ProxyTm: Large-Scale Protein Thermostability Data Generation

Diffuse Bio Team

Abstract

We present **ProxyTm**, a high-throughput platform for measuring protein thermostability at library scale. Building off of our RamaX platform, we generate ProxyTm labels for a library of ∼10,000 nanobody (VHH) variants in a single 1-week experiment, representing **100–1000× lower cost and higher throughput than conventional methods**. To verify that ProxyTm accurately measures melting temperature (T_m), we validate ProxyTm against a panel of 545 VHH reference sequences, and we demonstrate that this data can be used to train significantly more performant AI models for T_m prediction. We find T_m prediction performance scales log-linearly with ProxyTm dataset size, and we corroborate ProxyTm data corresponds with known biophysical VHH framework features.

These results establish ProxyTm as a practical platform for accurately measuring T_m across large libraries, producing the thermostability datasets needed to train next-generation protein engineering AI models.

We make our ProxyTm-finetuned model available via API.

1 Background

Diffuse is collecting massive biological datasets to train our next generation of AI protein engineering models. AI model performance scales with training data, yet current models for predicting protein developability properties are trained on small public datasets. The rapid collection of large, accurate biological datasets is critical for advancing AI-driven protein design.

In this report, we present **ProxyTm**, a high-throughput thermostability measurement platform built on RamaX [1] that enables rapid protein melting temperature (T_m) measurement at library scale.

Protein thermostability – the resistance of a folded protein to thermal denaturation – is a critical developability property in biotherapeutic development [2]. A protein's function depends on maintaining its native three-dimensional structure, which unfolds above a characteristic melting temperature (T_m), defined as the temperature at which 50% of the protein population is unfolded. For antibody and nanobody (VHH) therapeutics, low thermostability can lead to aggregation, loss of binding activity, and poor manufacturability [3, 4].

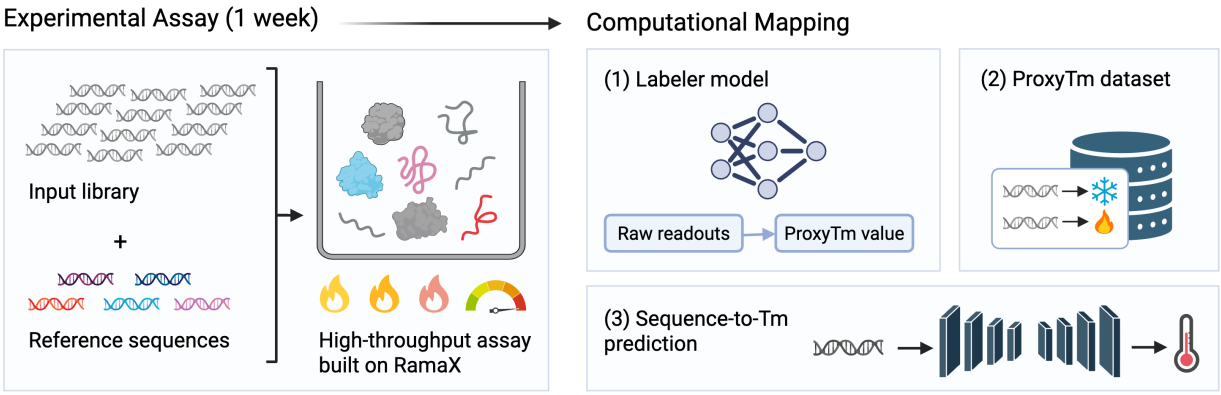


Figure 1 ProxyTm data generation overview. Starting with an input protein library and a small set of reference sequences with known T_m values, we run the ProxyTm assay built on the RamaX platform. A labeler model trained on reference sequences maps raw assay readouts to predicted T_m values. The labeler is then used to generate ProxyTm labels for the full library. ProxyTm-labeled data is used to train improved sequence-to-T_m models.

Consequently, T_m is routinely measured during lead optimization campaigns.

1.1 Current methods for T_m measurement

Standard methods for determining T_m are circular dichroism (CD) spectroscopy, differential scanning fluorimetry (DSF), nanoDSF, and differential scanning calorimetry (DSC) [6]. While these methods produce accurate measurements, they require purified protein and are practically limited to tens or hundreds of measurements per campaign. A typical nanoDSF measurement, including protein expression and purification, costs approximately $350 per datapoint and requires 3–4 weeks to express, purify, and measure ∼100–1000 proteins.

Previous efforts have sought to measure thermostability at scale — most notably the Meltome Atlas, which applied thermal proteome profiling to measure T_m values across entire proteomes [7]. TemBERTure, a language model trained on Meltome data, demonstrates that large-scale thermal stability measurements can improve predictive models [8]. In this report, we investigate whether models trained on such proteome-focused data generalize to VHH sequences and find that they do not. Furthermore, mass spectrometry-based approaches cannot resolve closely related variants within a protein family. In contrast, ProxyTm provides single-sequence resolution even across highly self-similar variants, a common need in antibody engineering.

1.2 Available VHH thermostability data

Public VHH-specific thermostability data remains extremely limited. The NbThermo database contains 564 VHH sequences with reported T_m values ranging from 40–100°C, collected across multiple studies using various methodologies including CD spectroscopy, DSF, nanoDSF, and DSC [10]. The TEMPRO dataset extends this to approximately ∼700 sequences by combining NbThermo with additional sequences from multiple studies [11].

1.3 Sequence-based VHH T_m prediction models

Predicting T_m directly from sequence enables *in silico* screening of vastly more candidates than can be characterized experimentally, expanding the design space that can be explored during lead optimization.

Several computational approaches have been developed to predict VHH melting temperature from sequence. NanoMelt uses an ensemble of regression models trained on four diverse protein language model embeddings, reporting Spearman $\rho = 0.83$, MAE $= 4.1°$C, and SDR $= 0.86$ on cross-validated test sets of ∼640 datapoints [12]. NBsTem combines molecular dynamics simulations with language model features, reporting $R = 0.83$ and MAE $= 2.3°$C [13]. TemBERTure fine-tunes a BERT-based model on the Meltome Atlas for general protein T_m prediction [8]. The recent NbBench benchmark systematically evaluated frozen antibody language model embeddings on nanobody T_m prediction, finding Spearman correlations of 0.39 (AntiBERTy) to 0.59 (AntiBERTa2) — underscoring the difficulty of the task with current public datasets [9].

A fundamental limitation shared by all current approaches is the scarcity of training data relevant for biotherapeutics. With fewer than ∼700 public VHH T_m measurements available, model performance is constrained by dataset size rather than model capacity. Scaling the collection of accurate thermostability data is critical for improving prediction models.

2 The ProxyTm Platform

Given the scarcity of large-scale thermostability data for training predictive models, we developed a new high-throughput assay to enable T_m measurement at large scales. Previously, we established the RamaX [1] platform to quickly screen large libraries for binder discovery. We built ProxyTm on top of RamaX to leverage its advantages in speed and scale, while introducing a simple neural network to map the raw readouts of the RamaX assay to a physical melting temperature value.

2.1 ProxyTm assay design

The ProxyTm workflow proceeds in three stages (Figure 1). First, a protein library of interest is screened on the RamaX platform under varying experimental conditions, yielding raw readout values for each sequence. Second, a simple labeler model is trained on the raw experimental features for a panel of 545 reference sequences with independently measured T_m values spiked into the experiment. Third, the trained labeler assigns ProxyTm values (in °C) to all remaining sequences in the library. These values are proxy T_m values, estimated from high-dimensional experimental readouts – hence the name of the method.

2.2 Scale, cost, and throughput

ProxyTm enables thermostability measurement at scales, speeds, and costs that are inaccessible to conventional methods:

	CD Spectroscopy	nanoDSF	ProxyTm	**Improvement**
Approx. datapoints per week	100	100	10K–1M	**>1,000×**
Approx. cost per datapoint	$500	$300	$0.10–$1	**100–1,000×**

Table 1 Comparison of throughput and cost across T_m measurement methods. Costs include protein expression and purification.

A single ProxyTm experiment in 1 week produces >10× more VHH T_m measurements than the entire

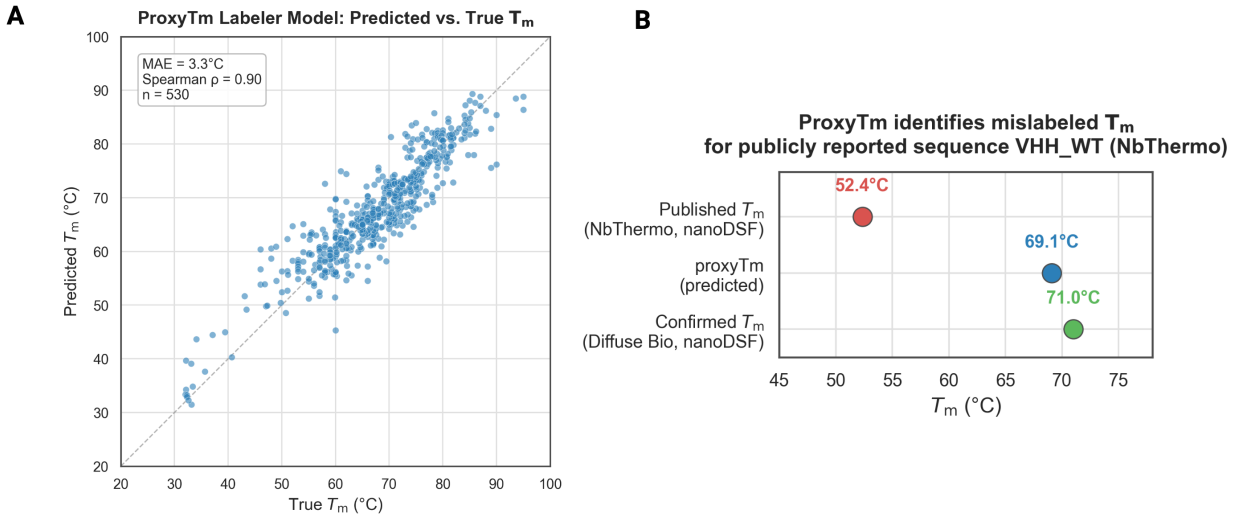


Figure 2 ProxyTm is an accurate proxy measure for true melting temperature. (A) ProxyTm labels (predicted T_m, °C) versus ground-truth T_m for 530 reference sequences. Plot shows 20-fold cross-validation; all points are hold-out predictions. (B) ProxyTm identifies a mislabeled entry in the NbThermo database. Sequence VHH_WT is reported at 52.35°C via nanoDSF. ProxyTm measures 69.09°C. Independent NanoTemper measurement confirms $T_m = 71.00$°C.

NbThermo database [10].

3 ProxyTm Accurately Measures Melting Temperature

3.1 Validating and calibrating ProxyTm with ground-truth T_m measurements

To validate the ProxyTm assay, we spiked 545 VHH sequences with known T_m values into the screening library. These reference sequences span a T_m range of approximately 30–90°C and include entries from the NbThermo and TEMPRO datasets [10, 11].

We trained a labeler model on the raw ProxyTm readouts for these reference sequences and assessed performance via 20-fold cross-validation after outlier removal (see Section A.1 for details). Each datapoint shown represents a hold-out prediction (Figure 2A). The labeler achieves an MAE of 3.3°C and Spearman $\rho = 0.90$ ($n = 530$), demonstrating that the ProxyTm assay produces readouts with strong, quantitative correspondence to true melting temperatures.

For inference on the full VHH library with unknown T_m values, predictions from all 20 fold-level models were averaged to yield a final predicted T_m per sequence. Sequences were retained only if both ensemble uncertainty (standard deviation across folds) and replicate uncertainty (mean absolute error across two experimental replicates) were below 3°C, yielding a final ProxyTm dataset of ∼8,000 datapoints.

3.2 Identification of mislabeled public data with ProxyTm

While training the labeler model, we identified cases where ProxyTm labels diverged substantially from publicly reported values. We investigated one case in detail: NbThermo sequence VHH_WT, reported in the NbThermo database with a T_m of 52.35°C measured via nanoDSF [10]. The ProxyTm label for this sequence was 69.09°C — a discrepancy of nearly 17°C.

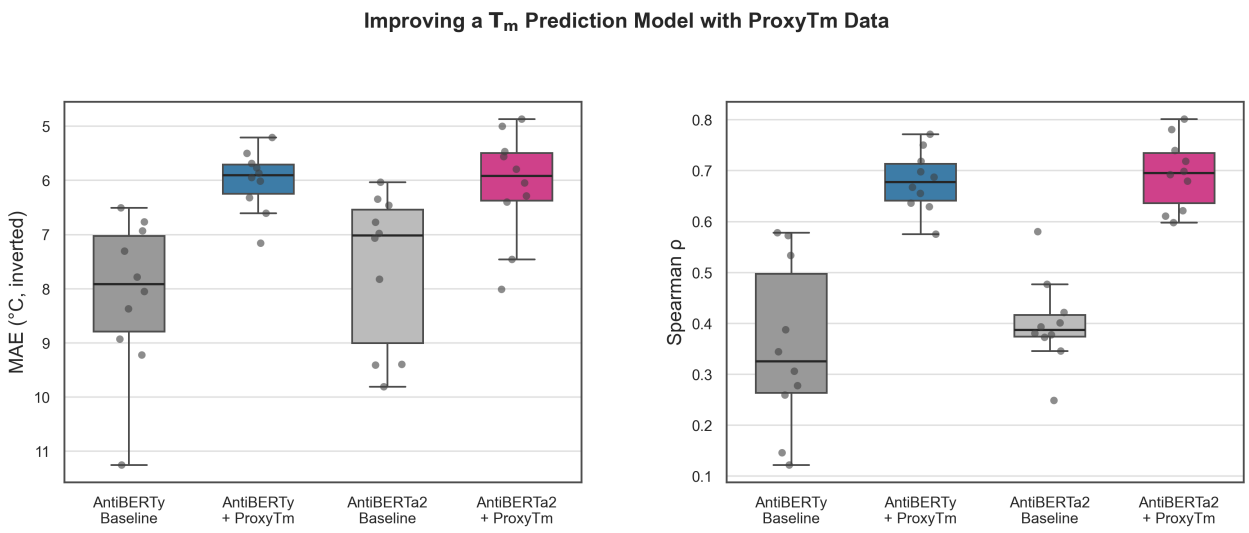


Figure 3 ProxyTm data trains significantly more performant T_m prediction models. (Left) MAE (°C, inverted axis — higher is better) and (Right) Spearman ρ (higher is better) across 10 seeds for AntiBERTy and AntiBERTa2, with and without ProxyTm augmentation.

To investigate this, we independently measured the T_m of VHH_WT again using nanoDSF[1] and obtained a value of 71.00°C, closely matching the ProxyTm prediction and contradicting the database entry (Figure 2B).

That ProxyTm was able to flag this mislabeled datapoint speaks to the accuracy of the underlying assay. We are continuing to investigate additional sequences where ProxyTm predictions diverge from reported values.

4 ProxyTm Data Trains Superior T_m Prediction Models

We next asked whether ProxyTm data, despite being a proxy measurement rather than a direct T_m determination, contains sufficient information to improve T_m prediction AI models beyond what is possible with public data alone.

4.1 Training on ProxyTm data leads to significant improvements in thermostability prediction

We fine-tuned two pretrained antibody language models — AntiBERTy [14] and AntiBERTa2 [15] — for T_m regression on the collected ∼8K ProxyTm dataset (see Section A.2 for details). We compared two training conditions per base model:

- **Baseline**: trained on aggregated public data (∼700 sequences, cluster-split 50/25/25 train/val/test) [10, 11].
- **Baseline + ProxyTm**: trained on the same public data plus ∼8,000 high-confidence ProxyTm datapoints, cluster-split 95/5.

Adding ProxyTm data substantially improves predictive performance across both base models (Figure 3, Supplementary Figure 1).

[1] Using NanoTemper Prometheus Panta instrument

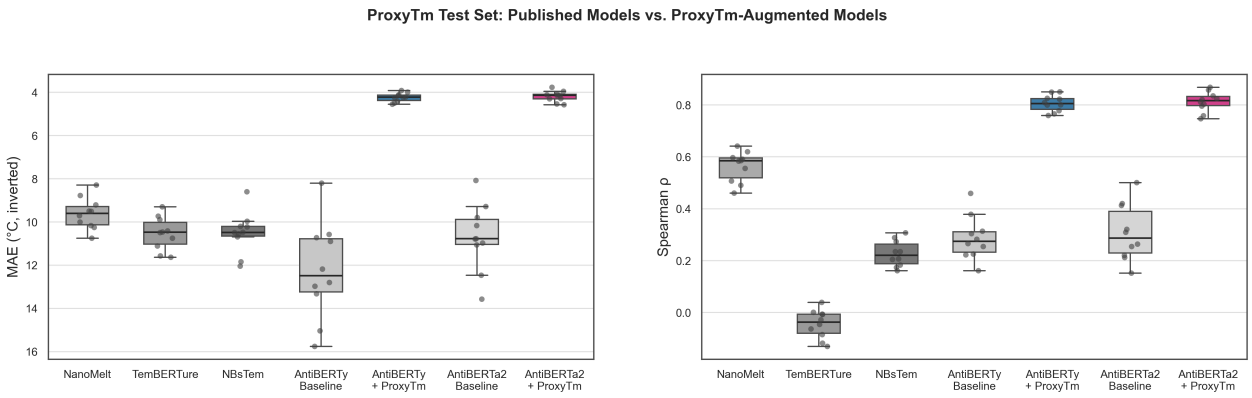


Figure 4 Current open-source VHH T_m prediction models fail to generalize to ProxyTm data. Models are evaluated on ProxyTm held-out data. (Left) MAE (°C, inverted — higher is better) and (Right) Spearman ρ (higher is better) across 10 seeds. Fine-tuning on ProxyTm data substantially improves both metrics for both base models, while published baselines trained without ProxyTm data show poor generalization to ProxyTm sequences.

The improvement is consistent across seeds, with substantially reduced variance in the ProxyTm-augmented models. These results demonstrate that ProxyTm data captures thermostability information beyond what is available in current public datasets, and this signal transfers effectively to improve predictive models.

We make the fine-tuned AntiBERTa2 model available via API.

4.2 Current published thermostability prediction models fail to generalize beyond small public datasets

We cannot directly compare our models to published baselines head-to-head, as each was trained on different splits of the available public data. Instead, we ask whether open-source models generalize to the larger, more diverse ProxyTm-labeled dataset.

ProxyTm values are noisier than gold-standard T_m measurements, but given the validated correspondence between ProxyTm labels and reported T_m (Section 3.1), we expect that T_m prediction models should still generalize to ProxyTm data.

We evaluated three publicly available open-source VHH T_m prediction models — NanoMelt [12], NBsTem [13], and TemBERTure [8] — on ProxyTm-labeled held-out data and compared their performance to our ProxyTm-augmented models.

On ProxyTm test data, the published models exhibit substantially higher error than our ProxyTm-augmented models (Figure 4, Supplementary Figure 2). By contrast, our ProxyTm-augmented models generalize well to hold-out ProxyTm data.

These published models report competitive performance on their respective test sets; however, they fail to generalize to the larger and more diverse ProxyTm dataset. TemBERTure [8], trained primarily on non-antibody proteins from the Meltome Atlas [7], shows near-zero rank correlation on VHH sequences.

A natural concern here is circular evaluation: models trained on ProxyTm labels outperform baselines on ProxyTm held-out data, but this could simply reflect that the models have learned to reproduce ProxyTm-specific biases rather than generalizable thermostability signal. While these biases almost certainly exist, we've already shown that ProxyTm labels correlate with ground-truth measurements (Figure 2), and that ProxyTm data captures real thermostability signal that lets us train better T_m prediction models (Figure 3). Therefore, we expect that an accurate T_m prediction model should still generalize to ProxyTm data.

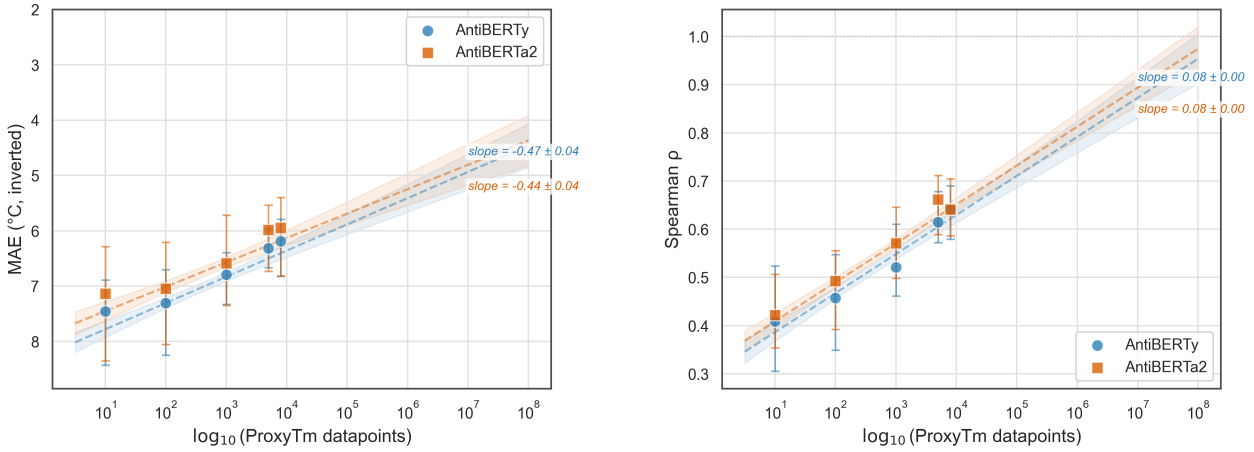


Figure 5 Thermostability predictive performance scales log-linearly with ProxyTm data. MAE and Spearman ρ as a function of ProxyTm fine-tuning samples (log scale) for AntiBERTy and AntiBERTa2. 100 seeds per data scale. Dashed lines show log-linear fits.

The generalization gap for models trained on limited public data underscores the need for larger, more diverse training datasets. ProxyTm provides a practical path to generating such datasets.

4.3 Thermostability prediction performance scales with ProxyTm data

A key advantage of ProxyTm is the ability to generate large datasets efficiently. To understand how model performance scales with data volume, we trained models on subsets of the ProxyTm data at logarithmically spaced intervals: 10, 100, 1,000, 5,000, and the full ∼8,000-datapoint set, comparing against the public-data-only baseline. We report metric performance across 100 seeds per data scale (Figure 5).

Both MAE and Spearman ρ improve approximately log-linearly with the number of ProxyTm fine-tuning samples, and this relationship holds across both AntiBERTy and AntiBERTa2 base models with nearly identical slopes.

These scaling properties suggest that further data collection — readily achievable with the ProxyTm platform — will yield continued performance gains. The consistency of the scaling law across architectures indicates that the improvement is driven by data quality and diversity rather than model-specific effects.

5 Large-Scale ProxyTm Dataset Elucidates Framework and CDR Contributions to VHH Thermostability

The scale of the ProxyTm dataset opens the door to analyses that require large, diverse sequence coverage. As an initial investigation, we asked how framework and CDR regions each contribute to VHH thermostability, leveraging ∼13,600 ProxyTm-labeled sequences across >3,200 framework clusters (Figure 6, Section A.3).

For any given framework cluster, we see wide variance in measured T_m — typically spanning 20–40°C — demonstrating that CDR sequences substantially influence thermostability. Even the most stable framework cluster (cluster 7, mean $T_m = 77$°C) spans a range of >20°C across its CDR variants. A predictive model must therefore attend to both framework and CDR regions.

Comparing the highest- and lowest-stability framework clusters with ≥10 sequences reveals intuitive findings that recapitulate known structural biology in a quantitative manner:

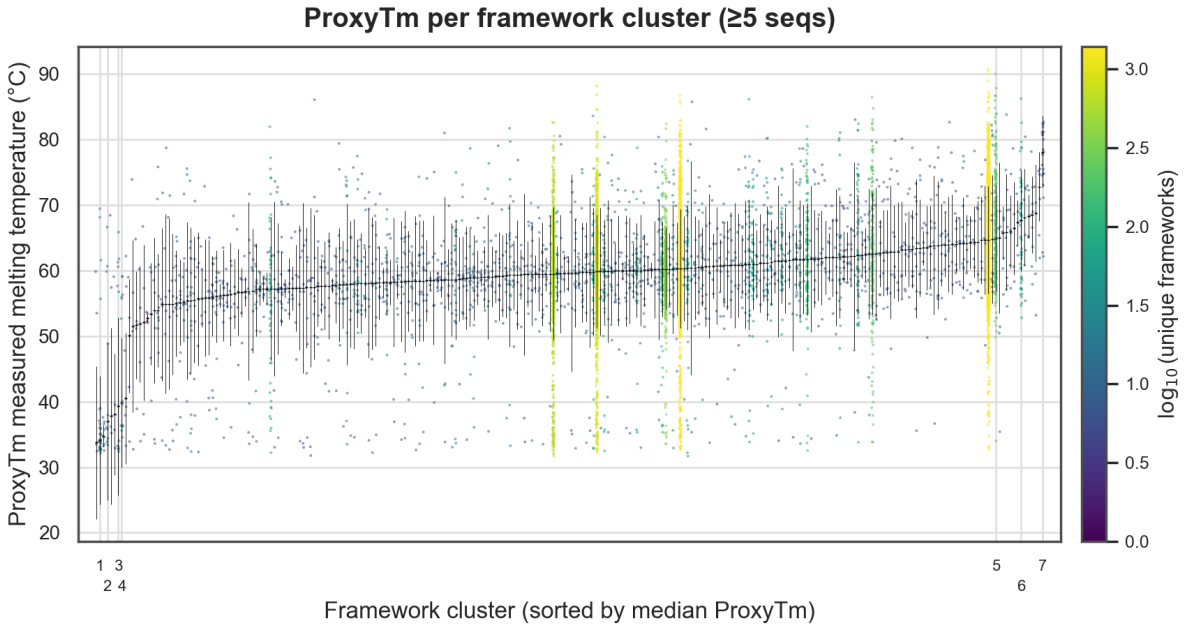


Figure 6 ProxyTm distributions per framework cluster ($\geq$5 sequences), sorted by median T_m. Points are individual VHH sequences colored by $\log_{10}$(unique frameworks per cluster); black lines indicate cluster median $\pm$ standard deviation. Labeled clusters are referenced in the text. Wide per-cluster T_m variance demonstrates that CDR composition substantially modulates stability within any given framework.

- **Critical structural features (>20°C penalty upon loss).** Frameworks lacking the conserved Cys23–Cys104 intradomain disulfide bond (clusters 1, 2) have low predicted T_m values, regardless of other framework composition. Truncation of FR4 (cluster 2) is similarly catastrophic, removing the conserved β-barrel cap.
- **VHH versus VH framework identity (∼15–20°C effect).** Frameworks with VH-type FR2 residues (clusters 3, 4) — which evolved to pair with a VL domain and expose hydrophobic residues at the former interface — are systematically less stable than frameworks with canonical VHH hallmark residues (clusters 5, 6). This is the dominant stability modulator among structurally intact frameworks, consistent with the known requirement for VHH-specific adaptations for folding [3, 4].
- **Core packing optimization (∼5–10°C effect).** The most stable cluster (cluster 7, mean 77°C) is distinguished by framework substitution R78A that improves hydrophobic core packing and reduces buried charge penalties, yielding an additional 5–10°C over other intact VHH frameworks. Secondary contributors include Val at IMGT 24 for improved packing around the disulfide anchor and Ala at IMGT 83 for reduced steric strain.

The scale of ProxyTm data surfaces clear, intuitive trends in framework thermostability — from the dominant effect of the internal disulfide to subtler contributions of VHH-type FR2 hallmark residues and core packing variants. Yet the wide T_m distributions within even the most conserved framework clusters underscore that stability is ultimately determined by the interplay of framework and CDR residues, not framework identity alone. These analyses illustrate how ProxyTm enables data-driven investigation of sequence–stability relationships at library scale.

6 Discussion

In this report, we introduce ProxyTm, a novel high-throughput assay for measuring protein thermostability at scale. The assay produces accurate T_m measurements, identifies errors in public databases, and generates data that meaningfully improves AI T_m prediction models — all at 100–1,000× lower cost and higher throughput than conventional methods.

ProxyTm is a proxy measurement, not a direct biophysical determination, and inherits noise from the underlying experimental assay. While this noise does not prevent training substantially improved thermostability models, it may limit applications requiring sub-degree resolution, such as ranking closely related point mutants. We are actively working to improve the resolution of the assay.

For future work, we plan to scale ProxyTm to larger library sizes, as well as to extend beyond VHHs to other protein modalities. We are also developing assays for additional developability properties (e.g., polyreactivity, expression) on the same platform. Rapid, cost-effective generation of large biophysical datasets will be critical for powering the next generation of AI protein design models — much as large-scale data collection has driven progress in other areas of AI.

Contributions and Acknowledgements

Contributors, ordered alphabetically by first name: Allison Cooke, Carla Perez*, Davian Ho, Jack Wang, Jessica Pan, Jonas Lindquist, Michael Young*, Namrata Anand, Tarun Prasad (*Lead contributors).

Figures created in BioRender (https://BioRender.com).

A Methods

A.1 Labeler model training

A reference panel of 545 VHH sequences (2 replicates each) with independently measured T_m values was spiked into each ProxyTm screening library. Reference sequences span ∼30–90°C and include entries from the NbThermo database [10] and the TEMPRO dataset [11]. The labeler is a simple MLP regression model, mapping raw ProxyTm readouts to predicted T_m.

Performance was assessed via 20-fold cross-validation. Replicates were treated as separate datapoints and mapped to the same fold.

After the initial cross-validation pass, controls with absolute prediction error exceeding 15°C were flagged as outliers and removed, with the exception of sequences with reference T_m below 40°C, which were retained to preserve coverage of the low-T_m regime. The full 20-fold procedure was then re-executed on the cleaned dataset. Outlier removal was performed once and not iterated.

For inference, predictions from all 20 fold-level models were averaged to yield a final predicted T_m per sequence. Two uncertainty metrics were computed: the standard deviation across fold-level predictions (ensemble uncertainty) and the mean absolute error between experimental replicates (replicate uncertainty). Predictions were retained only if both values were below 3°C; sequences exceeding either threshold were excluded from downstream analysis.

A.2 Model fine-tuning and evaluation

We fine-tuned two pretrained antibody language models — AntiBERTy [14] and AntiBERTa2 [15] — for T_m regression on the collected ∼8K ProxyTm dataset:

AntiBERTy [14]: an 8-layer BERT model (hidden dimension 512, ∼26M parameters) pretrained on paired and unpaired antibody sequences.

AntiBERTa2 [15]: a 12-layer RoFormer-based model (hidden dimension 1024, ∼202M parameters) pretrained on antibody sequences. The same training procedure was applied, with the MLP head adjusted for input dimension: Linear(1024, 16) → ReLU → Dropout → Linear(16, 1).

For both base models, we fully fine-tuned all backbone parameters with an MLP regression head, using mean pooling over sequence tokens and MSE loss.

Both models were trained with AdamW (weight decay 0.01) using a differential learning rate — 1×10^{-5} for backbone parameters and 1×10^{-4} (10×) for the head — with cosine annealing over 20 epochs at batch size 16. To prevent data leakage between related sequences, we clustered the baseline public dataset (∼700 sequences from NbThermo/TEMPRO) using MMseqs2 with minimum sequence identity 0.80 and coverage 0.80. Clusters were assigned to train/val/test splits (50/25/25) by seed hash, ensuring that all members of a cluster appear in the same split. When adding ProxyTm data, new sequences were assigned to existing clusters where possible; sequences forming novel clusters not represented in the original public data were split by cluster 95/5 train/test. This asymmetric split reflects the abundance of ProxyTm data relative to public data — validation and hyperparameter selection were performed on the public-data validation set to maintain a consistent evaluation baseline across conditions.

A.3 Framework T_m analysis

We leveraged ∼13,600 ProxyTm-labeled VHH sequences to investigate the relative contributions of framework and CDR regions to thermostability.

We extracted framework regions (FR1–FR4) from each sequence using IMGT numbering [17] via ANARCI [18], removing CDR1–3 to yield 12,794 unique framework sequences. Frameworks were grouped into 3,238 clusters at 90% sequence identity using greedy incremental clustering on Levenshtein distance. We then examined the distribution of ProxyTm values within each cluster (Figure 6).

References

[1] Diffuse Bio Team. RamaX: Ultra-fast and accurate protein binder discovery, screening, and optimization. Technical report, 2025. https://ramax.diffuse.bio/public/diffuse_ramax_report.pdf.

[2] Jain, T. et al. Biophysical properties of the clinical-stage antibody landscape. Proc. Natl. Acad. Sci. **114**, 944–949 (2017).

[3] Dumoulin, M. et al. Single-domain antibody fragments with high conformational stability. Protein Sci. **11**, 500–515 (2002).

[4] Ewert, S., Honegger, A. & Plückthun, A. Stability improvement of antibodies for extracellular application: CDR grafting to stable frameworks and structure-based framework engineering Methods **34**, 184–199 (2004).

[5] Greenfield, N. J. Using circular dichroism spectra to estimate protein secondary structure. Nat. Protoc. **1**, 2876–2890 (2006).

[6] Niesen, F. H., Berglund, H. & Vedadi, M. The use of differential scanning fluorimetry to detect ligand interactions that promote protein stability. Nat. Protoc. **2**, 2212–2221 (2007).

[7] Jarzab, A. et al. Meltome atlas — thermal proteome stability across the tree of life. Nat. Methods **17**, 495–503 (2020).

[8] Rodella, C., Lazaridi, S. & Lemmin, T. TemBERTure: advancing protein thermostability prediction with deep learning and attention mechanisms. Bioinform. Adv. **4**, vbae103 (2024).

[9] Zhang, Y. and Tsuda, K. NbBench: Benchmarking Language Models for Comprehensive Nanobody Tasks. *arXiv preprint arXiv:2505.02022*, 2025.

[10] Valdés-Tresanco, M. S. et al. NbThermo: a new thermostability database for nanobodies. Database **2023**, baad021 (2023).

[11] Alvarez, J.A.E. & Dean, S.N. TEMPRO: nanobody melting temperature estimation model using protein embeddings. Sci. Rep. **14**, 19074 (2024).

[12] Ramon, A., Ni, D., Predeina, A., Gaffey, C., Kunz, L., Onuoha, S. & Sormanni, P. Prediction of protein biophysical traits from limited data: a case study on nanobody thermostability through NanoMelt. mAbs **17**, 2442750 (2025).

[13] Mao, J. et al. Thermostability prediction powered by synergistic deep learning at experimental and theoretical levels for nanobodies. ACS Appl. Mater. Interfaces (2026).

[14] Ruffolo, J. A. et al. Deciphering antibody affinity maturation with language models and weakly supervised learning. arXiv:2112.07782 (2021).

[15] Barton, J. et al. Enhancing antibody language models with structural information. bioRxiv, 2023.12.12.569610 (2024).

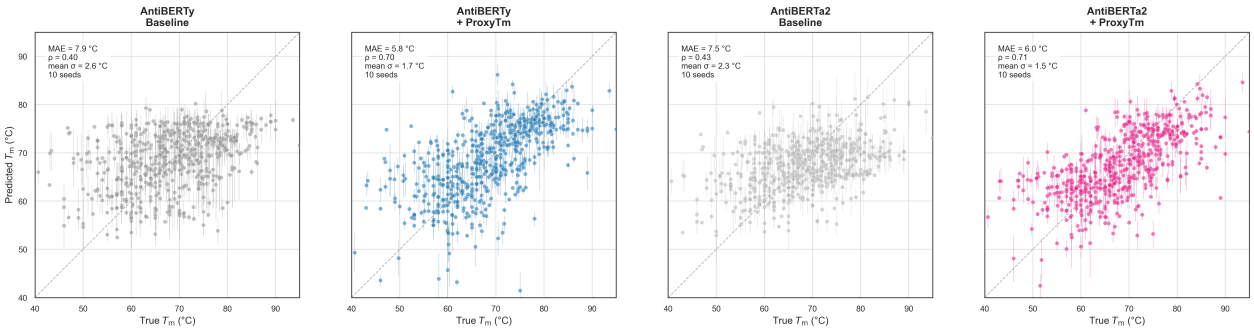
[16] Notin, P. et al. ProteinGym: large-scale benchmarks for protein fitness prediction and design. Adv. Neural Inf. Process. Syst. **36** (2023).

[17] Lefranc, M.-P. et al. IMGT, the international ImMunoGeneTics information system. Nucleic Acids Res. **37**, D1006–D1012 (2009).

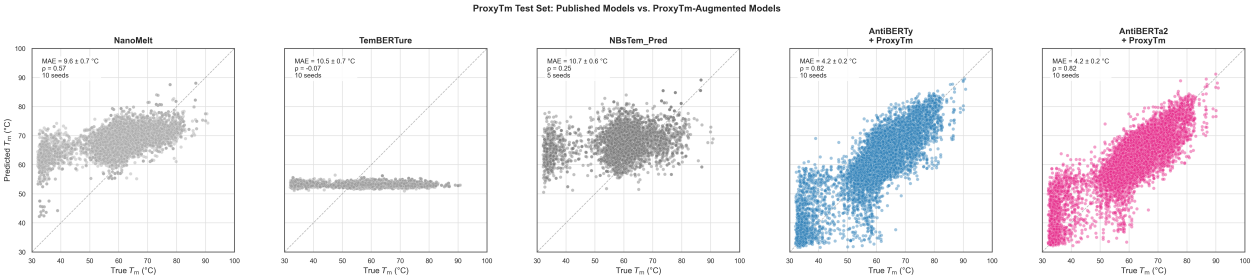
[18] Dunbar, J. & Deane, C. M. ANARCI: antigen receptor numbering and receptor classification. Bioinformatics **32**, 298–300 (2016).

[19] Vincke, C. et al. General strategy to humanize a camelid single-domain antibody and identification of a universal humanized nanobody scaffold. J. Biol. Chem. **284**, 3273–3284 (2009).

[20] Saerens, D. et al. Identification of a universal VHH framework to graft non-canonical antigen-binding loops of camel single-domain antibodies. J. Mol. Biol. **352**, 597–607 (2005).



Supplementary Figure 1 **Predicted versus true** T_m (°C) for each model condition evaluated on the public test set (TEMPRO, 50/25/25 cluster split). Each panel shows predictions from 10 independent seeds overlaid, with error bars indicating the standard deviation across seeds for each sequence. (From left to right) AntiBERTy Baseline, AntiBERTy + ProxyTm, AntiBERTa2 Baseline, AntiBERTa2 + ProxyTm. ProxyTm-augmented models show tighter clustering around the diagonal and reduced inter-seed variance.



Supplementary Figure 2 **Predicted versus true** T_m (°C) on ProxyTm held-out test data for published models and ProxyTm-augmented models. From left: NanoMelt, TemBERTure, NBsTem, AntiBERTy Baseline, AntiBERTy + ProxyTm, AntiBERTa2 Baseline, AntiBERTa2 + ProxyTm. Published models show large scatter and systematic bias, while ProxyTm-augmented models cluster tightly around the diagonal across the full T_m range. We do see evidence of ProxyTm assay bias and noise at lower T_m.