declare no competing financial interest.

REFERENCES

1. Whaley, K. J., and L. Zeitlin. 2022. Emerging antibody-based products for infectious diseases: Planning for metric ton manufacturing. *Hum. Vaccines Immunother.* 18, 1930847.
2. Kaplon, H., A. Chenoweth, …, J. M. Reichert. 2022. Antibodies to watch in 2022. *mAbs.* 14, 2014296.
3. Schlander, M., K. Hernandez-Villafuerte, …, M. Baumann. 2021. How Much Does It Cost to Research and Develop a New Drug? A Systematic Review and Assessment. *Pharmacoeconomics.* 39:1243–1269.
4. Modernizing Drug Discovery, Development & Approval March 31, 2016. https://phrma.org/-/media/Project/PhRMA/PhRMA-Org/PhRMA-Org/PDF/P-R/proactive-policy-drug-discovery.pdf.
5. Vermeer, A. W., and W. Norde. 2000. The thermal stability of immunoglobulin: unfolding and aggregation of a multi-domain protein. *Biophys. J.* 78:394–404.
6. Garber, E., and S. J. Demarest. 2007. A broad range of Fab stabilities within a host of therapeutic IgGs. *Biochem. Biophys. Res. Commun.* 355:751–757.
7. Kim, S. H., H. J. Yoo, …, D. H. Na. 2021. Nano Differential Scanning Fluorimetry-Based Thermal Stability Screening and Optimal Buffer Selection for Immunoglobulin G. *Pharmaceuticals.* 15:29.
8. Jain, T., T. Sun, …, K. D. Wittrup. 2017. Biophysical properties of the clinical-stage antibody landscape. *Proc. Natl. Acad. Sci. USA.* 114:944–949.
9. Shehata, L., D. P. Maurer, …, L. M. Walker. 2019. Affinity Maturation Enhances Antibody Specificity but Compromises Conformational Stability. *Cell Rep.* 28:3300–3308.e4.
10. Bailly, M., C. Mieczkowski, …, L. Fayadat-Dilman. 2020. Predicting Antibody Developability Profiles Through Early Stage Discovery Screening. *mAbs.* 12, 1743053.
11. Harmalkar, A., R. Rao, …, K. Y. Wei. 2023. Toward generalizable prediction of antibody thermostability using machine learning on sequence and structure features. *mAbs.* 15, 2163584.
12. Widatalla, T., Z. A. Rollins, …, A. Cheng. 2023. AbPROP: Language and Graph Deep Learning for Antibody Property Prediction. *ICML Workshop Comput. Biol.* https://icml-compbio.github.io/2023/papers/WCBICML2023_paper53.pdf.
13. Jia, L., M. Jain, and Y. Sun. 2022. Improving antibody thermostability based on statistical analysis of sequence and structural consensus data. *Antib. Ther.* 5:202–210.
14. Warszawski, S., A. Borenstein Katz, …, S. J. Fleishman. 2019. Optimizing antibody affinity and stability by the automated design of the variable light-heavy chain interfaces. *PLoS Comput. Biol.* 15, e1007207.
15. Bekker, G.-J., B. Ma, and N. Kamiya. 2019. Thermal stability of single-domain antibodies estimated by molecular dynamics simulations. *Protein Sci.* 28:429–438.
16. Waight, A. B., D. Prihoda, …, L. Fayadat-Dilman. 2023. A machine learning strategy for the identification of key in silico descriptors and prediction models for IgG monoclonal antibody developability properties. *mAbs.* 15, 2248671.
17. Zeiske, T., K. A. Stafford, and A. G. Palmer, 3rd. 2016. Thermostability of Enzymes from Molecular Dynamics Simulations. *J. Chem. Theor. Comput.* 12:2489–2492.
18. Molecular Operating Environment (MOE) 2022. Chemical Computing Group ULC, 910-1010 Sherbrooke St. W., Montreal, QC H3A 2R7. Chemical Computing Group Inc.
19. Hornak, V., R. Abel, …, C. Simmerling. 2006. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins.* 65:712–725.
20. Olsson, M. H. M., C. R. Søndergaard, …, J. H. Jensen. 2011. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. *J. Chem. Theor. Comput.* 7:525–537.
21. Søndergaard, C. R., M. H. M. Olsson, …, J. H. Jensen. 2011. Improved Treatment of Ligands and Coupling Effects in Empirical Calculation and Rationalization of pKa Values. *J. Chem. Theor. Comput.* 7:2284–2295.
22. Jorgensen, W. L., J. Chandrasekhar, …, M. L. Klein. 1983. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* 79:926–935.
23. Van Der Spoel, D., E. Lindahl, …, H. J. C. Berendsen. 2005. GROMACS: fast, flexible, and free. *J. Comput. Chem.* 26:1701–1718.

24. MacKerell, A. D., D. Bashford, …, M. Karplus. 1998. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B.* 102:3586–3616.

25. Wong, S. E., B. D. Sellers, and M. P. Jacobson. 2011. Effects of somatic mutations on CDR loop flexibility during affinity maturation. *Proteins.* 79:821–829.

26. Jeliazkov, J. R., A. Sljoka, …, J. J. Gray. 2018. Repertoire Analysis of Antibody CDR-H3 Loops Suggests Affinity Maturation Does Not Typically Result in Rigidification. *Front. Immunol.* 9, 413.

27. Kroon, G. J. A., H. Mo, …, P. E. Wright. 2003. Changes in structure and dynamics of the Fv fragment of a catalytic antibody upon binding of inhibitor. *Protein Sci.* 12:1386–1394.

28. Schoenle, M. V., Y. Li, …, R. Page. 2021. NMR Based SARS-CoV-2 Antibody Screening. *J. Am. Chem. Soc.* 143:7930–7934.

29. Lindorff-Larsen, K., N. Trbovic, …, D. E. Shaw. 2012. Structure and Dynamics of an Unfolded Protein Examined by Molecular Dynamics Simulation. *J. Am. Chem. Soc.* 134:3787–3791.

30. Berendsen, H. J. C., J. P. M. Postma, …, J. R. Haak. 1984. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* 81:3684–3690.

31. Evans, D. J., and B. L. Holian. 1985. The Nose–Hoover thermostat. *J. Chem. Phys.* 83:4069–4074.

32. Parrinello, M., and A. Rahman. 1981. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* 52:7182–7190.

33. Di Pierro, M., R. Elber, and B. Leimkuhler. 2015. A Stochastic Algorithm for the Isobaric-Isothermal Ensemble with Ewald Summations for all Long Range Forces. *J. Chem. Theor. Comput.* 11:5624–5637.

34. Ewald, P. P. 1921. Die Berechnung optischer und elektrostatischer Gitterpotentiale. *Ann. Phys.* 369:253–287.

35. Hess, B., H. Bekker, …, J. G. E. M. Fraaije. 1997. LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.* 18:1463–1472.

36. Ryckaert, J.-P., G. Ciccotti, and H. J. Berendsen. 1977. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* 23:327–341.

37. Miyamoto, S., and P. A. Kollman. 1992. Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J. Comput. Chem.* 13:952–962.

38. Chodera, J. D. 2016. A Simple Method for Automated Equilibration Detection in Molecular Simulations. *J. Chem. Theor. Comput.* 12:1799–1805.

39. Lefranc, M.-P., V. Giudicelli, …, G. Lefranc. 2005. IMGT, the international ImMunoGeneTics information system. *Nucleic Acids Res.* 33:D593–D597.

40. Dunbar, J., and C. M. Deane. 2016. ANARCI: antigen receptor numbering and receptor classification. *Bioinformatics.* 32:298–300.

41. Eisenhaber, F., P. Lijnzaad, …, M. Scharf. 1995. The double cubic lattice method: Efficient approaches to numerical integration of surface area and volume and to dot surface contouring of molecular assemblies. *J. Comput. Chem.* 16:273–284.

42. Shrake, A., and J. A. Rupley. 1973. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *J. Mol. Biol.* 79:351–371.

43. Stafford, K. A., N. Trbovic, …, A. G. Palmer. 2015. Conformational preferences underlying reduced activity of a thermophilic ribonuclease H. *J. Mol. Biol.* 427:853–866.

44. Chen, Y., S. L. Campbell, and N. V. Dokholyan. 2007. Deciphering Protein Dynamics from NMR Data Using Explicit Structure Sampling and Selection. *Biophys. J.* 93:2300–2306.

45. Korendovych, I. V. 2018. Rational and Semirational Protein Design. *Methods Mol. Biol.* 1685:15–23.

46. Yabuki, S. 2017. How to Lengthen the Long-Term Stability of Enzyme Membranes: Trends and Strategies. *Catalysts.* 7:36.

47. Bae, S.-H., H. J. Dyson, and P. E. Wright. 2009. Prediction of the rotational tumbling time for proteins with disordered segments. *J. Am. Chem. Soc.* 131:6814–6821.

48. Sutthibutpong, T., T. Rattanarojpong, and P. Khunrae. 2018. Effects of helix and fingertip mutations on the thermostability of xyn11A investigated by molecular dynamics simulations and enzyme activity assays. *J. Biomol. Struct. Dyn.* 36:3978–3992.

49. Li, Q., Y. Zheng, …, J. Tian. 2022. Computational design of a cutinase for plastic biodegradation by mining molecular dynamics simulations trajectories. *Comput. Struct. Biotechnol. J.* 20:459–470.

50. Sharp, K. A., E. O'Brien, …, A. J. Wand. 2015. On the relationship between NMR-derived amide order parameters and protein backbone entropy changes. *Proteins.* 83:922–930.

51. Hsu, A. 2020. The Critical Assessment of Protein Dynamics Using Molecular Dynamics (MD) Simulations and Nuclear Magnetic Resonance (NMR) Spectroscopy Experimentation.

52. Vugmeyster, L., O. Trott, …, A. G. Palmer. 2002. Temperature-dependent Dynamics of the Villin Headpiece Helical Subdomain, An Unusually Small Thermostable Protein. *J. Mol. Biol.* 320:841–854.

53. Johnson, E., A. G. Palmer, and M. Rance. 2007. Temperature dependence of the NMR generalized order parameter. *Proteins.* 66:796–803.

54. Massi, F., and A. G. Palmer. 2003. Temperature dependence of NMR order parameters and protein dynamics. *J. Am. Chem. Soc.* 125:11158–11159.

55. Harris, C. R., K. J. Millman, …, T. E. Oliphant. 2020. Array programming with NumPy. *Nature.* 585:357–362.

56. McKinney, W. 2010. Data Structures for Statistical Computing in Python. Austin, Texas, pp. 56–61.

57. Hunter, J. D. 2007. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* 9:90–95.

58. Virtanen, P., R. Gommers; …, SciPy 10 Contributors. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods.* 17:261–272.

59. Cock, P. J. A., T. Antao, …, M. J. L. De Hoon. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics.* 25:1422–1423.

60. Gowers, R., M. Linke, …, O. Beckstein. 2016. MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations. Austin, Texas, pp. 98–105.

61. McGibbon, R. T., K. A. Beauchamp, …, V. S. Pande. 2015. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* 109:1528–1532.

62. Miller, R. G. 1974. The Jackknife–A Review. *Biometrika.* 61:1–15.

63. Daoud, J. I. 2017. Multicollinearity and Regression Analysis. *J. Phys, Conf. Ser.* 949, 012009.

64. Strobl, C., A.-L. Boulesteix, …, T. Hothorn. 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinf.* 8:25.

65. Boulesteix, A.-L., S. Janitza, …, I. R. König. 2012. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *WIREs Data Min. &. Knowl.* 2:493–507.

66. Gregorutti, B., B. Michel, and P. Saint-Pierre. 2017. Correlation and variable importance in random forests. *Stat. Comput.* 27:659–678.

67. Chicco, D., M. J. Warrens, and G. Jurman. 2021. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ. Comput. Sci.* 7:e623.

68. Waldmann, P. 2019. On the Use of the Pearson Correlation Coefficient for Model Evaluation in Genome-Wide Prediction. *Front. Genet.* 10, 899.

69. Benevenuta, S., and P. Fariselli. 2019. On the Upper Bounds of the Real-Valued Predictions. *Bioinf. Biol. Insights.* 13, 1177932219 871263.

70. Rollins, Z. A., J. Huang, …, S. C. George. 2022. A computational algorithm to assess the physiochemical determinants of T cell receptor dissociation kinetics. *Comput. Struct. Biotechnol. J.* 20:3473–3481.

71. Pedregosa, F., G. Varoquaux, …, É. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* 12:2825–2830.

72. Scikit-Optimize Sequential Model-Based Optimization in Python — Scikit-Optimize 0.8.1 Documentation. https://scikit-optimize.github.io/stable/user_guide.html.

73. Chen, T., and C. Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. Preprint at arXiv. https://doi.org/10.48550/arXiv.1603.02754.

74. Raschka, S. 2018. MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack. *J. Open Source Softw.* 3:638.

75. Dunbar, J., A. Fuchs, …, C. M. Deane. 2013. ABangle: characterising the VH–VL orientation in antibodies. *Protein Eng. Des. Sel.* 26:611–620.

76. Olsen, T. H., I. H. Moal, and C. M. Deane. 2022. AbLang: an antibody language model for completing antibody sequences. *Bioinform. Adv.* 2, vbac046.

77. Devlin, J., M.-W. Chang, …, K. Toutanova. 2019. 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). Association for Computational Linguistics, Minneapolis, Minnesota, pp. 4171–4186.

78. Olsen, T. H., F. Boyles, and C. M. Deane. 2022. Observed Antibody Space: A diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. *Protein Sci.* 31:141–146.

79. Salgado, J. C., I. Rapaport, and J. A. Asenjo. 2006. Predicting the behaviour of proteins in hydrophobic interaction chromatography. 2. Using a statistical description of their surface amino acid distribution. *J. Chromatogr. A.* 1107:120–129.

80. Velegol, D., J. D. Feick, and L. R. Collins. 2000. Electrophoresis of Spherical Particles with a Random Distribution of Zeta Potential or Surface Charge. *J. Colloid Interface Sci.* 230:114–121.

81. Sillero, A., and J. M. Ribeiro. 1989. Isoelectric points of proteins: theoretical determination. *Anal. Biochem.* 179:319–325.

82. Buitinck, L., G. Louppe, …, G. Varoquaux. 2013. API design for machine learning software: experiences from the scikit-learn project. Preprint at arXiv. https://doi.org/10.48550/arXiv.1309.0238.

83. Yang, J. S., S. Wallin, and E. I. Shakhnovich. 2008. Universality and diversity of folding mechanics for three-helix bundle proteins. *Proc. Natl. Acad. Sci. USA.* 105:895–900.

84. Piana, S., K. Lindorff-Larsen, and D. E. Shaw. 2013. Atomic-level description of ubiquitin folding. *Proc. Natl. Acad. Sci. USA.* 110:5915–5920.

85. Ereño-Orbea, J., T. Sicard, …, J.-P. Julien. 2017. Molecular basis of human CD22 function and therapeutic targeting. *Nat. Commun.* 8:764.

86. Abanades, B., W. K. Wong, …, C. M. Deane. 2023. ImmuneBuilder: Deep-Learning models for predicting the structures of immune proteins. *Commun. Biol.* 6:575–578.

87. Fernández-Quintero, M. L., N. D. Pomarici, …, K. R. Liedl. 2020. Antibodies exhibit multiple paratope states influencing VH–VL domain orientations. *Commun. Biol.* 3:1–14.

88. Fernández-Quintero, M. L., B. A. Math, …, K. R. Liedl. 2019. Transitions of CDR-L3 Loop Canonical Cluster Conformations on the Micro-to-Millisecond Timescale. *Front. Immunol.* 10

89. Yang, K. K., Z. Wu, and F. H. Arnold. 2019. Machine-learning-guided directed evolution for protein engineering. *Nat. Methods.* 16:687–694.

90. Hie, B. L., and K. K. Yang. 2022. Adaptive machine learning for protein engineering. *Curr. Opin. Struct. Biol.* 72:145–152.

91. Yuan, H., C. Ni, …, M. Wang. 2022. Bandit theory and thompson sampling-guided directed evolution for sequence optimization. *Preprint at arXiv.* https://doi.org/10.48550/arXiv.2206.02092.