Probabilistic thermal stability prediction through sparsity promoting transformer representation

Yevgen Zainchkovskyy[*]
DTU Compute & Novo Nordisk A/S
yezs@novonordisk.com

Jesper Ferkinghoff-Borg
Novo Nordisk A/S
jfgb@novonordisk.com

Anja Bennett
Novo Nordisk A/S
zabx@novonordisk.com

Thomas Egebjerg
Novo Nordisk A/S
tegb@novonordisk.com

Nikolai Lorenzen
Novo Nordisk A/S
nlz@novonordisk.com

Per Jr. Greisen
Novo Nordisk A/S
pjug@novonordisk.com

Søren Hauberg
DTU Compute
sohau@dtu.dk

Carsten Stahlhut
Novo Nordisk A/S
ctqs@novonordisk.com

Abstract

Pre-trained protein language models have demonstrated significant applicability in different protein engineering task [1, 2]. A general usage of these pre-trained transformer models producing latent representation is to use a mean pool across residue positions to reduce the feature dimensions to further downstream tasks such as predicting bio-physical properties or other functional behaviours. In this paper we provide a two-fold contribution to machine learning (ML) driven drug design. Firstly, we demonstrate the power of sparsity by promoting penalization of pre-trained transformer models to secure more robust and accurate melting temperature (Tm) prediction of single-chain variable fragments with a mean absolute error of $0.23°$C. Secondly, we demonstrate the power of framing our prediction problem in a probabilistic framework. Specifically, we advocate for the need of adopting probabilistic frameworks especially in the context of ML driven drug design.

1 Introduction

Peptide and protein engineering is the process of optimizing peptide and proteins towards desired and valuable features for technological or medical applications [3]. In protein engineering, we seek to optimize the function of a protein with respect to e.g. its expression level, solubility, or thermal stability. Their functional behavior is directly determined by their amino acid sequence. Thus, to develop new or optimize desired properties for e.g., biomedical applications require to invert the relationship of the function given the sequence [4], also generally known in statistics and machine learning as the inverse problem. However, existing design methods have serious problems in distinguishing the functional levels of closely related proteins [5, 6]. While both protein engineering and design is a NP-hard problem [7], a direct search in the protein space simply becomes an overwhelming and intractable approach in linear time. Directed evolution has successfully demonstrated its applicability of mapping peptide and protein sequencing to functional behavior. However, it is highly limited by the fact that even high-throughput techniques only can sample a minor fraction of sequences constructed from diversification methods [8]. Among others, Bedbrook and co-workers

[*] Corresponding author

Workshop on Robustness in Sequence Modeling, 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

have demonstrated the direct applicability of utilizing machine learning for optimizing a property that would not have been possible to engineer through directed evolution alone, [6, 9, 10]. On the other hand, machine learning models are heavily dependent on learning from data - a crucial part in designing a machine learning driven drug design pipeline is therefore the accessibility of relevant functional data for the task at hand.

In this contribution, we demonstrate and discuss classical challenges in both designing compounds with dedicated properties from a minimal set of observations and data sets with quite scarce diversity. While we at one hand wish to provide as diverse molecules to maximize the coverage of the chemical search space and on the other hand seek to ensure optimized properties within minimum number of design rounds (experiments) - we are facing a typical active learning problem balancing explore vs exploit steps through the usage of model uncertainty estimates. In section 2, we provide a unified probabilistic framework for integrating compact latent pre-trained transformer features with Gaussian Process (GP) regression models, [11]. Through careful variant design train, development (dev), and test splits we demonstrate the applicability of uncertainty estimates to assess the models own notion of what it does not know. We examine the effect of training data with 1-5 mutations away from a wild-type sequence and the models ability to reason of its own predictive power to generalize to multiple mutations. While we in this paper limit ourselves to the quantification of predictive performance of the models, the GPs can be utilized as the surrogate model in a Bayesian Optimization framework for optimizing and searching the sequence space.

2 Background

Here, we motivate our problem and provide a brief overview of the core architecture of our probabilistic models utilizing a transformer architecture as input to our downstream regression models. Significant improvement and applicability of protein language models have been demonstrated over the last years, where among others the UniRep [12], Evolutionary Scale Modeling (ESM) [2], ProtBert [13] models can be mentioned. In [12], they utilize pretrained language model representation UniRep to generalize the representation to unseen regions of sequence space. Furthermore, Vig & Rao argues for attention in the transformer models corresponds to known biological properties like structure and binding sites that can enable contact prediction [14, 15]. In a drug design setting, we are especially interested in enabling pretrained representations in our protein engineering tasks for designing improved drugs as we are highly limited by the number of experiments we can conduct relative to the enormous sequence space at hand, e.g. 10^{130} for proteins of 100 amino-acids length. Thus, searching the space intelligently is needed even when utilizing high-throughput experimental setups.

While designing or engineering proteins, we are faced with the problem of optimizing towards specific functions of the molecules and effectively only interested in a tiny subspace of sequence space. The main challenge is naturally how can we utilize the general protein representation for a direct fine tuning to the downstream functional optimization task at hand. In this contribution, we seek to build a model for predicting the thermal stability of the antibody format single chain variable fragment (scFv). Due to the small sizes and the stranded nature of scFvs, these are commonly used as building blocks to construct recombinant multi-specific antibody formats, [16]. Unfortunately, the scFvs has been reported to be less termostable than larger antibody formats and thus more likely to lead to undesired aggregation and low Tms when utilized in a multi-specific format, [17]. To improve biophysical behavior of the scFV, our goal is to build a predictive model of the experimental measured Tm values determined by nano differential scanning fluorimetry (nanoDSF) [18]. Having an accurate model for prediction the melting temperature is needed to assess which variants to test experimentally for increased thermal stability. To quantify the accurate of our predictions, we follow a probabilistic approach where we not only obtain our mean predictions but just as importantly can provide uncertainty estimates on the predictions. Uncertainty estimates is critically needed for providing quantitative and directed search strategies balancing both exploitation and exploration.

2.1 Transformer-based models

Transformers are revolutionising NLP, have recently been repurposed to model biological sequences.In the core of a transformer architecture is the attention mechanism allowing to capture long-range dependencies between positions in a sequence. Originating as a solution to classic sequence-to-sequence (seq2seq) models, attention mechanism shows better performance and scaling characteristics than traditional RNNs or LSTMs. Common to those architectures is a context vector comprising of a hidden state of the network being carried through subsequent propagation, resulting in degraded performance with increased sequence lengths. On the other hand, transformers utilize self-attention, which allows processing of the whole sequence while still focusing on specific parts of it.

In the context of biological sequence modelling, the hidden state of a Transformer model corresponds to individual amino acid residues and represents the given amino acid in its context as a point in a high dimensional space (embedding). Thus, similar sequences are assigned similar representations by the network and are mapped to nearby points in space.

In this contribution, we use the ESM1-b variant of a Transformer protein language model from Facebook AI Research [2] encoding each of our sequences to an embedding $\mathbf{x} \in \mathbb{R}^{250\times1280}$.

2.2 Gaussian process regression

A Gaussian Processes (GP) is a powerful probabilistic framework enabling nonparametric, nonlinear Bayesian models [11]. A GP defines a prior distribution over the set of function $f(\mathbf{x})$ mapping the relation between our M-dimensional feature representation of our protein sequences to the target property $y = f(\mathbf{x}) + \epsilon$. Here ϵ represents additive observation noise. Using a standard zero-mean GP prior we obtain

$$p\left(f\left(\mathbf{X}\right)\right) = p\left(\mathbf{f}_X\right) = \mathcal{N}\left(0, \mathbf{K}\right), \tag{1}$$

where $\mathbf{K}$ is the covariance matrix between our training input features $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_N]$ such that $K_{ij} = k\left(\mathbf{x}_i, \mathbf{x}_j\right)$ defines the covariance function between input $\mathbf{x}_i$ and $\mathbf{x}_j$. We utilize one of the typically applied kernels for GP regression, Matern $\frac{5}{2}$ covariance function, with shared length-scale parameters for each input dimension σ_l, yielding

$$k\left(\mathbf{x}_i, \mathbf{x}_j\right) = \sigma_f^2 \left(1 + \frac{\sqrt{3}r}{\sigma_l}\right) \exp\left(-\frac{\sqrt{3}r}{\sigma_l}\right) \quad \text{where} \quad r = \sqrt{\left(\mathbf{x}_i - \mathbf{x}_j\right)^\top \left(\mathbf{x}_i - \mathbf{x}_j\right)} \tag{2}$$

Assuming additive independent identically distributed Gaussian noise with variance σ_ϵ^2 our predictive distribution for our new test proteins $\mathbf{Z}$ reads, [11],

$$\begin{aligned} p\left(f_Z | \mathbf{X}, \mathbf{y}, \mathbf{Z}\right) &= \mathcal{N}\left(\mu_z, \mathbf{\Sigma}_z\right), \quad \text{where} \quad \mu_z = k\left(\mathbf{Z}, \mathbf{X}\right)\left(\mathbf{K} + \sigma_\epsilon^2 \mathbf{I}\right)^{-1} \mathbf{y} \\ \mathbf{\Sigma}_z &= k\left(\mathbf{Z}, \mathbf{Z}\right) - k\left(\mathbf{Z}, \mathbf{X}\right)\left(\mathbf{K} + \sigma_\epsilon^2 \mathbf{I}\right)^{-1} k\left(\mathbf{X}, \mathbf{Z}\right). \end{aligned} \tag{3}$$

3 Driving sparsity through learned masks

Having extracted an embedding for each residue in a sequence of length P, a typical approach used to represent a complete protein as a single vector $\hat{\mathbf{x}}$ is by averaging across the transformer's hidden representation $\mathbf{x}$ at each sequence position p (mean pooling):

$$\hat{\mathbf{x}} = \frac{1}{P} \sum\nolimits_{p=1}^{P} \mathbf{x}_p \tag{4}$$

While significantly reducing the overall dimensionality of the embedding and allowing to represent proteins of different lengths, this approach inevitably results in loss of information. Intuitively, averaging assigns equal weight to all residues in the sequence, while in reality, only a handful of positions might influence the target of interest.

In this work, we investigate 3 different positional-weighted approaches to the typical averaging: a positively constrained **Learned mask** $\mathbf{w_l}$, sparsity promoting **Sigmoid-transformed mask** $\mathbf{w_s}$ and a Half-Cauchy **Prior based mask** $\mathbf{w_p}$:

$$\hat{\mathbf{x}}_\mathbf{l} = \frac{\sum_{p=1}^{P} \exp(\mathbf{w_l}_p)\mathbf{x}_p}{\sum \exp(\mathbf{w_l})} \quad (5) \qquad \hat{\mathbf{x}}_\mathbf{s} = \frac{\sum_{p=1}^{P} \mathrm{S}(\mathbf{w_s}_p)\mathbf{x}_p}{\sum \mathrm{S}(\mathbf{w_s})} \quad \text{where} \quad \mathrm{S}(x) = \frac{1}{1 + \exp(-x)} \quad (6)$$

$$\hat{\mathbf{x}}_\mathbf{p} = \frac{\sum_{p=1}^{P} \mathbf{w_p}_p \mathbf{x}_p}{\sum \mathbf{w_p}} \quad \text{where} \quad \mathbf{w_p} \sim \text{Half-Cauchy}(0, \sigma) \tag{7}$$

From a practical perspective, for all three approaches, we learn the masks as a part of the standard gradient based GP Maximum Log-Likelihood maximization procedure.

4 Results

Using Mean Absolute Error (MAE) as our metric, and partitioning the data set, we ran 3 sets of experiments corresponding to 3 different splits. First, we tested on the subset of the training set where single-site mutations were used as the validation (1MUT), next we used a random sample of the training set as validation (Uniform Shuffle) and finally evaluated the proposed methods on the hold-out test set itself. The reason for this partitioning is the spread of the positions of mutations in the wild-type sequence and relative sizes of training and validation set. Naturally, having learned a specific mask on the smaller training set, performance will degrade if mask does not reflect the mutated positions in the bigger test-set. This effect is simulated for the "Uniform Shuffle" split where the size of validation set was 24 samples (vs. 10 samples for the 1-MUT split).

Table 1: Evaluation of the Baseline and the proposed methods. For each split and method we report mean ± std over 64 trials. Bold values denote statistical significance against Baseline ($p < 0.05$).

	1-MUT Shuffle	Uniform Shuffle	Test Set (Hold out)
Baseline	1.216 ± 0.306	0.819 ± 0.157	0.273 ± 0.006
Learned mask	$\mathbf{1.110 \pm 0.338}$	0.836 ± 0.231	$\mathbf{0.227 \pm 0.010}$
Learned mask (Sigmoid)	$\mathbf{1.150 \pm 0.347}$	0.813 ± 0.161	$\mathbf{0.227 \pm 0.008}$
Learned mask (Prior)	$\mathbf{1.090 \pm 0.314}$	0.841 ± 0.158	$\mathbf{0.226 \pm 0.010}$

Gross metrics are reported in Table 1. Here, we see that the **Learned Mask (prior)** proposed method outperforms the standard averaging approach referenced as *Baseline* in the cases of balanced train/validation splits. Naturally, this method also depends on the actual value of the prior: σ - the second moment of Half-Cauchy distribution. In our case, a prior of $\sigma = 0.15$ was chosen as a result of a parameter sweep on the 1-MUT Shuffle resulting in lowest MAE. That prior is then reused for all **Learned Mask (prior)** runs.

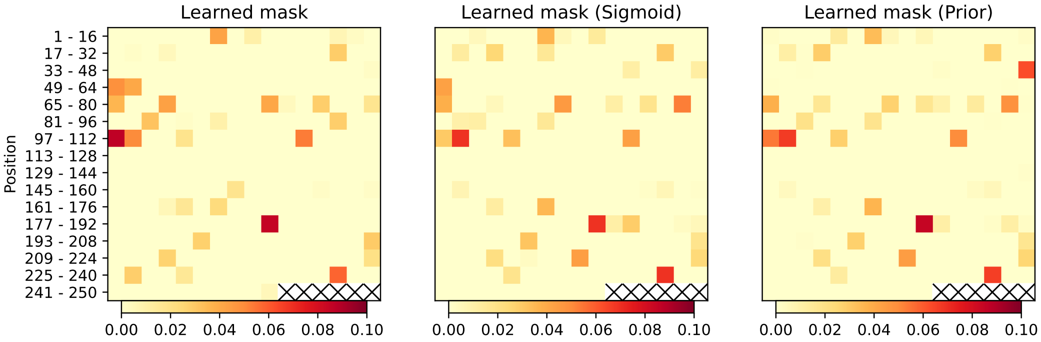

Figure 1: Masks learned on the hold out set. The number of zero entries (values below $1e^{-5}$ threshold) are: **208 for Learned mask**, **203 for Learned mask (Sigmoid)** and **9 for Learned mask (Prior)**. Unused entries are marked with a cross ×.

Learned masks are shown in Figure 1. In terms of the importance on melting temperature they emphasize roughly same positions in the sequence. Interestingly, the addition of Half-Cauchy prior, results in a more *dense* mask, as seen on the number of zero-entries. This appear to help generalization.

We summarize the individual test-set predictions and their corresponding uncertainties in Figure 2 (full version in Appendix, Figure 6).

Additionally, we train a model on a subset of the training-data comprising of only 80 samples all being single site mutations. While this restricted model does not perform particularly well in terms of our metric (resulting MAE $= 1.33$), it does great job in terms of uncertainties as shown in Figure 3 (full version in Appendix, Figure 7). Note that this limited model is consistently underestimating the target, as it is unable to capture additive effects of mutations.

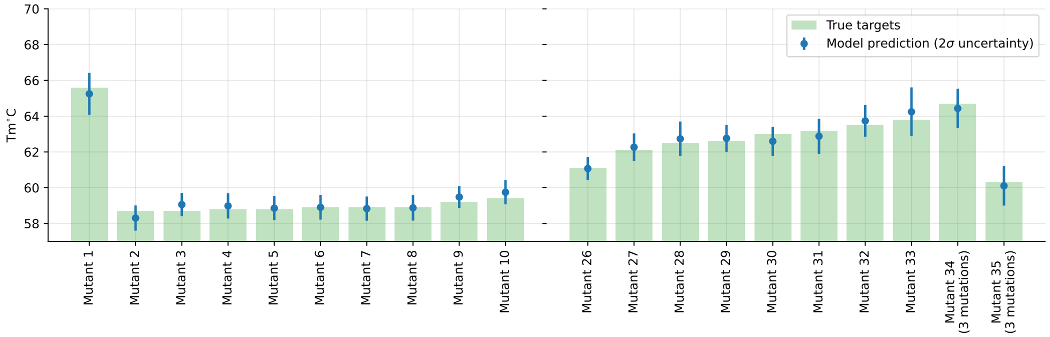

Figure 2: A subset of melting temperature prediction and corresponding uncertainties on the test-set.

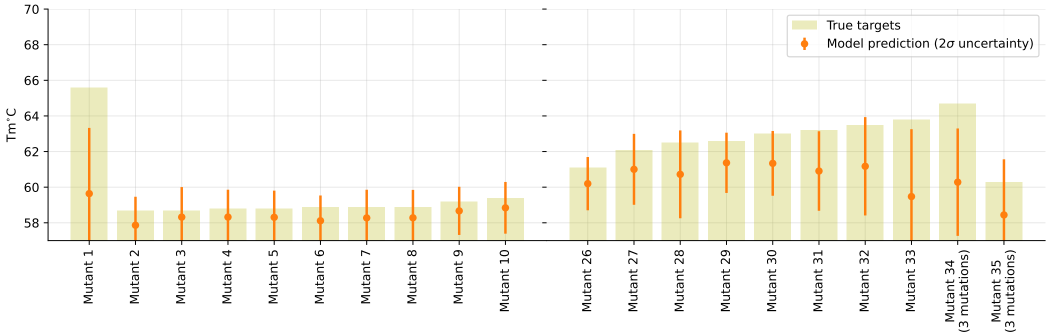

Figure 3: A subset of melting temperature prediction and corresponding uncertainties on the test-set when model is trained only on the sigle-mutations $N\text{train} = 80$

5 Conclusions

Utilization of machine learning driven techniques for navigating the drug design process towards optimized properties requires good compact representation of the molecule space of interest. We have demonstrated that sparsity promoting concentration of the larger pre-trained latent space provided by the protein language model, ESM-1b, leads to more robust estimate of a dedicated thermal stability optimization task for scFvs. We have proposed three variants of sparsity promoting effects through GP regression models integrating learned masks. In general all three models leads to improved predictive performance relative to a standard mean pooled feature representation. Even though our sparsity promoting models outperform our baseline model without mask on our final test data, our validation data indicates that the learned masks are sensitive to too aggressive sparsity when validation data is out-of-distribution of the training data. In fact, this makes sense as the sparse models exactly will seek to favor sparse representations given the training data at hand. Thus, if we seek to utilize the models solely for optimization in regions (residue positions) outside the support of previous seen data, care should be taken in utilizing the masks. From a Bayesian Optimization perspective, this type of evaluation would correspond to the explorative evaluation and thus the mean prediction evaluation is not suitable here. Instead of providing the mean prediction for evaluation we would seek opportunities to enrich the model support towards new regions e.g. through the upper-confidence-bound as acquisition function evaluation. Future work will examine the applicability sparsity promoting models in the context of steering explorative searches.

Acknowledgments and Disclosure of Funding

This project was supported in part by the Innovation Fund Denmark via the Industrial PhD Programme (grant no. 0153-00007B), Novo Nordisk R&D STAR Fellowship Programme and the Novo Nordisk Foundation (grant no. NNF20OC0062606) via the Center for Basic Machine Learning Research in Life Science (MLLS, `https://www.mlls.dk`).

References

[1] Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Peter Chen, John Canny, Pieter Abbeel, and Yun Song. Evaluating protein transfer learning with tape. *Advances in neural information processing systems*, 32, 2019.

[2] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *bioRxiv*, 2019.

[3] Jesse D Bloom, Sy T Labthavikul, Christopher R Otey, and Frances H Arnold. Protein stability promotes evolvability. *Proceedings of the National Academy of Sciences*, 103(15):5869–5874, 2006.

[4] Simona Cocco, Christoph Feinauer, Matteo Figliuzzi, Rémi Monasson, and Martin Weigt. Inverse statistical physics of protein sequences: a key issues review. *Reports on Progress in Physics*, 81(3):032601, 2018.

[5] Jiayi Dou, Lindsey Doyle, Per Jr Greisen, Alberto Schena, Hahnbeom Park, Kai Johnsson, Barry L Stoddard, and David Baker. Sampling and energy evaluation challenges in ligand binding protein design. *Protein Science*, 26(12):2426–2437, 2017.

[6] Kevin K Yang, Zachary Wu, and Frances H Arnold. Machine-learning-guided directed evolution for protein engineering. *Nature methods*, 16(8):687–694, 2019.

[7] Niles A Pierce and Erik Winfree. Protein design is np-hard. *Protein engineering*, 15(10):779–782, 2002.

[8] Nobuhiko Tokuriki and Dan S Tawfik. Stability effects of mutations and protein evolvability. *Current opinion in structural biology*, 19(5):596–604, 2009.

[9] Claire N Bedbrook, Kevin K Yang, J Elliott Robinson, Elisha D Mackey, Viviana Gradinaru, and Frances H Arnold. Machine learning-guided channelrhodopsin engineering enables minimally invasive optogenetics. *Nature methods*, 16(11):1176–1184, 2019.

[10] Florian Richter, Andrew Leaver-Fay, Sagar D. Khare, Sinisa Bjelic, and David Baker. De novo enzyme design using rosetta3. *PLOS ONE*, 6(5):1–12, 05 2011.

[11] Carl Edward Rasmussen. Gaussian processes in machine learning. In *Summer school on machine learning*, pages 63–71. Springer, 2003.

[12] Ethan C Alley, Grigory Khimulya, Surojit Biswas, Mohammed AlQuraishi, and George M Church. Unified rational protein engineering with sequence-based deep representation learning. *Nature methods*, 16(12):1315–1322, 2019.

[13] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, et al. Prottrans: towards cracking the language of lifes code through self-supervised deep learning and high performance computing. *IEEE transactions on pattern analysis and machine intelligence*, 2021.

[14] Jesse Vig, Ali Madani, Lav R Varshney, Caiming Xiong, Richard Socher, and Nazneen Fatema Rajani. Bertology meets biology: interpreting attention in protein language models. *arXiv preprint arXiv:2006.15222*, 2020.

[15] Roshan Rao, Joshua Meier, Tom Sercu, Sergey Ovchinnikov, and Alexander Rives. Transformer protein language models are unsupervised structure learners. *Biorxiv*, 2020.

[16] Robert E Bird, Karl D Hardman, James W Jacobson, Syd Johnson, Bennett M Kaufman, Shwu-Maan Lee, Timothy Lee, Sharon H Pope, Gary S Riordan, and Marc Whitlow. Single-chain antigen-binding proteins. *Science*, 242(4877):423–426, 1988.

[17] Anthony L Fink. Protein aggregation: folding aggregates, inclusion bodies and amyloid. *Folding and design*, 3(1):R9–R23, 1998.

[18] Wyatt Strutz. Exploring protein stability by nanodsf. *Biophysical Journal*, 110(3):393a, 2016.

A Appendix

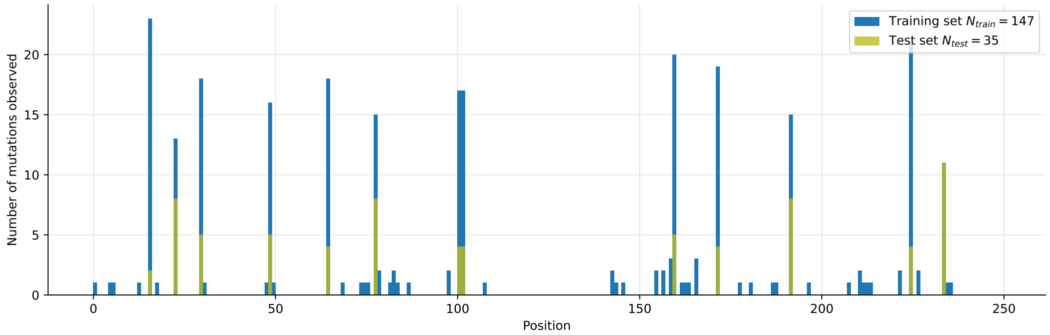

Figure 4: Histogram of the mutations occurring at the respective positions in training and test set.

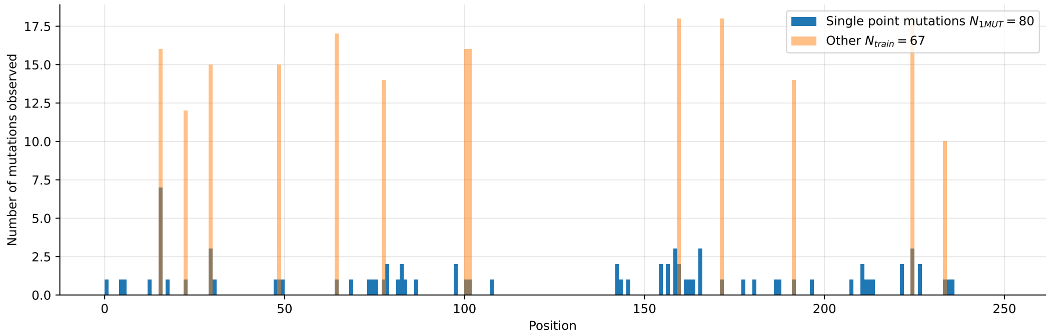

Figure 5: Histogram of the single site mutations of the training set.

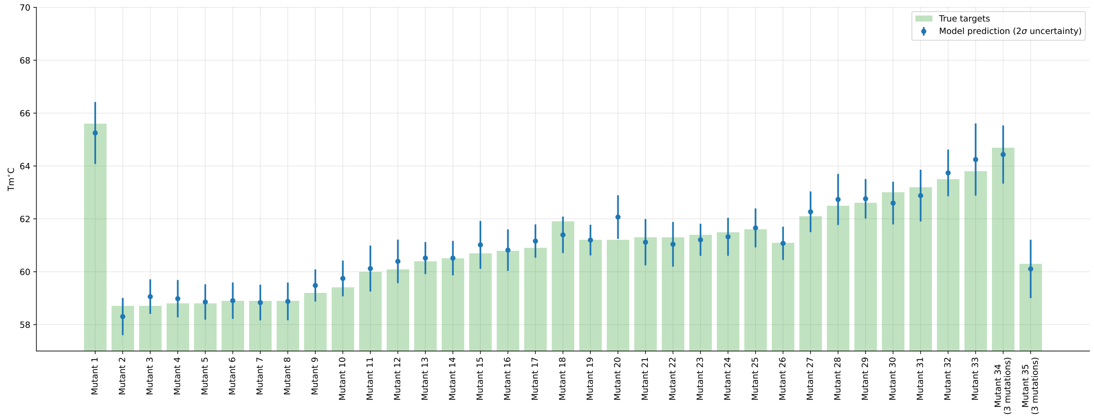

Figure 6: Complete test-set melting temperature prediction and corresponding uncertainties.

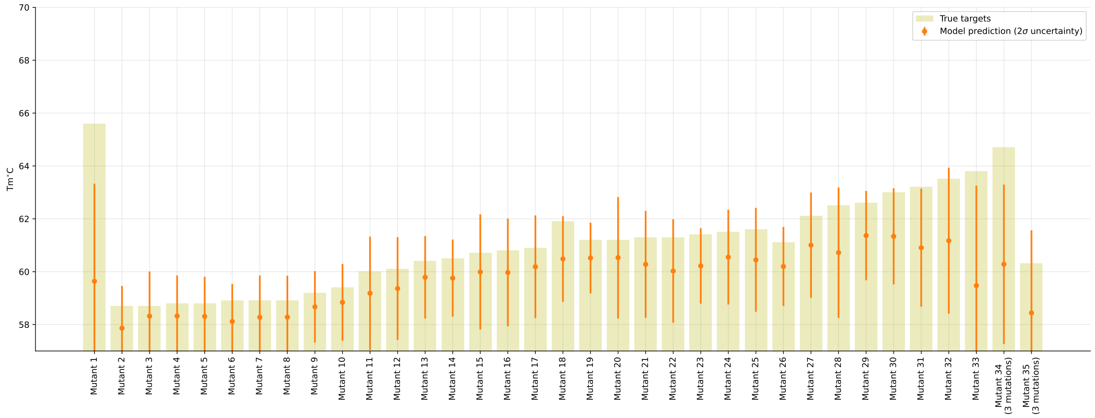

Figure 7: Complete test-set melting temperature prediction and corresponding uncertainties.