MABS
2023, VOL. 15, NO. 1, 2163584
https://doi.org/10.1080/19420862.2022.2163584

Taylor & Francis
Taylor & Francis Group

REPORT

OPEN ACCESS Check for updates

Toward generalizable prediction of antibody thermostability using machine learning on sequence and structure features

Ameya Harmalkar[a*,1] Roshan Rao[b*,1] Yuxuan Richard Xie[c*], Jonas Honer[d], Wibke Deisting[d], Jonas Anlahr[d], Anja Hoenig[d], Julia Czwikla[d], Eva Sienz-Widmann[d], Doris Rau[d], Austin J. Rice[e], Timothy P. Riley[e], Danqing Li[e], Hannah B. Catterall[e], Christine E. Tinberg[f], Jeffrey J. Gray[a], and Kathy Y. Wei[f]

[a]Department of Chemical and Biomolecular Engineering, The Johns Hopkins University, Baltimore, MD, USA; [b]Electrical Engineering and Computer Science, University of California, Berkeley, CA, USA; [c]Department of Bioengineering and Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, IL, USA; [d]Therapeutic Discovery, Amgen Research (Munich) GmbH, Munich, Germany; [e]Therapeutic Discovery, Amgen Research, Amgen Inc, Thousand Oaks, CA, USA; [f]Therapeutic Discovery, Amgen Research, Amgen Inc, South San Francisco, CA, USA

ABSTRACT
Over the last three decades, the appeal for monoclonal antibodies (mAbs) as therapeutics has been steadily increasing as evident with FDA's recent landmark approval of the 100th mAb. Unlike mAbs that bind to single targets, multispecific biologics (msAbs) have garnered particular interest owing to the advantage of engaging distinct targets. One important modular component of msAbs is the single-chain variable fragment (scFv). Despite the exquisite specificity and affinity of these scFv modules, their relatively poor thermostability often hampers their development as a potential therapeutic drug. In recent years, engineering antibody sequences to enhance their stability by mutations has gained considerable momentum. As experimental methods for antibody engineering are time-intensive, laborious and expensive, computational methods serve as a fast and inexpensive alternative to conventional routes. In this work, we show two machine learning approaches – one with pre-trained language models (PTLM) capturing functional effects of sequence variation, and second, a supervised convolutional neural network (CNN) trained with Rosetta energetic features – to better classify thermostable scFv variants from sequence. Both of these models are trained over temperature-specific data (TS50 measurements) derived from multiple libraries of scFv sequences. On out-of-distribution (refers to the fact that the out-of-distribution sequnes are blind to the algorithm) sequences, we show that a sufficiently simple CNN model performs better than general pre-trained language models trained on diverse protein sequences (average Spearman correlation coefficient, ρ, of 0.4 as opposed to 0.15). On the other hand, an antibody-specific language model performs comparatively better than the CNN model on the same task ($\rho =$ 0.52). Further, we demonstrate that for an independent mAb with available thermal melting temperatures for 20 experimentally characterized thermostable mutations, these models trained on TS50 data could identify 18 residue positions and 5 identical amino-acid mutations showing remarkable generalizability. Our results suggest that such models can be broadly applicable for improving the biological characteristics of antibodies. Further, transferring such models for alternative physicochemical properties of scFvs can have potential applications in optimizing large-scale production and delivery of mAbs or bsAbs.

ARTICLE HISTORY
Received 16 June 2022
Revised 14 December 2022
Accepted 26 December 2022

KEYWORDS
Antibody design; thermostability prediction; machine learning; unsupervised and supervised learning

1. Introduction

Monoclonal antibodies (mAbs) represent a large class of therapeutic agents, with more than 100 FDA-approved products marketed in the US (www.antibodysociety.org). Despite their widespread prevalence in drug development, mAbs are limited in biological scope because they bind only a single target. Multispecific biologics (bsAbs) engaging more than one target or epitope on the same target are of growing importance for accessing novel, therapeutically relevant pathways and mechanisms of action. In recent years, several multispecific biologics are approved for use and many more are in clinical and preclinical development.[1,2]

A common building block for the construction of multispecific biologics is the single-chain variable fragment (scFv), consisting of the target-engaging variable heavy chain (VH) linked to the variable light chain (VL) via a flexible linker. Multispecific format platforms such as the BiTE,[3] IgG-scFv,[4] and XmAb[5] incorporate scFv modules. Although scFvs are prevalent in multispecific biologic candidates, they may display sub-optimal physical properties relative to conventional mAbs and generally require sequence modifications to produce a developable asset. One property that is used to gauge the potential developability of a scFv module or scFv-containing multispecific is thermostability – scFv candidates are experimentally screened and/or optimized for thermostability to identify suitable modules.[6,7] However, these

CONTACT Kathy Y. Wei kwei@amgen.com Therapeutic Discovery, Amgen Research, Amgen Inc, South San Francisco, CA 94080, USA

*Work performed during internship at Amgen Research, Amgen Inc

[1]A.H and R.R share equal author contribution

Supplemental data for this article can be accessed online at https://doi.org/10.1080/19420862.2022.2163584 experiments are resource intensive and time-consuming. Accurate computational methods to predict scFv thermostability from primary amino acid sequence for scFv candidate selection/deselection (and to predict mutations to guide thermostability engineering efforts) would be invaluable to multispecific drug development.

Over the past decade, the use of computational tools to predict stability-enhancing mutations has gained considerable momentum, albeit with limited success. Protein consensus design, a state-of-the-art approach, uses phylogenetic information from multiple sequence alignments (MSAs) to obtain the most frequent amino acid for a residue position.[8] However, these residues have improved thermostability in only 50% of the cases, with even poorer performance for antibody sequences with highly conserved framework regions. Structure-based approaches, such as PROSS[9] and AbLIFT,[10] have used thermodynamic energies (Rosetta $\Delta\Delta$ G) and structural information (CDR-dependent PSSMs) to improve binding and stability, however their application to scFvs is often limited by the relative scarcity of structural data. Alternatively, machine learning approaches have employed large datasets such as ProTherm,[11,12] that collate mutant effects on protein stability from mesophilic and thermophilic sequences, to predict thermostability.[13–15] Unfortunately, the plethora of publicly available thermostability data excludes mAb or scFv sequences, limiting their generalization. Predicting thermally enhancing scFv sequences or designing thermostable mutations within approved scFv candidates is still a challenge. Currently, there is a need for computational approaches to leverage sequence information to predict biophysical attributes, such as thermostability, with high accuracy and generalizability enabling efficient protein design. We address this critical need by equipping unsupervised and supervised learning approaches over a thermostability prediction task tuned specifically for generated scFv sequences.

We demonstrate the use of deep learning approaches to infer thermostability attributes from scFv sequences generated experimentally and screened for their temperature sensitivity. First, with pre-trained language models (PTLMs), we assess the ability of unsupervised networks to predict thermostability with zero-shot and fine-tuned predictions. Then, we utilize supervised learning to train simple, predictive CNN architectures. To provide structural context to the supervised networks, we also feed the network with thermodynamic information via Rosetta energies. Further, we examine whether these networks could provide insights toward design of thermostable mutants, thereby improving biologics engineering. With this work, we present a proof-of-concept study of utilizing PTLMs and thermodynamic features toward relatively niche problems in protein informatics.

Results

Machine learning tasks can be tuned to identify scFv thermostability

The biological problem of thermostability prediction at the sequence-level involves identifying which sequences could result in a highly thermostable biomolecular structure; for antibodies and scFvs, this implies conservation of the folded structural state and/or antigen binding upon high-temperature stress (Figure 1a). Thermal stability of proteins depends on residue-level biophysical attributes. However, for antibodies and scFvs, owing to the high consensus in sequences, deciphering heuristic or empirical rules based solely on sequence patterns for distinguishing thermostable and unstable sequences is challenging. Machine learning (ML) models have shown potential to extract higher-order relationships mapping sequences to function in the absence of underlying biophysical pathways, and they perform well on classification tasks. Leveraging sequence and structural information as features, ML approaches applied on a plethora of prediction tasks, e.g., fluorescence landscapes,[16–18] intrinsic stability,[19,20] missense variant effects,[21] protein fitness,[22] antigen-specificity,[23] have shown high success rates. With the availability of an explicit dataset with temperature-level information, we could extrapolate ML methods for thermostability prediction tasks.

To learn temperature-specific contextual patterns in sequence-data, we sought to develop and train machine-learning models for thermostability prediction using scFv sequences. We collected temperature data (TS50, temperature at half-max binding) from various engineering studies for developing thermostable single-chain variable fragment (scFv) molecules. The experimental data used in this study were curated from historical therapeutic programs, and data collection was not devised with the explicit intent of training a predictive thermostability model, resulting in non-uniform distributions in our data. The sequence data contained scFv sequences assembled by performing mutations to heavy and light chains from multiple germlines (Methods and Sup. Figure. S1). We collated 2,700 scFv sequences from 17 projects that target different antigens (further referred to as experimental sets) to constitute the sequence data. Additionally, sequences from another scFv study (currently under clinical trials) and an isolated scFv dataset form out-of-distribution, blind test sets. For each sequence, thermostability is evaluated with a TS50 measurement representing the temperature at half-maxima of target binding, and this measurement serves as the temperature annotation (Figure 1(b)). For the isolated scFv dataset, thermostability is evaluated with a Tm measurement representing the first transition from folding to unfolding as temperature is increased.

The temperature measurement (TS50) for 2,700 scFv sequences is used for two thermostability prediction tasks: (1) Regression: Prediction of TS50 measurement of a scFv sequence and (2) Classification: Prediction of whether a given sequence corresponds to a thermally stable scFv. For the regression task, absolute values of TS50 measurements are used, whereas, for the classification task, the TS50 data are divided into four bins, namely under-50°C, 50°C–60°C, 60°C–70°C and 70°C-up. With these training data, we have trained two models (Figure 1c): (1) Pre-trained language models (PTLMs), unsupervised BERT-like model architectures,[24,25] trained over large sequence corpus spanning evolutionary diversity. These models are trained to extrapolate learned representations of protein structures, function and biological

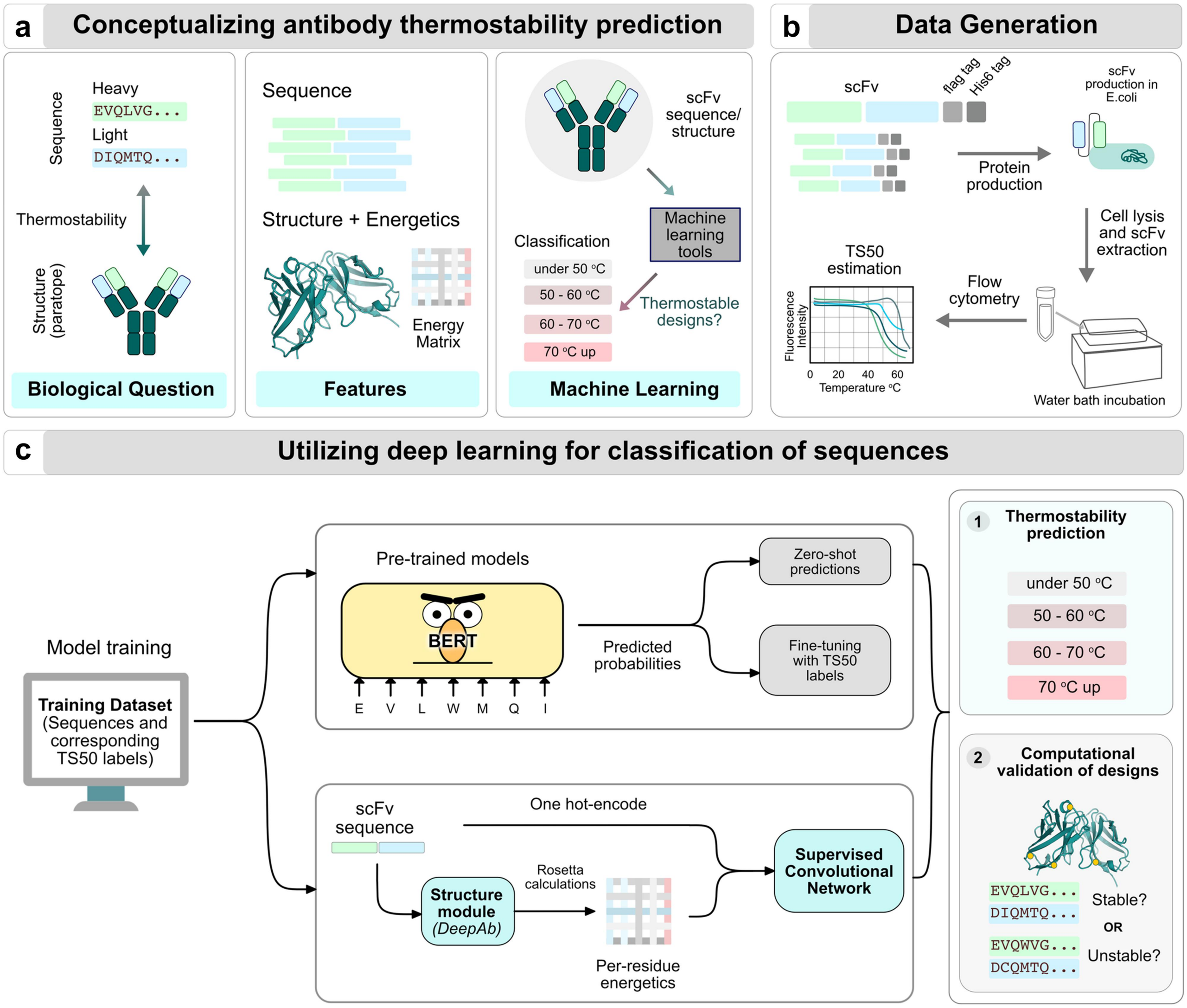


Figure 1. A pipeline to identify scFv thermostability using deep learning. (a) *The biological challenge of antibody thermostability prediction from sequences*. Antibody thermostability can determine the manufacturability of antibodies in downstream processes. The biological question that we AIM to tackle is whether we can predict the thermal characteristics of an scFv, given its sequence. Available data for this challenge can comprise of the amino acid sequences, structures and calculated energetics. Leveraging antibodies with pre-determined temperature characteristics is paramount, however, the availability is scarce for such a dataset. (b) *Thermostability data generation*. To generate a dataset of scFv sequences with known temperature-specific features, we determined the loss of target binding of the scFv post high temperature stress to obtain a TS50 measurement. (c) *Training a classification network for predicting TS50 bins*. One of the approaches is transfer learning with unsupervised models (top branch). We utilized pre-trained BERT-like models (such as ESM1-b, ESM1-v, etc) to make (1) Zero-shot predictions and (2) Fine-tuned predictions with the labeled TS50 dataset. Another approach is to train a supervised model with calculated thermodynamic energies (bottom branch). We used sequence and structure-based features for supervised learning using simple convolutional models to train a classifier. The outcome of such trained ML models can be employed either for predicting thermostability of generated antibody sequences or to computationally validate experimental designs.

activity at a sequence-level, and they can make zero-shot predictions or be fine-tuned with thermal stability data. (2) Supervised convolutional models that equip simple deep-convolutional networks utilizing annotated thermostability data with feature encodings at a sequence-level (i.e. one-hot encoded amino-acid types) and an energy-level (i.e. thermodynamic energies obtained from putative three-dimensional structural models generated with DeepAb.[26])

With both models, our aim is to predict thermostability of a given scFv sequence. By developing a model that can effectively filter and screen scFv sequences, we could significantly accelerate the identification of better variants for stable, manufacturable, multispecific biologics.

Fine-tuning pre-trained sequence models with thermostability data improves classification performance over zero-shot predictions

Unsupervised models trained on a large corpus of protein sequences are reported to infer evolutionary relationships and statistical patterns about protein structure and function.[25,27] PTLMs have reportedly shown successful performance in downstream prediction tasks (e.g., predicting mutational landscapes, secondary structure and tertiary contacts[24]) without any additional supervision i.e. in a zero-shot setting where inference is performed directly on the input sequence. To assess whether zero-shot learning from the PTLMs could be extrapolated for the thermostability tasks, we evaluate likelihood-based zero-shot predictions[25] from the ESM-1 v[25] and UniRep[20] language models trained on diverse protein sequences and the AntiBERTy language model[28] trained on antibody sequences from the Observed Antibody Space (OAS) (Figure 2a).[29] Figure 2c,e show that zero-shot predictions do not in general correlate well with thermostability, either on TS50 data or on blind test sets. These results are contrary to those reported in prior work.[25,27,30] It is important to note that the setting is quite different; prior work largely evaluates single mutations of a parent protein, whereas our datasets consist of

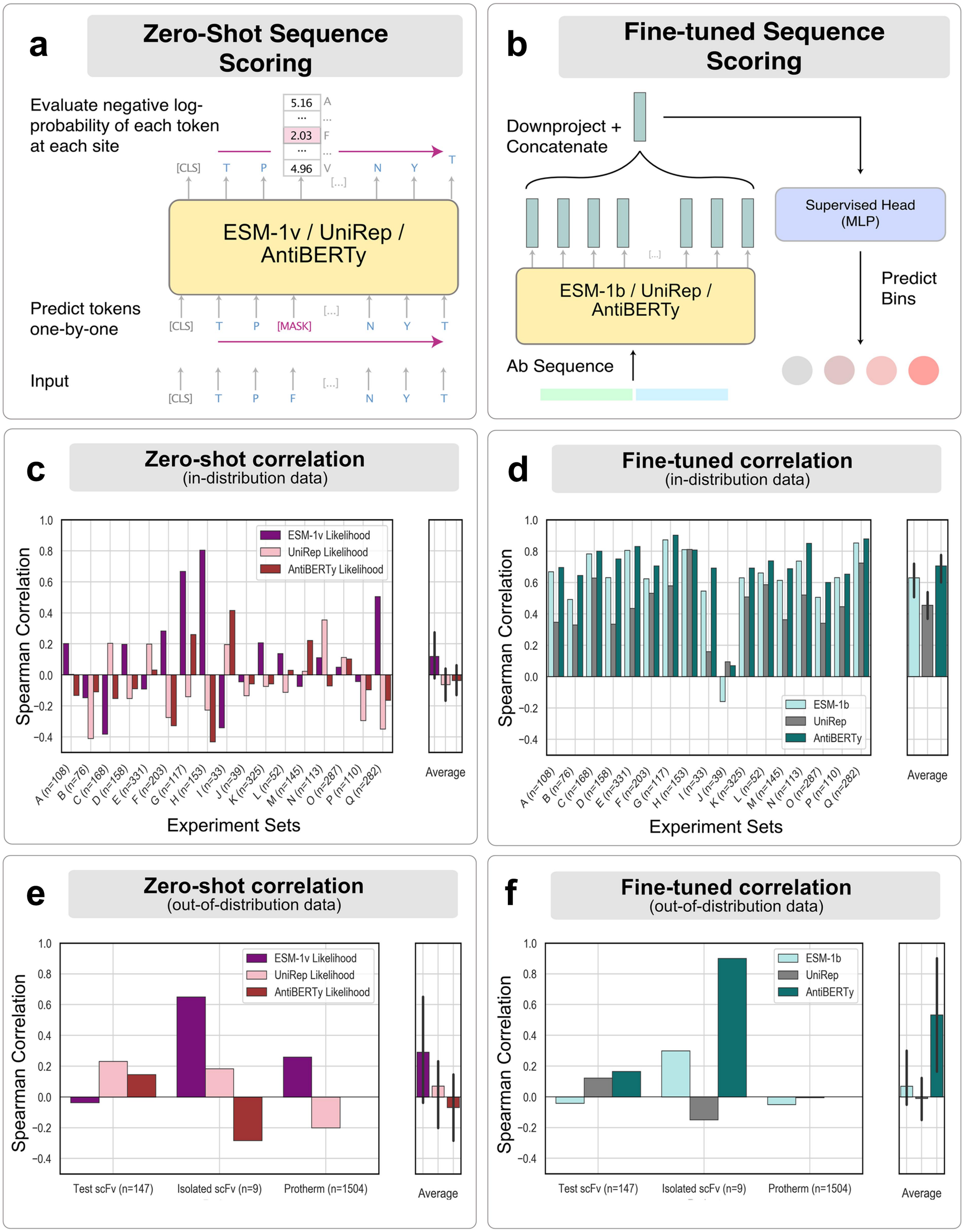


Figure 2. **Fine-tuning over pre-trained unsupervised models improves correlation on withheld targets**. (a) Zero-shot and (b) Fine-tuned sequence scoring methods for thermostability prediction. (c) Zero-shot likelihood-based predictions with pre-trained models do not correlate strongly with the TS50 datasets. (d) Fine-tuning the pre-trained models on TS50 data from $n - 1$ targets significantly improves correlation on the held-out target. (e) Zero-shot likelihood-based predictions on blind test sets. (f) Models fine-tuned on TS50 data do not generalize well to blind test sets.

multiple mutations (including insertions and deletions) and are derived from multiple parent proteins. For the antibody-specific language model, the correlation is worse than the other PTLMs trained on diverse protein sequences. This is in agreement with the recent work by Nijkamp *et al.*, demonstrating that large-scale, antibody-specific language models are relatively poor predictors for general properties such as thermostability.[31]

Next, we fine-tuned the pre-trained features from the ESM-1b, UniRep and AntiBERTy language models (Figure 2b) specifically on our downstream thermostability prediction task to assess if that improves performance. For this task, we froze the pre-trained weights for the models and trained the parameters of the classification head – a multilayer perceptron (Methods) that predicts the TS50 temperature bins. Fine-tuning the PTLMs on $(n-1)$ sets significantly improve correlation for the held-out target as evident in Figure 2d. Fine-tuned predictions from ESM-1b, UniRep and AntiBERTy models achieve moderate-to-high average Spearman correlation on held-out targets when trained on TS50 data (0.63, 0.45 and 0.71 respectively) (Figure 2d). However, the ESM-1b and UniRep predictions do not generalize well to blind test sets (Figure 2f). On the other hand, the AntiBERTy-finetuned model generalizes better on the out-of-distribution dataset. Unlike ESM-1 v and Unirep language models, AntiBERTy encodes antibody-specific features, which can be then fine-tuned for prediction on antibody-specific landscapes, as demonstrated here for scFv thermostability. By clustering the embeddings from the AntiBERTy-finetuned model via t-distributed stochastic neighbor embedding (t-SNE), we found that even after fine-tuning, the sequences stay clustered by their experimental sets (Sup. Figure. S2). This suggests that the model exploits some underlying relationship in the sequences to make predictions. Even though the performance generalizes to new datasets, it might serve as a potential caveat when utilizing the fine-tuned model for designing/validating new thermostable sequences.

Supervised network trained with energy features improves generalizability across out-of-distribution datasets

Unlike pre-trained models utilizing large sequential information, Shanehsazzadeh *et al.* demonstrated that small supervised models could achieve competitive performance on downstream prediction tasks benchmarked in TAPE.[18] Similar predictive performance was also reported for antigen-specificity prediction with supervised convolutional networks.[23] To evaluate whether small, supervised networks would perform and generalize better, we built a supervised convolutional model with the scFv sequences for the thermostability prediction task. Figure 3a shows the detailed architecture of our supervised CNN deep-learning model. Since the sample size of the experimental dataset was relatively small (2,700 scFv sequences), we decided to supplement the network with structure-specific information. Conventional structure-based approaches (such as evaluating based on Rosetta Δ G) showed poor correlation with TS50 (Sup. Fig. S3), primarily owing to a variable sequence length and structural contributions affecting global energies. To feed the network in a localized structural context, we incorporated the energetics as a two-dimensional *i-j* residue energy matrix. The residue energy matrix serves as a contact map that provides a reduced representation of protein-free energies rather than a global-free energy estimation. For each scFv sequence, we generated a structural model with DeepAb[26] and evaluated the thermodynamic features (total energy split into one-body, *i-i*, and two-body, *i-j*, residue energies) calculated using Rosetta ref2015 energy function.[32] The contributions of i^{th} residue with every j^{th} residue (where $j \in 1, N$ such that $N =$ total number of residues) were tabulated and binned in an *i-j* matrix that constituted the energy features. Finally, we equipped the model with two branches: (1) sequence branch with one-hot encoded amino-acid sequences and (2) energetics branch with pairwise *i-j* residue-residue energy matrix. The final model architecture and hyperparameters are reported in Figure 3a, and this model was trained (Sets A-O) and evaluated with the available sequences (held-out sets P and Q and two scFv sets from different studies). The model architecture was built such that contributions from either of the two branches could be turned off to obtain sequence or energy dependence over the classification performance.

In spite of the sequence diversity in the experimental data, we wanted to investigate whether there was an underlying relationship between the sequences; whether the experimental sets from which the sequences were derived had an impact over prediction accuracies. We analyzed the representation learned by the sequence-only model and the energetics-only model by projecting the embeddings from the dense layer for each sequence into two dimensions via t-distributed stochastic neighbor embedding (t-SNE) Figure 3b. The sequence-only model embeddings were clustered by their experimental set, as evident by the aggregation of colored points in Figure 3b.iv. On the other hand, the energetics-only model embeddings were independent of any clustering based on the experimental set as demonstrated by the noisy embedding for energetics (Figure 3b) ii. Thus, fine-tuned and supervised models trained only on sequence-features are able to infer the underlying experimental origin of the sequences and skew thermostability predictions, making them less generalizable toward newer, blind datasets.

Since the energetics-only model is independent of any sequence-specific information, we assessed the performance of this supervised model by constructing a receiver-operating-characteristic (ROC) curve derived from the prediction of the 70-up bin (Figure 3c). As we aim to identify thermostable sequences, the prediction accuracy of the 70-up bin is most important. We evaluated the ROC for four test datasets: two held-out (Sets P and Q) and two blind datasets representing a test scFv and an isolated scFv. The area under ROC is over 0.7, denoting a high classification accuracy. Figure 3d shows the Spearman correlation coefficient for all four test datasets, with the energetic-only, sequence-only and energetics + sequences models, respectively. On held-out datasets (Set P and Q), among the supervised models, the coefficients are over 0.5 for energetics-only model, with energetics + sequences model

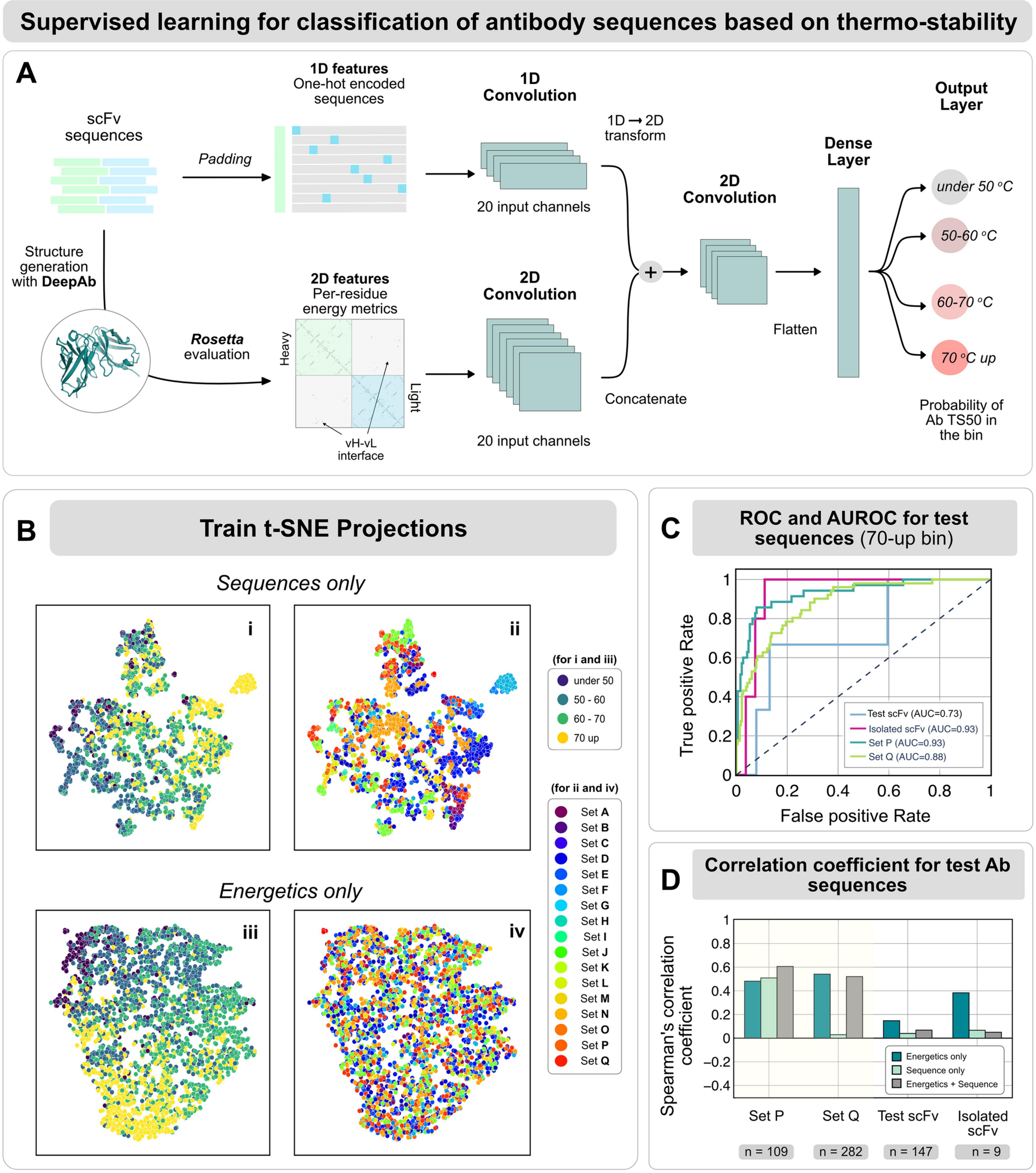


Figure 3. Energy features can extract 'generalizable' information of thermostability. (a) The supervised convolutional network architecture for classification of antibody sequences. The input scFv sequences pass a structure-generation module with DeepAb followed by Rosetta-based evaluation to estimate per-residue energies for each amino acid residue in the scFv structure. The sequences are one-hot encoded (top branch) and the energetic features, represented as an i-j matrix (bottom branch), are provided to the network. The output from the sequence branch and the energy branch are passed through a dense-layer to generate the probabilities of the sequence to lie in each of the temperature bins. (b) t-stochastic neighbor embeddings from the energetics-only model colored by the temperature bins. (c) Receiver-operating characteristic curve to demonstrate the classification of the test sequences for the above-70 bin with the energetics-only model. Note that Test scFv and Isolated scFv have a smaller sample size, explaining the relatively less rugged nature of the curves. (d) The model's performance metrics for the classification task on completely blind test scFv sequences is reported with the Spearman's correlation coefficient.

showing a slightly better performance. However, on blind datasets, the performance drops for energetics + sequences and sequences-only (coefficients under 0.1). The energetics-only model shows better correlation for the blind datasets (0.2 and 0.4 respectively) than the other supervised models. In comparison with the PTLM performance in Figure 2c-f, the AntiBERTy-finetuned model shows better correlation (average correlation of 0.52 versus 0.29 for SCNN trained on Sets A-Q and 0.4 for ensemble of SCNNs, see Sup. Fig. S6-S7). However, it is important to note that the language models are inherently skewed toward the experimental datasets.

Finally, as a control, we randomly initialized the weights in the SCNN for the classification task and found that it is unable to distinguish sequences based on thermostability (Sup. Fig. S1). Further, on the test sets, we performed weighted random predictions, i.e. we predicted the classification bin label with a weighted random choice, with sample size in each bin set as the weights (Sup. Fig. S4-S5). In both these tests, where the energetics-only SCNNs were able to decipher some relationship between the energetics of the scFv and the thermostability, the randomly initialized models could not demonstrate any discernible relationship, demonstrating the significance of learned representations from supervised data.

Networks trained with experimental TS50 data can distinguish thermal stability-enhancing designs

With improved thermal stability predictions on scFv sequences, we aspire to optimize and engineer antibody or scFv molecules for specific biomolecular applications. Thus, we sought to evaluate the ability of our predictive models in discriminating between thermostable and thermally degenerate mutations. Prior studies by Koenig *et al.*[33] and Warszawski *et al.*[10] detail thermal aggregation experiments on point mutations for an anti-VEGF antibody (PDB ID: 2 FJG/2FJF[34]). In both the studies, a deep mutational scanning (DMS) experiment was performed for the antibody and selected point mutations that improved binding enrichment over wildtype were analyzed for their fragment antigen-binding (Fab) melting temperature (T_m). These point mutants (20 mutations compiled from both studies) serve as a test case to evaluate whether the networks trained on TS50 temperature measurements could obtain insights about related temperature-dependent attributes such as thermal aggregation and whether they distinguish the thermally enhancing and thermally hampering mutations. Although the model is better suited to classify diverse sequences rather than point mutations, with this test, we sought to understand the generalizability of our models trained on scFv sequences and TS50 measurements to antibody sequences and melting temperature measurements.

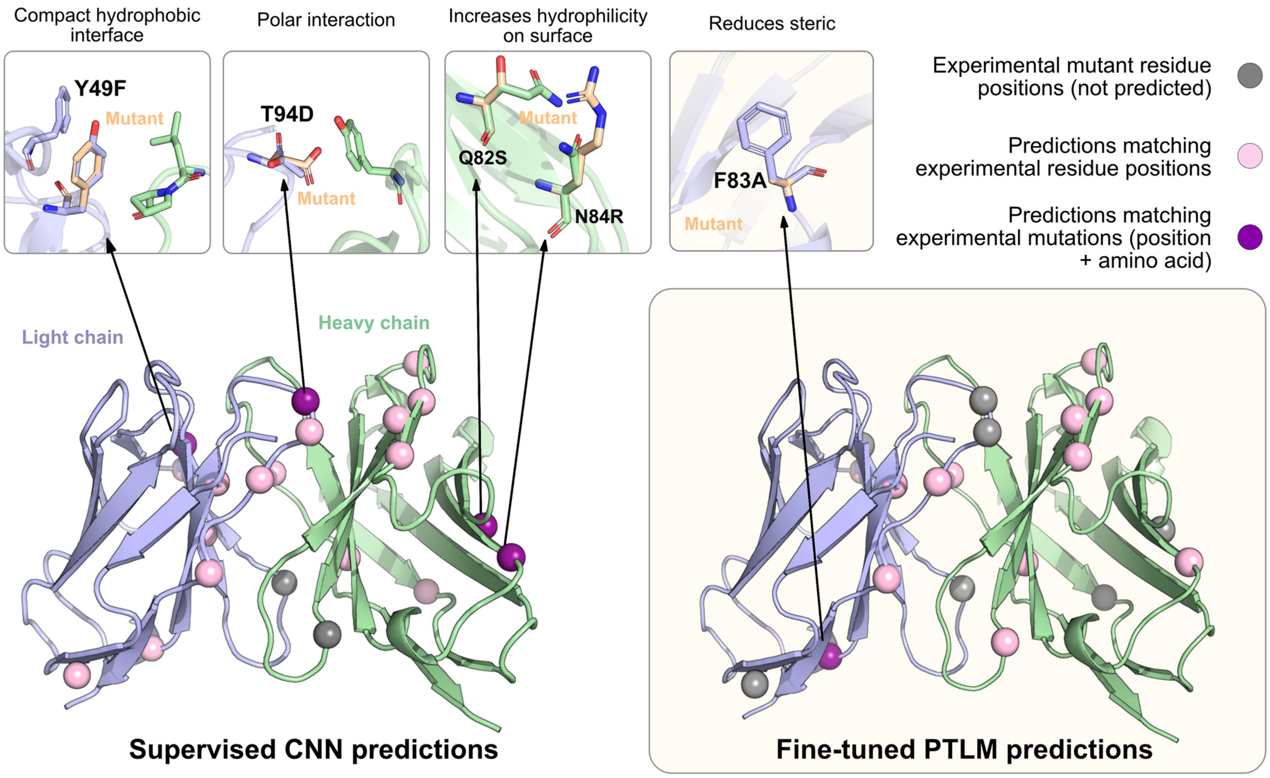
To perform an unbiased analysis, we performed point mutations over the antibody (a computational DMS) and analyzed the classification performance of our PTLM models (zero-shot and fine-tuned) and our SCNN model (ensemble of energetics-only CNNs) on these point mutations. Out of the language models, the UniRep and AntiBERTy fine-tuned models failed to differentiate between point mutations and attributed over 99% mutants to the 70-up bin. Only the ESM-1b model could differentiate between the point mutants well, i.e. had predictions for all temperature bins. Figure 4 compares our 70-up bin predictions with the experimental thermostable mutants for the heavy and light chains with the two models (SCNN and ESM-1b fine-tuned PTLM). The 20 mutation positions that were validated experimentally are highlighted as spheres in the cartoon representations. Our networks identify five out of the 20 mutations correctly (highlighted in magenta, 4 mutants identified by SCNN and one by ESM-1b finetuned model). Surprisingly, all five of these mutations comprise the framework residues. Further, for 18 out of 20 mutations, the SCNNs could identify the residue position correctly, albeit predicting different amino-acid mutations as most thermostable.

Out of 4540 point mutations analyzed (N_{res} = 227 residues, 20 amino acids per residue), experimental data were available for only 20 point mutations. Since only 0.44% of the total possible mutations in the anti-VEGF antibody were assessed for melting temperatures experimentally, the validation dataset for thermostability is sparse. Further, in spite of being temperature-specific attributes, TS50 and T_m are different experimental measurements and do not correlate exactly. Conventional structure-based approaches, such as $\Delta\Delta$ G from Rosetta[35] or FoldX,[36] could hardly predict thermostable point mutations (Sup. Fig. S8). It is, therefore, remarkable that our networks could predict the thermostable residue positions in 90% of the cases, with 25% successful predictions (correct residue positions as well as amino-acid residues).

By extrapolating the networks trained on TS50 measurements over alternative thermal aggregation experiments (T_m in this case), we demonstrate that intrinsic thermal attributes could be captured by such models. Moreover, on comparing the residue positions violating the germline consensus sequence for the anti-VEGF Ab, we observed different amino acid mutations, highlighting the ability of these models to provide mutations orthogonal to traditional germlining approaches (Sup. Fig. S9-S10). With a more diverse and larger training dataset, it would be possible to develop a more robust model. Our results suggest that these networks could serve as a useful tool for screening or filtering scFv (or even antibody) sequences for temperature-specific antibody design pipelines.

Discussion

Thermostability is an important determinant of developability. To address the limitations in developing thermostable biological candidates, antibody engineering efforts are directed toward identifying and screening for sequences that can improve thermostability. In this work, we have tested two approaches for prediction of thermostable scFv sequences from features learned with a sequential and thermodynamic context. As the corpus of sequence databases is vast (billions of sequences from diverse protein families), we equipped the unsupervised learned representations via pre-trained language models to classify sequences into temperature-specific bins quantifying their thermostability. Unlike conventional machine-learning approaches that use sequence or structural-coordinate features, we incorporated enriched information with thermodynamic features. Further, we tested the performance of using energetic features on small, supervised CNN models for the classification tasks. Finally, we demonstrated the applicability of our work for antibody engineering efforts by identifying experimentally validated melting temperature (Tm) enhancing mutations on an anti-VEGF antibody. While the primary objective of this work was to study proof-of-principle for scFv thermostability classification with machine-learning models, the secondary objective was to identify


Compact hydrophobic interface
Polar interaction
Increases hydrophilicity on surface
Reduces steric
Y49F
Mutant
T94D
Mutant
Mutant
Q82S
N84R
F83A
Mutant
Experimental mutant residue positions (not predicted) ●
Predictions matching experimental residue positions ●
Predictions matching experimental mutations (position + amino acid) ●
Light chain
Heavy chain
Supervised CNN predictions
Fine-tuned PTLM predictions

Heavy chain				Light chain			
Wildtype	Experimental	SCNN	PTLM	Wildtype	Experimental	SCNN	PTLM
S7*	L	Y	-	S12*	M	W	-
A40*	E	-	C, P, S	A43	P	-	-
A54	R	G	D, E	Y49	F	***F***, W	-
T58	E	D, M, Y	N	Y55	F	H, L, N	G, T
Y59	H	L, M	F, V, S	F83*	A	D, K, Q	***A***, T, Y
Y60*	T	H, I, R	I, S	Q89	L	R	G, N, P, S
Q82*	S	H, N, ***S***	-	T94	D, V	***D***, H, M	-
N84*	R	H, K, ***R***, Y, W	S, G	P95	T	L, F	-
Y95	F	D, H, M	A, G	T97	M	Y	K, R
M107	Y	D	F, H, K	Q100	P	H	A

Figure 4. **Computational deep mutational scan of an antibody variable fragment shows agreement with experimental thermal denaturation data**. Validating all the point mutants with our SCNN and ESM-1b finetuned models for anti-VEGF antibody (PDB ID: 2FJG(bound) and PDB ID: 2FJF (unbound)), we observed synergies in mutants predicted in the over-70°C bin and the experimental thermal denaturation data available from prior work.[10,33] Spheres indicate the experimentally validated mutants that improved Tm; pink indicates predictions from the network with the same residue position, but different amino acid mutation; red indicates the predictions matching experimental data and gray indicates mutations which were not observed in computational predictions. Thumbnails highlight the mutations in agreement with experiments and potential interactions. The table illustrates the comparison with the experimental and computational predictions.

'generalizable' feature representations that can aid in creating a pipeline for rapid, computational screening and validation of scFv and antibody sequences based on their thermal characteristics.

First, we extrapolated the zero-shot learning and fine-tuning principles of large-scale PTLMs for inferring temperature-dependent biophysical attributes of scFv sequences. We acknowledge the limitations of extrapolating language models trained on massive sets of protein sequences to scFv-specific sequence data. Unlike natural proteins that evolve across many species under selective or evolutionary pressures, antibodies are often selected for binding toward a particular antigen in one organism. Since models trained on a huge corpus of natural proteins might be unsuitable for capturing scFv or antibody sequence information, we also employed an antibody-specific language model, AntiBERTy,[28] for the thermostability prediction task. We found that although zero-shot embeddings hardly clustered based on their thermal attributes, fine-tuning over the antibody-specific embeddings improved correlation with thermostability and its generalizability to newer, blind datasets. By equipping antibody-specific language models (e.g.

AbLang,[29] Progen2-OAS,[31] AntiBERTa,[37]) or AntiBERTy,[28]) we can gauge and fine-tune the antibody feature landscape to better explore the biophysical attributes for scFv and antibodies.

Recent studies have demonstrated that structure-specific information improves prediction quality for antigen–antibody binding tasks.[38,39] Extrapolating this for thermostability prediction, we incorporated structural context to our supervised models. To better learn generalizable representations, we sought to enrich the structural context of scFv sequences with thermodynamic (energy) features. We focused on the residue–residue energy features for each putative structure generated from the scFv sequence and used these energy-dependent features to train our CNN model. Validation of this small, supervised CNN network demonstrates the ability of energy-features to be more generalizable and predictive toward thermostability. For the relatively small dataset of 2,700 scFv sequences, we found that thermodynamic context could infer biophysical attributes independent of sequence origin (i.e. experimental sets from different germlines). Note that our training data are non-uniform, with some experimental sets skewed largely toward higher temperature bins owing to the selection procedure for generating the scFv sequences. With stringent selection and screening of sequences for uniform distribution in the temperature bins, along with incorporation of negative data (i.e. variant sequences that did not express, showed drastically low TS50 values or unfolded at room temperature), we could generate more well-distributed datasets for training ML models. Additionally, utilizing accurate three-dimensional antibody structures could refine the energetic input and potentially improve performance. Better quality of experimental and structural data may provide better predictions of thermostability.

In spite of the networks being trained for classification tasks, there are avenues to extend these models toward biologics engineering and design. With this work, we demonstrated how we could use this classification network to filter and suggest thermostability-enhancing point mutations. Although the temperature point mutant dataset was sparse, our models demonstrated substantial predictive accuracy toward mutants with higher thermal aggregation (Tm). One might argue that our network is trained on TS50 measurements, and so evaluating mutants with improved melting temperature is not plausible. However, as both the metrics evaluate the thermal attributes of the sequence, we can extend the patterns learned by our networks over alternative temperature-dependent data (i.e. thermal aggregation). Among the PTLMs and the SCNNs, we observed that the language models failed to strongly discriminate the point mutants. As language models utilize the underlying relationship in sequences, the small changes from point mutations might go unpenalized. On the other hand, SCNNs trained with energy features showed a significant effect of point mutations on thermostability prediction, as each point mutation affects local pairwise interactions in its vicinity.

With this work, even with training on sparse experimental data, we want to highlight the use of ML models toward evaluating an essential biophysical characteristic. Energy features represent a refined, information-rich resource that can add thermodynamic context that ML models are often deprived of. We have demonstrated a proof-of-concept of using PTLMs and SCNN architectures for thermostability prediction. Moreover, these models can also be equipped with experimental information derived from alternate physical properties (e.g. viscosity, binding enrichment, etc.), thereby enabling the engineering and design of antibodies and broad-scale biologics.

Materials and methods

Experimental methods

Generation of scFvs

scFvs with a (G4S)3 linker were cloned as a single construct into a pTT vector with a puromycin selection marker. Constructs were transfected into a mammalian CHO-K1 cell line and stably expressed at a 4 mL scale. After 21 days post transfection, VCD and viability were measured and the expression level of secreted proteins in conditioned medium were analyzed by non-reduced SDS PAGE gel. Cells were further incubated with magnetic beads coupled with either proA (for scFvs with lambda variable domains) or proL (for scFvs with kappa variable domains) overnight. The beads were separated from cell media and followed by washed with PBS for three times and water for 2 times. scFvs were eluted from the magnetic beads with a low pH buffer (100 mM glycine, pH2.7) and neutralized with 3 M Tris (pH11). Differential Scanning Fluorimetry (DSF) was carried out to determine the melting temperature of the purified material. Briefly, molecules were heated at 1.0°C/min on a nanoDSF instrument. Changes in tryptophan fluorescence were monitored to evaluate protein unfolding and aggregation. The Tm is reported as the midpoint between the unfolding onset and the max unfolded state.

TS50 screening assay

The thermostability of scFvs was screened by determining the loss of target binding after high-temperature stress. To this end, soluble scFvs (VH-(G4S)3-VL) containing a C-terminal FLAG-tag (DYKDDDDK) and a 6xHis-tag were produced in *E. coli* TG1 (Agilent, Santa Clara, USA) in 10 mL LB cultures. Protein production was induced with 1 mM IPTG. Bacteria were then centrifuged, and the cell pellet was re-suspended in 1 mL Gibco™ DPBS. Cells were lysed with four freeze/thaw cycles, and residual cells and cell debris were removed by two centrifugation steps. 100 μ L of these crude extracts were transferred into 0.2 mL tubes and subjected for 5 min to different temperatures in water-baths (4°C, 50°C, 60°C, 70°C). After incubation, the tubes were directly transferred on ice and human target transfected CHO-cells were incubated with 50 μ L of the lysates. Bound scFvs were detected and analyzed by flow cytometry. Median fluorescence intensity values were determined and plotted. The temperature corresponding to half maximal binding of each scFv was calculated (TS50). The scFv sequences were further binned into sets based on the identity of the antigen they bind (not random). Note that since the sets were not curated for an ML task, there is a lack of a uniform distribution across sets.

nanoDSF T_m method

Thermal melting (T_m) temperatures were determined by running a Trp Shift Study on the Prometheus, NT.48. A thermal ramp was applied at 1.0°C/min with start temperature 25°C and stop temperature with 95°C. Unfolding was measured by the fluorescence ratio 350 nm/330 nm. Data analysis and T_m determination were performed using PR. ThermControl v2.0.4. Samples were normalized to 1.0 mg/mL in formulation buffer prior to Tm analysis.

Dataset distribution

The thermostability data (TS50 measurements, thermal melting [T_m] temperatures) for this study was collected from historical single-chain variable fragment (scFv) therapeutic generation campaigns. The distribution of the scFv sequences across the experimental sets and the test dataset is illustrated in Sup. Fig. S1. Within the same experimental sets, there were occasional replicates (same sequence, different TS50 measurements) which were excluded from the training data. There is no sequence redundancy across experimental sets i.e. same sequence does not show up in different sets. The range of TS50 measurements are illustrated in Sup. Fig. S1.B and span from 25°C to 75°C. Since the TS50 corresponds to half-max binding temperature, the TS50 values for the sequences are discrete (Sup. Fig. S1) and are amenable toward a classification task. For alternative physical property measurements such as melting temperature which tend to be continuous, a regression model could be trained. For this work, owing to the discrete nature of the data, we have trained classification models to distinguish based on thermostability.

The test data (out-of-distribution dataset) includes test and isolated scFv sequences. Temperature measurements to test scFv sequences were performed with TS50, whereas, for isolated scFv, melting temperatures were obtained. We also test the general protein language model on a set of protein sequences from ProTherm.[13] Note that antibody-specific models, i.e. AntiBERTy-based language model and supervised language models, are not trained on the general protein sequences from ProTherm.

Pretrained language models

We evaluate four large-scale pretrained language models on their ability to predict the stability of scFv sequences. The first model, UniRep,[20] is an mLSTM[40] with 1900 hidden units pretrained on the Pfam database.[41] Following the "evotuning" methodology proposed by the authors, we collect MSAs for each sequence in our TS50 set, combine all sequences into a single dataset, and further pretrain the model on this evolutionarily related set of sequences using the implementation from.[42] Further, we consider both the ESM-1b[24] and ESM-1 v[25] transformer models. Both are 33 layer, 650 M parameter transformer models, pretrained with masked language modeling on the Uniref database.[43] The primary difference is that ESM-1b is trained on a 50% sequence identity filtered dataset (Uniref50), while ESM-1 v is trained on a 90% sequence identity filtered dataset (Uniref90). ESM-1 v is specifically designed to improve zero-shot likelihood evaluation of protein sequences. Finally, we also evaluate an antibody-specific language model, AntiBERTy[28] – a bidirectional transformer trained on 558 M natural antibody sequences from the Observed Antibody Space.[29]

Zero-shot evaluation

One approach to predicting stability with pretrained language models is to directly use model likelihood or pseudolikelihood.[25,27] Sequences which are more likely under a model are predicted to be more stable. Suppose $x = x_1 x_2 \ldots x_n$ is a protein sequence with each x_i representing a residue. UniRep models the probability of each residue given all preceding residues. As a result, the likelihood of a sequence can be efficiently evaluated as

$$\mathcal{L}_{\mathrm{UniRep}}(x) = \prod_{i=1}^{n} p(x_i | x_j \quad j < i) \tag{1}$$

ESM-1v models the probability of masked residues given unmasked residues. It is not possible to efficiently decompose this probability and obtain an exact likelihood. However, it is possible to obtain the pseudo-likelihood of a sequence:

$$\mathrm{pseudo} - \mathcal{L}_{\mathrm{ESM-1v}}(x) = \prod_{i=1}^{n} p(x_i | x_j \quad j \neq i) \tag{2}$$

In practice, the log-likelihood and pseudo-log-likelihood are evaluated for numerical stability. Additionally, ESM-1v comes as an ensemble of five models trained with different random seeds. The predictions from all five models are averaged to obtain the final pseudo-log-likelihood.

For AntiBERTy, we use the final layer embeddings to obtain prediction logits by iteratively masking each residue. The logits for each masked residue are used to estimate the categorical cross entropy (CCE) loss with respect to the actual token (amino acid residue at the masked position) and summed over the sequence to evaluate the pseudo-log-likelihood. Since AntiBERTy inputs are limited to heavy and light chains of antibodies, we split each scFv sequence across the linkers into heavy- and light-chain sequences. Prior to evaluating CCE loss, the prediction logits are concatenated to obtain a pseudo-likelihood for the entire scFv sequence.

Finetuned evaluation

The other approach to predicting stability with pretrained language models is to finetune a task-specific model using supervised data. For the UniRep model, we use the methodology suggested by the authors and take the final hidden state along with the average of previous hidden states as a fixed-length vector representation of 3900 hidden units. For the ESM-1b and the AntiBERTy model, we follow the methodology suggested by Detlefsen *et al.*[44] and downproject each per-residue representation to four dimensions, followed by a concatenation. This results in a fixed-length embedding of size $4L$, where L is the maximum sequence length in the TS50 dataset. If a sequence has length less than L, it is padded with zeros. Additionally, we also implemented attention-weighted pooling to summarize information from all residue positions into a fixed sized tensor thus skipping padding. However, the performance of attention-weighted pool was relatively worse than concatenation.

These embeddings are passed through a linear layer with a hidden dimension of 512, followed by tanh activation, then to a final layer to predict class logits. Parameters of the UniRep and ESM-1b models are frozen during training. Parameters of the head model (including the initial down-projection for ESM-1b) are trained with the Adam optimizer and a learning rate of 10^{-3}.

For TS50 data, models are trained on all but one target and evaluations are made on the held out target. For non-TS50 data, an ensemble of TS50 models (one for each holdout target) is used to make predictions.

Supervised models

Dataset curation

Sequence inputs: Datasets for TS50 measurements of scFvs from all experimental sets were aggregated to form a single dataset. The scFv sequences comprised a heavy and a light chain linked together with a Glycine-Serine (G4S) $_x$ linker. We chose to create a dataset with the scFv sequences, split into their respective heavy- and light-chain sequences, and instead of classifying sequences based on their thermostability, we included their TS50 measurements. Sets P and Q were removed, along with the sequences of the test scFv and the isolated scFv to constitute the held-out set. The amino acid sequences were one-hot encoded to form an input of dimension $(V_H + V_L + 3) \times 21$, where V_H and V_L correspond to the heavy- and light-chain sequences, respectively. The additional token to the amino acids' one-hot encoding corresponds to the delimiter at the start and end positions of the scFv sequence and between heavy and light chains to indicate a chain-break.

Energetic inputs: To obtain the energetics input, the sequences were first passed through a structural module, i.e. the DeepAb protocol for antibody structure prediction. For each predicted structure, we ran a Rosetta Relaxation and Refinement protocol for side-chain repacking (XML scripts in the Supplementary). Energy estimation in Rosetta starts with an energy relaxation step to reduce steric clashes (Rosetta Relax) with constraints to the start coordinates so that the accuracy of backbone structure (predicted by DeepAb) is not diminished. The all-atom model is refined further with four cycles of side-chain packing to obtain a robust structure, and the lowest energy structure is chosen for further calculations. For each refined model, we then estimated the residue–residue interaction energies with the residue_energy_breakdown application. We converted these one-body and two-body energies to a two-dimensional *i-j* matrix that served as the energetic information for training in the supervised CNN models. The energy values in the *i-j* matrix were also further binned together in 20 bins between the lower-end and upper-end energies of $[-25, 10]$ REU, respectively. We included an additional bin for the start, end and chain-break tokens, respectively. The dimension of the pairwise energy data is this, $L \times L \times 21$.

The energies are evaluated with the Rosetta ref2015 energy function.[32] Each energy value in the *i-j* matrix represents a contribution of that interaction to the total energy. The Rosetta total energy is determined by a linear combination of energy terms dependent on physical (LJ attractive and repulsive energies,[45] solvation,[46] electrostatics,[47]); empirical (hydrogen and disulfide bond energies[48,49]); statistical (backbone and side-chain torsion preferences[47,50]); and knowledge-based (amino-acid propensities,[50] rotamer energies[51]) parameters. Further details about the energy function and the energy terms are provided in *Supplementary Methods (Energy contributions).*

Model architecture

We evaluated supervised convolutional networks with sequence, energetics and sequence + energetic features to predict thermostability of scFv sequences. We use the scFv sequences to predict the structure (DeepAb[26]) and obtain energetic features with Rosetta. All the sequence and energy input is converted to a fixed length embedding of size L and $L \times L$ respectively, where L represents the maximum sequence length in the dataset, such that V_H and V_L are the maximum lengths of the heavy and light chains, respectively. The sequences less than L are padded with zeros. While padding the sequence and energy embedding are fed to two parallel branches of the model, one with a 1D convolutional layer and other with a 2D convolutional layer. The sequence and energetic input are fed to the network, such that sequences pass through a 1D CNN and energies pass through a 2D CNN, followed by concatenation. The output is then passed through another 2D convolutional layer, and then a final layer to produce the logits. We perform a *softmax* over the logits to obtain the class probabilities (Sup. Fig. S11). The parameters of the supervised model are trained with Adam optimizer with categorical cross entropy (CCE) loss and a learning rate of $10.^{-3}$

To estimate the predicted TS50 value (i.e. the regression task), the probabilities are weighed with the mean TS50 value in each bin. For the prediction of the temperature bins for a given sequence (prediction task), an *argmax* over the probabilities gives the expected thermostability (temperature bin).

$$\mathrm{TS50}_{\mathrm{predicted}}(x) = \sum_{i=1}^{n=4} p(x).\mathrm{TS50}_{\mathrm{bin}(i)} \tag{3}$$

Alternatively, we performed additional tests with different CNN architectures for the energy branch (1D-CNN by flattening the *i-j* matrix and 2D-CNN with absolute energy values, i.e. no binning). The architectures for these additional models were optimized by performing a randomized search for the parameters and variables, layers, dropout, batch size, number of filters, kernel size, strides, epochs and pooling size. The performance was assessed with Spearman's correlation coefficient, and we found that the 2D-CNN binned architecture for the energy branch worked better than other architectures. The 1D-CNNs resulted in loss of individual interactions as the average contribution by most of the residues is similar, and it was difficult for the network to discern useful context. For 2D-CNNs with absolute data, as better sequences are identified with lower, negative energies, the information was lost in the convolutions. 2D-CNN binned architecture resolves both the issues; it provided context of individual residue–residue interactions and the binning ensured that relevant energy information is retained through the convolutional layers. A comparison of these methods is further illustrated in the Sup. Fig. S2. As the CNN models are trained on a smaller dataset, to reduce variance, we use three CNN models trained on different sets to obtain an ensemble of CNNs, which we use for our predictions with the anti-VEGF antibody thermal denaturation data.

Extrapolating trained predictive models for design

The anti-VEGF DMS dataset was generated to enrich binding by designing multi-point variants. To determine whether our predictive models have potential in protein design, we created a computational DMS on the anti-VEGF antibody (PDB ID: 2FJG). Each residue position in the sequence was mutated to 19 other amino acids to obtain mutant sequences. Each sequence was one-hot encoded to obtain the sequence data, and the energetics dataset was generated following the procedure mentioned prior. This sequence and energy input were fed to the models, and the point mutants classified in the 70-up temperature bin by our predictions were cross-verified with the experimental results.

Acknowledgments

The authors thank Ai Ching Lim for her support of the project.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

AH and JJG were partially supported by the NIH R35 GM141881.

ORCID

Ameya Harmalkar http://orcid.org/0000-0001-6863-9634
Roshan Rao http://orcid.org/0000-0003-4412-3742
Yuxuan Richard Xie http://orcid.org/0000-0003-1664-9114
Julia Czwikla http://orcid.org/0000-0001-7856-789X
Austin J. Rice http://orcid.org/0000-0002-4165-4241
Christine E. Tinberg http://orcid.org/0000-0002-6179-0435
Jeffrey J. Gray http://orcid.org/0000-0001-6380-2324
Kathy Y. Wei http://orcid.org/0000-0002-8794-1385

Data availability statement

The source code for TherML (zero-shot, fine-tuned and supervised models) is available at https://github.com/AmeyaHarmalkar/therML for non-commercial use only. The experimental thermostability data and sequences are from internal antibody engineering studies and cannot be made available as the sequences are an intellectual property of Amgen. Any additional information required to reanalyze the data reported in this paper is available from the lead author upon request.

References

1. Spiess C, Zhai Q, Carter PJ. Alternative molecular formats and therapeutic applications for bispecific antibodies. Mol Immunol. 2015;67:95–106. doi:10.1016/j.molimm.2015.01.003.
2. Zhong X, D'Antona AM. Recent advances in the molecular design and applications of multispecific biotherapeutics. Antibodies (Basel, Switzerland). 2021;10. doi:10.3390/antib10020013.
3. Klinger M, Benjamin J, Kischel R, Stienen S, Zugmaier G. Harnessing T cells to fight cancer with BiTE® antibody constructs–past developments and future directions. Immunol Rev. 2016;270:193–208. doi:10.1111/imr.12393.
4. Dong J, Sereno A, Aivazian D, Langley E, Miller BR, Snyder WB, Chan E, Cantele M, Morena R, Joseph IBJK, et al. A stable IgG-like bispecific antibody targeting the epidermal growth factor receptor and the type I insulin-like growth factor receptor demonstrates superior anti-tumor activity. mAbs. 2011;3:273–88. doi:10.4161/mabs.3.3.15188.
5. Moore GL, Bernett MJ, Rashid R, Pong EW, Nguyen DHT, Jacinto J, Eivazi A, Nisthal A, Diaz JE, Chu SY, et al. A robust heterodimeric Fc platform engineered for efficient development of bispecific antibodies of multiple formats. Methods (San Diego, Calif). 2019;154:38–50. doi:10.1016/j.ymeth.2018.10.006.
6. Sawant MS, Streu CN, Wu L, Tessier PM. Toward drug-like multispecific antibodies by design. Int J Mol Sci. 2020;21:7496. doi:10.3390/ijms21207496.
7. Miller BR, Demarest SJ, Lugovskoy A, Huang F, Wu X, Snyder WB, Croner LJ, Wang N, Amatucci A, Michaelson JS, et al. Stability engineering of scFvs for the development of bispecific and multivalent antibodies. Protein eng des sel: PEDS. 2010;23:549–57.
8. Lapidoth GD, Baran D, Pszolla GM, Norn C, Alon A, Tyka MD, Fleishman SJ. AbDesign: an algorithm for combinatorial backbone design guided by natural conformations and sequences. Proteins: Struct Funct Bioinform. 2015;83:1385–406. doi:10.1002/prot.24779.
9. Goldenzweig A, Goldsmith M, Hill SE, Gertman O, Laurino P, Ashani Y, Dym O, Unger T, Albeck S, Prilusky J, et al. Automated structure- and sequence-based design of proteins for high bacterial expression and stability. Mol Cell. 2016;63:337–46. doi:10.1016/j.molcel.2016.06.012.
10. Warszawski S, Katz AB, Lipsh R, Khmelnitsky L, Nissan GB, Javitt G, Dym O, Unger T, Knop O, Albeck S, et al. Optimizing antibody affinity and stability by the automated design of the variable light-heavy chain interfaces. PLoS Comput Biol. 2019;15:1–24. doi:10.1371/journal.pcbi.1007207.
11. Kumar MDS, Bava KA, Gromiha MM, Prabakaran P, Kitajima K, Uedaira H, Sarai A. ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. Nucleic Acids Res. 2006;34:D204–6. doi:10.1093/nar/gkj103.
12. Nikam R, Kulandaisamy A, Harini K, Sharma D, Gromiha MM. ProThermDB: thermodynamic database for proteins and mutants revisited after 15 years. Nucleic Acids Res. 2020;49:D420–D424. doi:10.1093/nar/gkaa1035.
13. Gromiha MM, Suresh MX. Discrimination of mesophilic and thermophilic proteins using machine learning algorithms. Proteins. 2008;70:1274–79. doi:10.1002/prot.21616.
14. Jia L, Yarlagadda R, Reed CC, Zhang Y. Structure based thermostability prediction models for protein single point mutations with machine learning tools. PLoS ONE. 2015;10:1–19. doi:10.1371/journal.pone.0138022.
15. Yang Y, Urolagin S, Niroula A, Ding X, Shen B, Vihinen M . Pontstab: protein variant stability predictor. importance of training data quality. Int J Mol Sci. 2018;19(4):1009. doi:10.3390/ijms19041009.
16. Sarkisyan KS, Bolotin DA, Meer MV, Usmanova DR, Mishin AS, Sharonov GV, Ivankov DN, Bozhanova NG, Baranov MS, Soylemez O, et al. Local fitness landscape of the green fluorescent protein. Nature. 2016;533:397–401. doi:10.1038/nature17995.

17. Rao R, Bhattacharya N, Thomas N, Duan Y, Chen X, Canny J, Abbeel P, Song YS. Evaluating protein transfer learning with TAPE. Adv Neural Inf Process Syst. 2019;32:9689–701.

18. Shanehsazzadeh A, Belanger D, Dohan D. Is transfer learning necessary for protein landscape prediction? 2020;1–10.

19. Rocklin GJ, Chidyausiku TM, Goreshnik I, Ford A, Houliston S, Lemak A, Carter L, Ravichandran R, Mulligan VK, Chevalier A, et al. Global analysis of protein folding using massively parallel design, synthesis, and testing. Science. 2017;357:168–75. doi:10.1126/science.aan0693.

20. Alley EC, Khimulya G, Biswas S, AlQuraishi M, Church GM. Unified rational protein engineering with sequence-based deep representation learning. Nat Methods. 2019;16:1315–22. doi:10.1038/s41592-019-0598-1.

21. Gray VE, Hause RJ, Luebeck J, Shendure J, Fowler DM. Quantitative missense variant effect prediction using large-scale mutagenesis data. Cell Systems. 2018;6:116–124.e3. doi:10.1016/j.cels.2017.11.003.

22. Hsu C, Nisonoff H, Fannjiang C, Listgarten J. Learning protein fitness models from evolutionary and assay-labeled data. Nat Biotechnol. 2022;40:1114–22. doi:10.1038/s41587-021-01146-5.

23. Mason DM, Friedensohn S, Weber CR, Jordi C, Wagner B, Meng SM, Ehling RA, Bonati L, Dahinden J, Gainza P, et al. Optimization of therapeutic antibodies by predicting antigen specificity from antibody sequence via deep learning. Nat Biomed Eng. 2021;5:600–12. doi:10.1038/s41551-021-00699-9.

24. Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, Guo D, Ott M, Zitnick CL, Ma J, et al., Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*; 2021; Vol. 118. doi: 10.1073/pnas.2016239118.

25. Meier J, Rao R, Verkuil R, Liu J, Sercu T, Rives A. Language models enable zero-shot prediction of the effects of mutations on protein function. Adv Neural Inf Process Syst. 2021;34:35. https://proceedings.neurips.cc/paper/2021/hash/f51338d736f95dd42427296047067694-Abstract.html

26. Ruffolo JA, Sulam J, Gray JJ. 2021. Antibody structure prediction using interpretable deep learning. bioRxiv. doi:10.1016/j.patter.2021.100406.

27. Hopf TA, Ingraham JB, Poelwijk FJ, Schärfe CP, Springer M, Sander C, Marks DS. Mutation effects predicted from sequence co-variation. Nat Biotechnol. 2017;35:128–35. doi:10.1038/nbt.3769.

28. Ruffolo JA, Gray JJ, Sulam J. Deciphering antibody affinity maturation with language models and weakly supervised learning. Patterns (New York, N.Y.). 2021;3. doi:10.1016/j.patter.2021.100406.

29. Kovaltsuk A, Leem J, Kelm S, Snowden J, Deane CM, Krawczyk K. Observed antibody space: a resource for data mining next-generation sequencing of antibody repertoires. J Immunol. 2018;201:2502–09. doi:10.4049/jimmunol.1800708.

30. Riesselman AJ, Ingraham JB, Marks DS. Deep generative models of genetic variation capture the effects of mutations. Nat Methods. 2018;15:816–22. doi:10.1038/s41592-018-0138-4.

31. Nijkamp E, Ruffolo J, Weinstein EN, Naik N, Madani A. ProGen2: exploring the boundaries of protein language models. arXiv. 2022;2206.13517. doi:10.48550/arXiv.2206.13517.

32. Alford RF, Leaver-Fay A, Jeliazkov JR, O'Meara MJ, DiMaio FP, Park H, Shapovalov MV, Renfrew PD, Mulligan VK, Kappel K, et al. The Rosetta all-atom energy function for macromolecular modeling and design. J Chem Theory Comput. 2017;13:3031–48. doi:10.1021/acs.jctc.7b00125.

33. Koenig P, Lee CV, Walters BT, Janakiraman V, Stinson J, Patapoff TW, Fuh G. Mutational landscape of antibody variable domains reveals a switch modulating the interdomain conformational dynamics and antigen binding. *Proceedings of the National Academy of Sciences of the United States of America*; 2017; Vol. 114, p. E486–E495. doi:10.1073/pnas.1613231114.

34. Fuh G, Wu P, Liang WC, Ultsch M, Lee CV, Moffat B, Wiesmann C. Structure-function studies of two synthetic anti-vascular endothelial growth factor Fabs and comparison with the Avastin Fab. J Biol Chem. 2006;281:6625–31. doi:10.1074/jbc.M507783200.

35. Huang P, Chu SKS, Frizzo HN, Connolly MP, Caster RW, Siegel JB. Evaluating protein engineering thermostability prediction tools using an independently generated dataset. ACS Omega. 2020;5:6487–93. doi:10.1021/acsomega.9b04105.

36. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. The FoldX web server: an online force field. Nucleic Acids Res. 2005;33:W382–W388. doi:10.1093/nar/gki387.

37. Leem J, Mitchell LS, Farmery JH, Barton J, Galson JD. Deciphering the language of antibodies using self-supervised learning. bioRxiv. 2021. doi:10.1016/j.patter.2022.100513.

38. Akbar R, Robert PA, Pavlović M, Jeliazkov JR, Snapkov I, Slabodkin A, Weber CR, Scheffer L, Miho E, Haff IH, et al. A compact vocabulary of paratope-epitope interactions enables predictability of antibody-antigen binding. Cell Rep. 2021;34:108856. doi:10.1016/j.celrep.2021.108856.

39. Schneider C, Buchanan A, Taddese B, Deane CM, Valencia A. DLAB: deep learning methods for structure-based virtual screening of antibodies. Bioinformatics. 2021;38:377–83. doi:10.1093/bioinformatics/btab660.

40. Krause B, Murray I, Renals S, Liang L. Multiplicative lstm for sequence modelling in *5th International Conference on Learning Representations*; 2017; p. 2872–80; Toulon, France.

41. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, et al. The pfam protein families database. Nucleic Acids Res. 2004;32:D138–D141. doi:10.1093/nar/gkh121.

42. Ma EJ, Kummer A. Reimplementing unirep in jax. bioRxiv. 2020.

43. Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. Uniref: comprehensive and non-redundant UniProt reference clusters. Bioinformatics. 2007;23:1282–88. doi:10.1093/bioinformatics/btm098.

44. Detlefsen NS, Hauberg S, Boomsma W. What is a meaningful representation of protein sequences? *ArXiv preprint arXiv:2012.02679* (2020).

45. Jones JE, Chapman S. On the determination of molecular fields. —II. From the equation of state of a gas. Proceedings of the Royal Society of London Series A, Containing Papers of a Mathematical and Physical Character. 1924;106:463–77.

46. Lazaridis T, Karplus M. Effective energy function for proteins in solution. Proteins: Struct Funct Bioinform. 1999;35:133–52. doi:10.1002/(SICI)1097-0134(19990501)35:2<133::AID-PROT1>3.0.CO;2-N.

47. Park H, Bradley P, Greisen P, Liu Y, Mulligan VK, Kim DE, Baker D, Dimaio F. Simultaneous optimization of biomolecular energy functions on features from small molecules and macromolecules. J Chem Theory Comput. 2016;12:6201–12. doi:10.1021/acs.jctc.6b00819.

48. Kortemme T, Morozov AV, Baker D. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. J Mol Biol. 2003;326:1239–59. doi:10.1016/S0022-2836(03)00021-4.

49. O'Meara MJ, Leaver-Fay A, Tyka MD, Stein A, Houlihan K, DiMaio F, Bradley P, Kortemme T, Baker D, Snoeyink J, et al. Combined covalent-electrostatic model of hydrogen bonding improves structure prediction with Rosetta. J Chem Theory Comput. 2015;11:609–22. doi:10.1021/ct500864r.

50. Leaver-Fay A, O'Meara MJ, Tyka M, Jacak R, Song Y, Kellogg EH, Thompson J, Davis IW, Pache RA, Lyskov S, et al. Scientific benchmarks for guiding macromolecular energy function improvement. Methods Enzymol. 2013;523:109–43.

51. Shapovalov MV, Dunbrack RLJ. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. Structure (London, England: 1993). 2011;19:844–58. doi:10.1016/j.str.2011.03.019.