Preprint published online

TemStaPro: protein thermostability prediction using sequence representations from protein language models

Ieva Pudžiuvelytė [1,3], Kliment Olechnovič [1], Egle Godliauskaite [2], Kristupas Sermokas [2], Tomas Urbaitis [2], Giedrius Gasiunas [1,2], Darius Kazlauskas [1] *

[1]Institute of Biotechnology, Life Sciences Center, Vilnius University, Saulėtekis av. 7, Vilnius, LT-10257, Lithuania [2]CasZyme, Saulėtekis av. 7c, Vilnius, LT-10257, Lithuania [3]Institute of Computer Science, Faculty of Mathematics and Informatics, Vilnius University, Didlaukio st. 47, Vilnius, LT-08303, Lithuania

ABSTRACT

Reliable prediction of protein thermostability from its sequence is valuable for both academic and industrial research. This prediction problem can be tackled using machine learning and by taking advantage of the recent blossoming of deep learning methods for sequence analysis. We propose applying the principle of transfer learning to predict protein thermostability using embeddings generated by protein language models (pLMs) from an input protein sequence. We used large pLMs that were pre-trained on hundreds of millions of known sequences. The embeddings from such models allowed us to efficiently train and validate a high-performing prediction method using over 2 million sequences that we collected from organisms with annotated growth temperatures. Our method, TemStaPro (Temperatures of Stability for Proteins), was used to predict thermostability of CRISPR-Cas Class II effector proteins (C2EPs). Predictions indicated sharp differences among groups of C2EPs in terms of thermostability and were largely in tune with previously published and our newly obtained experimental data. TemStaPro software is freely available from https://github.com/ievapudz/TemStaPro.

INTRODUCTION

Biotechnological research and development often involves searching for proteins that can remain stable (maintain their spatial structures) in a high-temperature setting. In many cases, the only information initially known about a protein is its sequence of amino acids. Therefore it is beneficial to have computational tools that can efficiently predict protein thermostability from a protein sequence alone. Several machine learning-based methods were developed for that in the past (1, 2, 3, 4, 5, 6) and in recent years (7, 8, 9, 10, 11), but these efforts did not focus on drastically increasing the amount of data used for training and validation, which could potentially lead to better-performing methods with an ability to distinguish multiple levels of thermostability.

In general, most of the current state-of-the-art sequence-based methods (7, 8, 9, 10) were trained and tested using protein sequences taken from datasets of proteins annotated with experimentally-determined thermal stability information. Such datasets are inevitably small because gathering experimental data on a per-protein basis is usually very expensive and time-consuming. There is an alternative way of collecting protein thermostability data — taking the available information about optimal growth temperatures of organisms that have sequenced genomes converted to proteomes, grouping the proteomes by the corresponding growth temperature intervals, collecting the protein sequences, and annotating them with temperature values (11, 12). This approach is not as precise as gathering experimental data for every protein separately, but the optimal growth temperature of an organism provides a reliable lower bound for the melting temperature of proteins in that organism (13). Most importantly, the proteomes-based data gathering can provide millions of sequences for machine learning. Nevertheless, even the most recent deep learning-based method (11) was trained only on a small subset (less that 1%) of sequences from available proteomes with known growth temperatures.

Training using big data when starting from raw amino acid sequences is extremely challenging due to the need to construct or learn a complex protein representation suitable for making predictions. However, there is a possibility to take a shortcut and apply a transfer learning approach — use protein representations generated by other methods trained for different tasks. More specifically, it is possible to use protein sequence embeddings generated by encoders of protein language models (pLMs) that were trained on hundreds of millions of natural protein sequences (14, 15). Such pLM embeddings are rich representations that were already shown to be suitable inputs for various predictive tasks (16).

In this work we propose using pLM embeddings for training simple binary classifiers that predict whether a protein remains stable above some temperature threshold. We collected over 2 million of protein sequences from organisms with known optimal growth temperatures and we used that data to train, validate, and test multiple classifiers for multiple temperature thresholds. We also showed that our classifiers: perform exceedingly well on a previously published benchmark dataset; show similar performance for both water-soluble

*To whom correspondence should be addressed. Email: darius.kazlauskas@bti.vu.lt

and membrane proteins; do not suffer performance drops when applied to longer proteins. We combined the classifiers into a software tool that, given a protein sequence as input, predicts protein stability for multiple temperature thresholds and checks if the predictions are not contradicting each other. The resulting method, *TemStaPro* (Temperatures of Stability for Proteins), is freely available as a standalone program.

We tested TemStaPro software to predict thermostability of CRISPR-Cas Class II effector proteins (C2EPs). C2EPs are usually found in bacteria that grow best in moderate temperatures (20 to 45 °C). However, there are few Cas9 and Cas12b variants that can function at temperatures above 60 °C (17, 18). Thermostable C2EPs are important because they can be used in conjunction with nucleic acid amplification methods to detect SARS-CoV-2 variants of concern in a single reaction (18), for genome engineering of thermophilic organisms (19), or in aid to increase lifetime of gene editing tools in human plasma (20). Our results indicate that thermostability differs among groups of C2EPs, for example Cas12f and TnpB-like proteins are more likely to function at higher temperatures than ones from Cas9 and Cas13 groups.

MATERIALS AND METHODS

Data preparation

The data source that was used to construct training, validation, and testing datasets was composed of 21 498 annotated organisms (12, 21). The taxonomy identifiers given in the data source were used to fetch UniParc (22) identifiers for the corresponding proteome. UniParc identifiers were used to download FASTA files with proteins composing the proteomes.

The main objective was to develop a binary protein classification model into thermostability classes: proteins that are stable (class '1') or not stable (class '0') at 65 °C and higher temperatures. The threshold of 65 °C was chosen because primary envisioned purpose of our method was to detect C2EP proteins that can withstand >60 °C temperatures during isothermal amplification step in a one-pot SARS-CoV-2 detection kit (23). However, to get a more universally informative output of the tool, the class '0' was divided into subintervals from 40 to 65 °C using a step of 5 degrees.

The collection of datasets (24) that were used to train, validate, and test the tool is given in Table 1. The data source of proteomes was filtered from eukaryotes and duplicate taxonomy identifiers. *TemStaPro-Minor* is composed of 162 proteomes and *TemStaPro-Major* - of 5491 proteome. Both datasets were constructed in a way so that there would be no duplicate protein sequences and that a single proteome identifier from the collection of proteomes would be present only in either training, validation, or testing subset, which have approximate proportions of 70%, 15%, and 15%, respectively (regarding numbers of protein sequences). Since the datasets were intended to be used for a classification model that uses ESM-1b (14) embeddings as input, all subsets originally contained sequences no longer than 1022 amino acids.

Due to the random sampling of proteomes, *TemStaPro-Minor* set contains protein sequences mostly from organisms, whose growth temperatures are in ranges 20-40 °C or 60-80 °C (Supplementary Figure S1), thus the cases, which are not included in these intervals, are not sufficiently covered. Since this set was small, it was convenient for preliminary investigations of choosing the best input representations and architecture for the neural network model.

TemStaPro-Major is a set designated to train the final version of the classifier. To include more data, it resulted as an imbalanced set with sequences to cover the range of temperatures between 0 and 100 °C (Supplementary Figure S2).

It is important to note that the collected data is not suitable for training multiclass classification (where classes represent non-overlapping temperature intervals) or regression models because the temperature values used for ground truth are only lower bounds for the possible temperatures of stability of proteins. For example, a protein from an organism living in 45 °C environment may (or may not) also be stable at 60 °C. This also means that a single binary classifier is not very versatile. A binary classifier trained using 45 °C threshold cannot tell if a protein predicted as stable at over 45 °C is also stable at over 60 °C — a classifier trained using at least 60 °C threshold is needed for that. Thus, it was decided to train and use multiple binary classifiers in TemStaPro.

The final tool is composed of 6 classifiers trained to distinguish between 6 different temperature thresholds. For a more accurate evaluation of each model, there were 6 balanced versions of the testing subset of *TemStaPro-Major* created. Balancing was done by the random sampling of proteomes in the class '0' to collect the number of protein sequences composing the class '1' and keep as many different proteomes of the class '0' as possible.

In the process of development, we tested whether the tool makes predictions for longer protein sequences no less accurately than for shorter sequences. Therefore the original *Major* testing subset was supplemented with 47 831 sequences longer than 1022 amino acids.

Additional data for benchmarking

SAPPHIRE dataset One more dataset that was used to assess our method is SAPPHIRE (9) testing dataset, which is balanced and has 742 sequences. This dataset was used to evaluate SCMTPP (8) and ThermoPred (4) tools as well. The SAPPHIRE dataset contains thermophilic and

Table 1. All datasets that were used in the development of the tool.

Dataset	Subset	All sequences	Class 0 sequences	Class 1 sequences	Max. length
TemStaPro-Minor-bal65	Training	283 360	141 600	141 760	1 022
	Validation	63 158	32 790	30 368	1 022
	Testing	73 308	37 739	35 569	1 021
TemStaPro-Major-imbal	Training	1 835 664	1 688 495	147 169	1 022
	Validation	395 964	365 370	30 594	1 022
	Testing	435 182	408 518	26 664	27 313*
TemStaPro-Major-bal65	Testing	52 872	26 436	26 436	1 022
TemStaPro-Major-bal60	Testing	68 186	34 093	34 093	1 022
TemStaPro-Major-bal55	Testing	116 340	58 170	58 170	1 022
TemStaPro-Major-bal50	Testing	165 414	82 707	82 707	1 022
TemStaPro-Major-bal45	Testing	207 108	103 554	103 554	1 022
TemStaPro-Major-bal40	Testing	233 640	116 820	116 820	1 022
SAPPHIRE	Testing	742	371	371	1 643

*UniProt entry of the protein sequence that determined the maximum length of the subset: A0A222VTR7.

non-thermophilic proteins, although the exact temperature threshold for group distinction is not mentioned: according to the creators of the dataset, proteins labelled as thermophilic (TPP) are those that are stable at 80-100 °C temperature range (8). The dataset's thermophilic proteins were taken from thermophilic organisms - organisms that grow at the mentioned temperature range, meanwhile non-thermophilic proteins were collected from non-thermophilic organisms.

Sequence dataset of Class II effector proteins Initial datasets of Cas12 and Cas9 were taken from (25) and (17), respectively. Cas13 sequence dataset was constructed by building HMMER (26) sequence profiles for Cas13 groups (27) and using them to search NR (28), UniRef100 (29), MGnify (30), and IMG/VR v4 (31) databases with hmmsearch (26). Only sequences with E-value $\leq$1e-20 were extracted. In case the same sequence was found using different queries, the hit having lower E-value was assigned to the group. Latter dataset was combined with Cas12 and Cas9 datasets to form final dataset of 16376 sequences (*SupplementaryFileC2EPsPredictions.tsv*). Thermostability predictions were done for all those sequences, but to check the thermostability of different C2EP groups, we only used sequences having more than 300 residues.

Protein language models

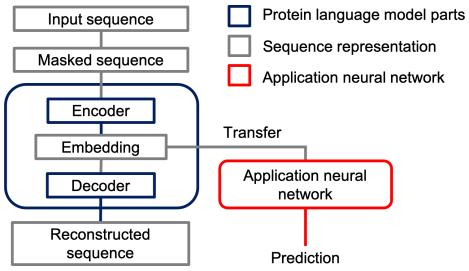

Figure 1. The scheme of embeddings from protein language model usage in the application neural network model.

ESM-1b (14) and ProtTrans (15) protein language models are transformer models trained on protein sequences. Due to the properties of transformer architecture, they are usually applied to natural language processing (NLP) tasks (32). However, since amino acid sequences can be considered as a particular language, transformer architectures were also applied to solve tasks related to protein biology. Attention mechanisms in models of transformer architecture, taking BERT-like model as an example (33), are capable to capture the folding structure, binding sites, and complex biophysical properties of proteins.

This work exploits the transfer learning by taking protein representations from the last layer of protein language models and passing them as input to the classification model (Figure 1).

The potential suitability of both ESM-1b and ProtTrans embeddings for thermostability classification was detected using principal component analysis (PCA) of mean embedding vectors. Plots of the first two principal components (Figure 2) demonstrated the distinct separation of points corresponding to different thermostability classes.

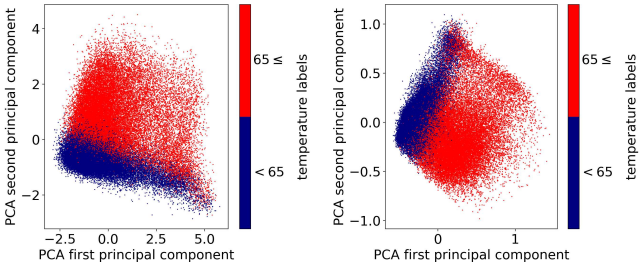

Figure 2. PCA visualizations of ESM-1b mean embeddings (left) and ProtTrans mean embeddings (right) computed for the validation subset of the *TemStaPro-Minor-bal65* dataset. Plots were generated using *Scikit-Learn Python* library (version 0.24.2).

A notable advantage of ProtTrans model is that it does not have a limit for protein's length, while ESM-1b does not work for protein sequences that are longer than 1022 amino acids.

Binary classifier design process

Classifiers in this study were implemented as feed-forward densely connected neural network (NN) models with up to two hidden layers. The classifier design process involved training, validation, and testing of multiple NN architectures using multiple types of input representation derived from pLM embeddings. In order to try more variations in a reasonable amount of time, the initial design stage was done using *TemStaPro-Minor-bal65* dataset.

The comparison of ESM-1b and ProtTrans embeddings was done extensively: not only mean, yet also other kinds of representations (Supplementary Table S1) were tested as input to the single-layer perceptron (SLP) classifier. The representation performance analysis, done using common metrics for binary classification (MCC, accuracy, precision, recall, and ROC AUC), demonstrated that ProtTrans representations, normalization of embeddings, and including more information about the distribution of embeddings' components in the representation (for instance, octiles embeddings) give the best results (Supplementary Table S2, Supplementary Figures S3 and S4).

In addition to the representation analysis, the search for the best model's architecture was executed as well. Architectures that were chosen to run experiments with had one or two hidden layers, whose sizes were chosen to be original embeddings size divided by several multiples of 2 (Supplementary Figure S5, Supplementary Table S3). The results of the architecture analysis showed that the bigger the predictor's architecture is taken, the more accurate predictions are done (Supplementary Figure S6). Such conclusion was made in the scope of the defined set of models.

The results of the analysis performed using *TemStaPro-Minor-bal65* dataset allowed to make the following decision: for further development of the final method ProtTrans mean and octiles embeddings were chosen. Although models that used mean ProtTrans embeddings did not provide as good results as the octiles representations, due to less resource-intensive training procedure with mean embeddings, both these representations were chosen to be used for the further training of binary classifiers on *TemStaPro-Major-imbal* dataset using multiple thresholds. The trained classifiers

were then tested on *TemStaPro-Major-bal* datasets, and the results (Supplementary Figure S7) showed no significant performance differences between using means and octiles for input. Since mean embeddings take up less storage space, they were chosen to be used in the final predictor tool.

Final binary classifiers training and testing

The functionality of the TemStaPro method is carried out by multiple binary classifiers that accept mean ProtT5-XL (from ProtTrans) representations (vectors of length 1024) as input. Classifiers were implemented as neural network models using *PyTorch*. Each model is a multi-layer perceptron with 2 fully-connected hidden layers of sizes 512 and 256. After each layer (except the last one), rectified linear unit (ReLU) activation function is applied. Sigmoid activation function is used after the last layer.

A single binary classifier was trained using mini-batch training principle (with batch size of 24) and Adam optimizer (34) with learning rate of 0.0001. Since the datasets were imbalanced, training and validation sets were loaded using weighted random sampler and the loss function was chosen to be weighted cross entropy. The loss function used weights that were calculated based on the training subset.

The validation of each binary classifier was done after each training epoch for the whole validation set. For each epoch model evaluation metrics were retrieved: MCC, accuracy, loss, precision, recall, ROC AUC, and precision-recall (PR) AUC.

Testing of each model was done for the whole chosen (imbalanced or balanced corresponding to the tested model) testing set. As in the validation stage, model evaluation metrics for the testing stage were calculated as well.

Predictor application

The TemStaPro user's input is a FASTA file with amino acid sequences with proteins' identifiers in the headers. For each protein in the FASTA file a mean ProtTrans (15) embedding is generated, which is the input of the classification model. In addition to this, there is an option available to pass embeddings of each residue to the model. Per-residue embeddings can also be averaged over a residue window of size k, which can be customized, to get per-segment embeddings. Then each segment (of size k) of amino acids gets a prediction.

The classification predictions are made by 6 ensembles each composed of 5 neural network models that were trained to make binary classification of proteins with respect to one of 6 temperature thresholds, which were chosen to be: 40, 45, 50, 55, 60, and 65. The output list of 6 predictions is created from averaged predictions of each ensemble.

Based on the sequence of all 6 classification predictions, each input protein is assigned two labels: left-hand and right-hand. These labels are determined by scanning the binary predictions starting from the left or right-hand side, respectively. For example, the left-hand label is assigned the temperature range, where the last positive prediction (class '1') is encountered. If outputs are only negative (class '0'), then the label is the first temperature range. On the other hand, the right-hand label is assigned by reading the outputs starting from the right: the label is assigned the temperature range, where the first '1' is encountered. The treatment of the '0'-only case coincides with the left-hand principle.

Since binary predictions are made independently, conflicts might occur between the outputs of the classifiers: for instance, the predictor of 40 °C threshold would predict that protein is not stable at 40 °C and higher temperatures, although the predictor of 50 °C threshold would state otherwise. When such conflict occurs, left-hand and right-hand labels differ. On the contrary, if the labels report the same temperature interval, that interval can be interpreted as the highest temperature range at which the protein was predicted to still be thermostable.

RESULTS

Performance of binary classifiers

Trained classifiers were tested with imbalanced and balanced (for each temperature threshold) testing sets (Table 2, Supplementary Tables S4 and S5).

In the imbalanced testing case, the maximum MCC score that was reached by the binary classifier for 50 °C temperature threshold was 0.691.

In the balanced testing for each temperature threshold case, the highest MCC score was achieved by the classifier adjusted for 65 °C threshold.

Misclassification analysis in terms of protein solubility

Since there is a script presented in ProtTrans GitHub repository (15) to predict, whether a protein is membrane-bound or water-soluble, it was attempted to use it to label proteins in *TemStaPro-Major* dataset in terms of solubility. We checked whether the thermostability predictors are biased towards one solubility class of proteins.

Binary classifiers were used to make thermostability predictions for the balanced (for 65 °C threshold) subset from *TemStaPro-Major* dataset, which consisted of 52872 proteins, of which 11961 were membrane-bound and 40911 were water soluble. After counting the misclassification cases and calculating the frequency of mistakes (Table 3), we observed that predictors are not biased to any group: on average 9 percent of proteins were misclassified in each case.

Sequence length effect analysis

Since the classifiers were trained using the dataset with sequences no longer than 1022 amino acids, we decided to assess the predictors with a testing set that includes longer

Table 2. TemStaPro models' MCC and precision-recall (PR) AUC scores after testing with imbalanced and balanced (for each temperature threshold) *TemStaPro-Major* datasets.

	TemStaPro-Major-imbal		*TemStaPro-Major-bal*	
Model	MCC	PR AUC	MCC	PR AUC
TemStaPro-t65	0.647	0.781	0.838	0.971
TemStaPro-t60	0.570	0.786	0.801	0.963
TemStaPro-t55	0.613	0.740	0.725	0.927
TemStaPro-t50	0.691	0.826	0.756	0.939
TemStaPro-t45	0.677	0.840	0.703	0.924
TemStaPro-t40	0.635	0.810	0.640	0.896

sequences to see if the classifier performs on longer sequences no worse than on the shorter ones.

The original *TemStaPro-Major* testing set was supplemented with sequences longer than 1022 amino acids. The predictions for this set were made using a standalone classifier for the temperature threshold of 65 °C trained on *TemStaPro-Major-imbal* training dataset and were analysed with respect to the length threshold, so that at least 5% of the sequences would fall into each bin - the last such bin consisted of sequences that were longer than 1200 amino acids. Additionally, each of the bins were balanced with respect to the temperature labels of 65 °C threshold. The plot (Figure 3) demonstrates that the alteration of the predictor's performance using longer sequences for testing was not substantial.

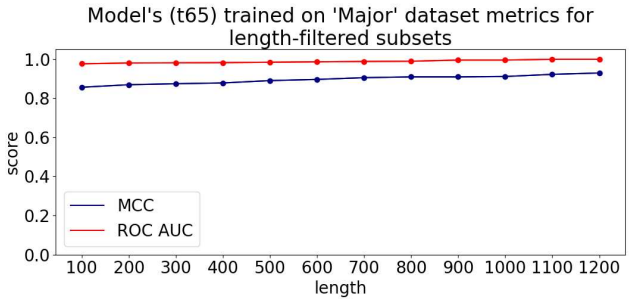

Figure 3. MCC and ROC AUC evaluation scores of the classifier trained on *TemStaPro-Major-bal65* at various length thresholds.

Performance comparison with other predictors

TemStaPro was also tested on SAPPHIRE tool's testing dataset (9), which was previously used to test SCMTPP (8) and ThermoPred (4) tools for protein thermostability prediction.

ThermoPred is a tool that uses support vector machine (SVM) model and amino acid along with dipeptide composition as input. This method was trained to identify thermophilic proteins by choosing the lower and upper bounds for optimal growth temperature of organism for which the protein belongs: 60 °C was a lower limit for thermophilic organisms and 30 °C - an upper limit for mesophilic organisms.

SCMTPP and SAPPHIRE are methods that are trained to identify thermophilic proteins that are considered to be stable at 80-100 °C temperature range. SCMTPP uses scoring card method (SCM) and dipeptide composition as input, whereas SAPPHIRE is a method that uses partial least squares (PLS) regression and 12-dimensional vectors as input. These vectors are constructed out of outputs of 5 different types of models combined with 8 types of composition, physicochemical property, composition-transition-distribution (CTD), and evolutionary information-based features.

The tool with the highest MCC score among other published predictors is SAPPHIRE. The results showed that TemStaPro predictors for temperature thresholds 55-65 °C perform better than SAPPHIRE (Table 4). The MCC scores of remaining predictors (for thresholds 40-50 °C) differed from SAPPHIRE tool's score by no more than 0.05, which makes the TemStaPro a prospective tool for accurate thermostability predictions.

Software tool

We implemented TemStaPro as a command-line software tool, it is freely available at https://github.com/ievapudz/TemStaPro. By default, TemStaPro provides global thermostability scoring for each input protein sequence. An example output table of the global scoring is given in Figure 4: each protein gets 6 raw thermostability scores from the predictor of every temperature threshold together with left-hand and right-hand labels.

Conflicting cases, when left-hand and right-hand labels differ, are marked with '*' in the 'clash' column. *TemStaPro-Major* testing set (without longer sequences included) and a standalone classifier for 65 °C threshold was used to check how frequently such conflicts occur: 18945 out of 387351 proteins had conflicting predictions, which makes up to 5% of all cases.

Besides the default global protein scoring, the user might opt for per-residue or per-segment predictions. For the per-residue case each amino acid in the protein sequence gets a distinct set of thermostability predictions (Supplementary Figure S8), similarly for per-segment option, where each full segment of the chosen size in the sequence gets its set of predictions. Additionally, there is an option to plot per-residue and per-segment predictions (Figure 5).

The speed of the TemStaPro software is mostly determined by whether the ProtTrans embeddings are produced on GPU or not. Using NVIDIA GeForce RTX 2080 Ti GPU, TemStaPro processes 10000 sequences with average length of 1000 residues in less than 2 hours. Without GPU, the operating time increases several times (up to 60 times if run on a laptop with Intel i7-8565U CPU).

Table 3. Frequency of protein misclassification cases.

Model	Frequency of mistakes among: membrane-bound proteins	soluble proteins
Model trained on *TemStaPro-Major-bal65*	0.098 (1171/11961)	0.087 (4730/40911)
Model trained on *TemStaPro-Major-imbal65*	0.092 (1095/11961)	0.084 (4518/40911)

Table 4. Models' scores after testing with an independent SAPPHIRE dataset.

Model	Accuracy	Sensitivity (Recall)	Specificity	MCC	ROC AUC
TemStaPro-t65	0.958	0.938	0.978	0.917	0.990
TemStaPro-t60	0.961	0.984	0.938	0.923	0.996
TemStaPro-t55	0.949	0.989	0.908	0.901	0.990
TemStaPro-t50	0.923	0.992	0.854	0.854	0.976
TemStaPro-t45	0.914	0.989	0.838	0.837	0.975
TemStaPro-t40	0.914	0.997	0.830	0.839	0.984
SAPPHIRE	0.942	0.951	0.933	0.884	0.980
SCMTPP	0.865	0.849	0.881	0.731	-
ThermoPred	0.860	0.938	0.782	0.729	-

protein_id	position	sequence	length	t40_binary	t40_raw	t45_binary	t45_raw	t50_binary	t50_raw	t55_binary	t55_raw	t60_binary	t60_raw	t65_binary	t65_raw	left_hand_label	right_hand_label	clash
YmeCas12a	-	MSKVNNG...FVLRNLS	1362	1	8.606E-01	0	4.829E-01	0	1.332E-03	0	4.124E-10	0	5.582E-07	0	7.915E-11	[40-45)	[40-45)	-
SauCas9	-	MKRNYIL...PQIIKKG	1053	0	3.019E-01	0	4.704E-01	1	6.093E-01	0	7.953E-02	0	1.469E-03	0	1.179E-09	<40	[50-55)	*
SpyCas9	-	MDKKYSI...LSQLGGD	1368	0	8.729E-05	0	5.426E-08	0	1.268E-08	0	6.861E-14	0	6.653E-08	0	2.441E-14	<40	<40	-
CaldoCas9	-	MRYKIGL...PLQSTRD	1087	1	6.183E-01	1	8.813E-01	1	9.999E-01	1	7.876E-01	0	2.369E-01	0	2.43E-08	[55-60)	[55-60)	-

Figure 4. An example tab-separated table that is the output of the global prediction mode of TemStaPro program. The main output of the method is a TSV table with 8 columns: 'protein_id' - a header taken from the FASTA file of the input protein; 'sequence' - an amino acid sequence of the protein; 'length' - a length of the protein's amino acid sequence; 't??_binary' - a binary prediction label for a given temperature threshold (one of the six thresholds is written in the place of question marks) - the label is assigned by rounding the raw prediction (see the next point) at this temperature threshold; 't??_raw' - a raw prediction value for a given temperature threshold (real numbers from the interval [0, 1]); 'left_hand_label' - a label of the highest temperature range, at which the protein was predicted to still be thermostable (possible labels of temperature ranges are: '<40', '[40-45)', '[45-50)', '[50-55)', '[55-60), '[60-65)', '≥65'); 'right_hand_label' - a label that is interpreted as 'left_hand_label', yet the label is assigned by reading the outputs starting from the right (possible values of the label coincide with the 'left_hand_label'); 'clash' - a Boolean identifier, whether a contradiction between the models' predictions was observed - the expected output is a decreasing sequence of binary predictions if the outputs are read from left to right in the increasing order of the temperature thresholds (expected output is labelled as '-' and other cases are assigned '*').

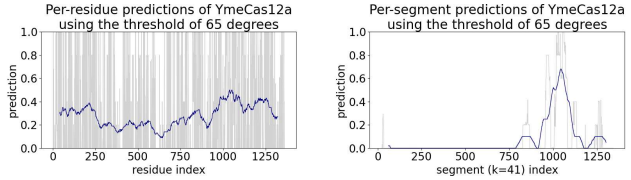

Figure 5. An example plot for the output of per-residue mode (left) and per-segment mode with default window size of 41 (right).

Thermostability of Class II effector proteins

To get a better view on thermal stability among different C2EP groups, we tested our method on a large dataset (16376 sequences) of Cas9, Cas12, TnpB, and Cas13 proteins (*SupplementaryFileC2EPsPredictions.tsv*). For further analysis, we considered only sequences longer than 300 residues. There were 11341 such sequences, 10324 (91%) of them had clash-free predictions (*SupplementaryC2EPClashAnalysis.xlsx*). Thermostability prediction varied greatly between groups of Cas12 and TnpB (Figure 6). This might be explained by the fact that members of Cas12 and TnpB differ greatly in sequence similarity and length (35, 36). The Cas12a group, which is currently actively studied and used in biotechnology applications (37), did not have thermostable (≥65 °C) sequences (Figure 6) as predicted by our method.

In contrast to Cas12a, more than half of the members of Cas12b group were predicted to function at 50 °C or higher temperature (Figure 6). Such observation corresponds to the experimental data because most of the characterized thermostable Cas12 proteins belong to the Cas12b group (23, 38). Interestingly, we predicted that most thermostable Cas12 groups are Cas12f1, Cas12f2, and Cas12g (Figure 6). However, the latter groups were not studied experimentally for thermostability. On the other hand, some of the members (e.g. Un1, Un2, Mi1, and Mi2) of Cas12f1 and Cas12f2 groups are found in archaea, which is an indication of possible thermostability of these C2EPs. Two groups of TnpBs (namely, TnpB2 and TnpB-Kra, which contains TnpB from *Ktedonobacter racemifer* (39)) showed higher predicted thermal stability compared to a group represented by TnpB from *Deinococcus radiodurans* ISDra2.

Just a few Cas9 groups (namely, Cas9-C3 and Cas9-C7; Supplementary Table S6) contain thermostable members. This observation is in tune with experimental data. The Cas9-C3 group contains characterized thermostable proteins CaldoCas9, GeoCas9, and ThermoCas9 (19, 20, 40). Cas9-C7 group includes NsaCas9 which was shown to function at temperatures above 60 °C in our previous study (17).

Cas13 groups tend to have predicted lower thermostability except for Cas13x (Figure 6), which contains only 8 members, thus it is too early to draw any conclusions about their thermal stability.

In rare cases our method predicted lower (differences >10 °C) than experimentally characterized temperatures for thermostable proteins (e.g. TccCas13a; Supplementary Table S6). However, these are exceptions, in 92% of the cases (35 out of 38) predicted thermostability varied no more than 10 °C from the experimental data (Supplementary Table S6).

Experimental validation of thermostability predictions

To validate the accuracy of the thermostability prediction model, we have experimentally characterized two potentially thermostable proteins. Ghy2Cas9 was previously identified and described in (17). This enzyme showed dsDNA cleavage activity in cell free lysates, but protein thermostability was not characterized. We also identified a putative thermostable Cas12b ortholog from *Clostridia bacterium*, CbaCas12b (NMA13999.1), in publicly available genome databases. We expressed the enzymes in *E. coli* and purified them. The proteins along with SpyCas9, as a control, were subjected to analysis by nano differential scanning fluorimetry (nanoDSF) to ascertain the temperatures at which they unfold. The enzymes were tested either without their guide RNA (apo) or with single guide RNA (RNP), except for CbaCas12b, for which we could not identify a tracrRNA. Ghy2Cas9 was shown to begin to unfold at around 54 °C, CbaCas12b at 57 °C, and SpyCas9 at 45 °C, with their respective RNPs unfolding at around 2-3 °C higher temperature (Figure 7a, Supplementary Figure S9). Predicted temperatures of thermal stability for both Ghy2Cas9 and SpyCas9 ([60-65) and <40 °C, respectively) did not differ more than 7 °C from their experimentally determined melting point temperatures. Following this, we evaluated the dsDNA cleavage activity of Ghy2Cas9 as well as SpyCas9 across a range of temperatures from 37 °C to 70 °C using fluorophore-labelled dsDNA substrates. As shown in Figure 7b, Ghy2Cas9 and SpyCas9 retained robust nuclease activity at temperatures up to 55 °C

and 50 °C, respectively, which correlates with the determined unfolding temperature of the RNPs by nanoDSF.

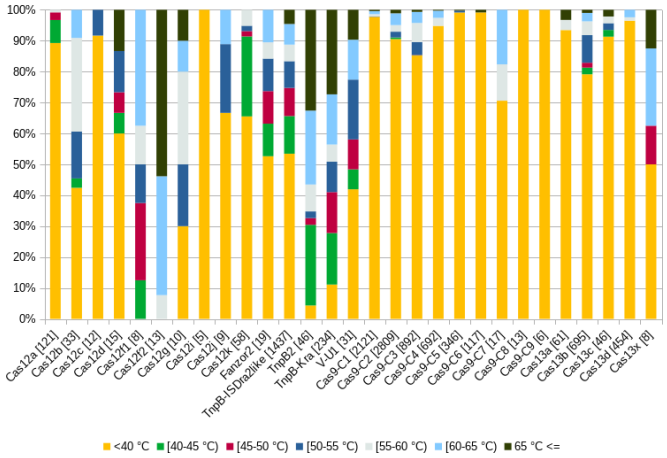

Figure 6. Predicted thermostability of various C2EP groups. Numbers in brackets correspond to the amount of sequences used for predictions.

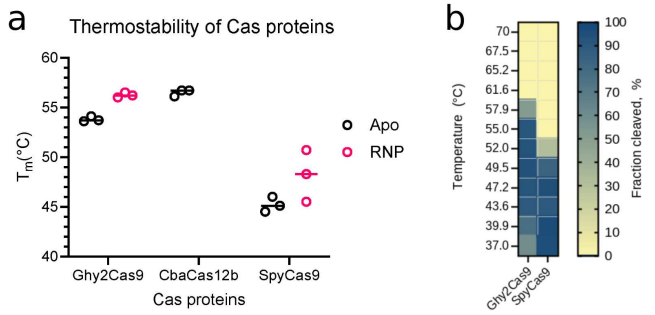

Figure 7. **(a)** Thermal stability of Cas proteins without guide RNA (apo) or loaded with sgRNA (Ghy2Cas9 and SpyCas9) (RNP). Protein unfolding was measured using nano differential scanning fluorimetry (nanoDSF) over a temperature range from 20 °C to 80 °C. Fluorescence was monitored as temperature increased at a rate of 1 °C per second. The inflection point of the fluorescent curve is interpreted as the unfolding point of the protein (T_m). Data points collected from replicate experiments are plotted as circles, the means are plotted as dashes. **(b)** The double-stranded DNA (dsDNA) cleavage activities of Ghy2Cas9 and SpyCas9 RNPs were measured using in vitro assays containing fluorophore-labeled dsDNA target substrates. Cleaved fragments were quantitated and are represented in a heatmap showing overall activity at temperatures ranging from 37 °C to 70 °C. The intensity of the blue colour indicates the fraction of substrate cleaved.

DISCUSSION AND CONCLUSIONS

Embeddings from pre-trained protein language models can be highly suitable for the task of protein thermostability prediction — this became nearly apparent even after our initial principal component analysis of ESM and ProtTrans embeddings. We further showed that a simple dense neural network can be efficiently trained to predict a protein thermal stability class from the mean of per-residue embedding vectors (we also showed that quantile values can be used as an alternative to mean values). With that established, we endeavored to make a better thermostability prediction method not by complicating the machine learning model, but rather by preparing and using more data for training and validation. We prepared and utilized a dataset of over 2 million sequences annotated with temperatures. The considerable amount and diversity of our data allowed us to train, validate, and test classifiers for multiple temperature thresholds (from 40 to 65 °C), and to establish that the performance of our predictors was not affected by the sequence length or protein solubility. When tested on a recent independent dataset, SAPPHIRE (9), our trained and validated method, named TemStaPro, performed better than state-of-the-art sequence-based predictors.

As the final TemStaPro version uses mean pLM embedding vectors as input, we added a possibility to consider not only the full-sequence mean, but also means on the level of subsequences and even single residues — the resulting local thermostability scoring can be used to visualize how different parts of the protein sequence contribute to the global classification outcome.

We tested our method on CRISPR-Cas Class II effector proteins. Interestingly, we saw large variation in thermal stability among groups of Cas12 and TnpB. For example, more than a half of the members of groups Cas12b, Cas12f1, Cas12f2, Cas12g, TnpB2 and TnpB-Kra have predicted temperatures of $\geq$50 °C (Figure 6). In contrast, members of Cas12a, Cas9, and Cas13 groups might function at lower temperatures.

We also observed that TemStaPro is a more pessimistic than optimistic predictor — it tends to slightly underestimate the highest temperature at which the protein is still stable. We attribute this trait to the particularity of the training data, where sequences were annotated not with exact melting temperature values, but with their lower bounds.

To conclude, considering that the large majority (92%) of our predictions for well-characterized proteins were confirmed experimentally, we believe that TemStaPro can be useful for pre-screening potentially thermostable candidate proteins and thus reducing the number of experiments needed to determine protein thermostability in biotechnology.

ACKNOWLEDGEMENTS

We thank Antanas Kiziela, Mindaugas Margelevičius, and Česlovas Venclovas for valuable comments about the manuscript and the software.

FUNDING

This project has received funding from European Regional Development Fund (project No 13.1.1-LMT-K-718-05-0021) under grant agreement with the Research Council of Lithuania (LMTLT). Funded as European Union's measure in response to COVID-19 pandemic.

Conflict of interest statement. EG, KS, TU, and GG are employees of CasZyme, GG has a financial interest in CasZyme. The remaining authors declare that they have no conflict of interest.

REFERENCES

1. Gromiha, M. M. and Suresh, M. X. (2008) Discrimination of mesophilic and thermophilic proteins using machine learning algorithms. *Proteins,* **70**(4), 1274–1279 [PubMed: 17876820] [doi: 10.1002/prot.21616].
2. Ku, T., Lu, P., Chan, C., Wang, T., Lai, S., Lyu, P., and Hsiao, N. (2009) Predicting melting temperature directly from protein sequences.

Comput Biol Chem, **33**(6), 445–450 [PubMed: 19896904] [doi: 10.1016/j.compbiolchem.2009.10.002].
3. Wu, L.-C., Lee, J.-X., Huang, H.-D., Liu, B.-J., and Horng, J.-T. (2009) An expert system to predict protein thermostability using decision tree. *Expert Systems with Applications,* **36**(5), 9007–9014 [doi: 10.1016/j.eswa.2008.12.020].
4. Lin, H. and Chen, W. (2011) Prediction of thermophilic proteins using feature selection technique. *J Microbiol Methods,* **84**(1), 67–70 [PubMed: 21044646] [doi: 10.1016/j.mimet.2010.10.013].
5. Nakariyakul, S., Liu, Z.-P., and Chen, L. (2012) Detecting thermophilic proteins through selecting amino acid and dipeptide composition features. *Amino Acids,* **42**(5), 1947–1953 [PubMed: 21547362] [doi: 10.1007/s00726-011-0923-1].
6. Fan, G.-L., Liu, Y.-L., and Wang, H. (2016) Identification of thermophilic proteins by incorporating evolutionary and acid dissociation information into Chou's general pseudo amino acid composition. *J Theor Biol,* **407**, 138–142 [PubMed: 27396359] [doi: 10.1016/j.jtbi.2016.07.010].
7. Feng, C., Ma, Z., Yang, D., Li, X., Zhang, J., and Li, Y. (2020) A Method for Prediction of Thermophilic Protein Based on Reduced Amino Acids and Mixed Features. *Front Bioeng Biotechnol,* **8**, 285 [PubMed: 32432088] [PubMed Central: PMC7214540] [doi: 10.3389/fbioe.2020.00285].
8. Charoenkwan, P., Chotpatiwetchkul, W., Lee, V. S., Nantasenamat, C., and Shoombuatong, W. (2021) A novel sequence-based predictor for identifying and characterizing thermophilic proteins using estimated propensity scores of dipeptides. *Sci. Rep.,* **11**(1), 23782 [PubMed: 34893688] [PubMed Central: PMC8664844] [doi: 10.1038/s41598-021-03293-w].
9. Charoenkwan, P., Schaduangrat, N., Moni, M. A., Lio', P., Manavalan, B., and Shoombuatong, W. (2022) SAPPHIRE: A stacking-based ensemble learning framework for accurate prediction of thermophilic proteins. *Comput Biol Med,* **146**, 105704 [PubMed: 35690478] [doi: 10.1016/j.compbiomed.2022.105704].
10. Ahmed, Z., Zulfiqar, H., Khan, A. A., Gul, I., Dao, F.-Y., Zhang, Z.-Y., Yu, X.-L., and Tang, L. (2022) iThermo: A Sequence-Based Model for Identifying Thermophilic Proteins Using a Multi-Feature Fusion Strategy. *Front Microbiol,* **13**, 790063 [PubMed: 35273581] [PubMed Central: PMC8902591] [doi: 10.3389/fmicb.2022.790063].
11. Zhao, J., Yan, W., and Yang, Y. (2023) DeepTP: A Deep Learning Model for Thermophilic Protein Prediction. *Int J Mol Sci,* **24**(3), 2217 [PubMed: 36768540] [PubMed Central: PMC9917291] [doi: 10.3390/ijms24032217].
12. Engqvist, M. K. M. (2018) Correlating enzyme annotations with a large set of microbial growth temperatures reveals metabolic adaptations to growth at diverse temperatures. *BMC Microbiol,* **18**(1), 177 [PubMed: 30400856] [PubMed Central: PMC6219164] [doi: 10.1186/s12866-018-1320-7].
13. Dehouck, Y., Folch, B., and Rooman, M. (2008) Revisiting the correlation between proteins' thermoresistance and organisms' thermophilicity. *Protein Eng Des Sel,* **21**(4), 275–278 [PubMed: 18245807] [doi: 10.1093/protein/gzn001].
14. Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., and Fergus, R. (2021) Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci U S A,* **118**(15), e2016239118 [PubMed: 33876751] [PubMed Central: PMC8053943] [doi: 10.1073/pnas.2016239118].
15. Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., Bhowmik, D., and Rost, B. (2022) ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *IEEE Trans Pattern Anal Mach Intell,* **44**(10), 7112–7127 [PubMed: 34232869] [doi: 10.1109/TPAMI.2021.3095381].
16. Fenoy, E., Edera, A. A., and Stegmayer, G. (2022) Transfer learning in proteins: evaluating novel protein learned representations for bioinformatics tasks. *Brief Bioinform,* **23**(4), bbac232 [PubMed: 35758229] [doi: 10.1093/bib/bbac232].
17. Gasiunas, G., Young, J. K., Karvelis, T., Kazlauskas, D., Urbaitis, T., Jasnauskaite, M., Grusyte, M. M., Paulraj, S., Wang, P.-H., Hou, Z., Dooley, S. K., Cigan, M., Alarcon, C., Chilcoat, N. D., Bigelyte, G., Curcuru, J. L., Mabuchi, M., Sun, Z., Fuchs, R. T., Schildkraut, E., Weigele, P. R., Jack, W. E., Robb, G. B., Venclovas, C., and Siksnys, V. (2020) A catalogue of biochemically diverse CRISPR-Cas9 orthologs. *Nat Commun,* **11**(1), 5512 [PubMed: 33139742] [PubMed Central: PMC7606464] [doi: 10.1038/s41467-020-19344-1].
18. Nguyen, L. T., Macaluso, N. C., Pizzano, B. L. M., Cash, M. N., Spacek, J., Karasek, J., Dinglasan, R. R., Salemi, M., and Jain, P. K. (2021) A Thermostable Cas12b from Brevibacillus Leverages One-pot Detection of SARS-CoV-2 Variants of Concern. *medRxiv,* p. 2021.10.15.21265066 [PubMed: 34704101] [PubMed Central: PMC8547533] [doi: 10.1101/2021.10.15.21265066].
19. Adalsteinsson, B. T., Kristjansdottir, T., Merre, W., Helleux, A., Dusaucy, J., Tourigny, M., Fridjonsson, O., and Hreggvidsson, G. O. (2021) Efficient genome editing of an extreme thermophile, Thermus thermophilus, using a thermostable Cas9 variant. *Sci. Rep.,* **11**(1), 9586 [PubMed: 33953310] [PubMed Central: PMC8100143] [doi: 10.1038/s41598-021-89029-2].
20. Harrington, L. B., Paez-Espino, D., Staahl, B. T., Chen, J. S., Ma, E., Kyrpides, N. C., and Doudna, J. A. (2017) A thermostable Cas9 with increased lifetime in human plasma. *Nat Commun,* **8**(1), 1424 [PubMed: 29127284] [PubMed Central: PMC5681539] [doi: 10.1038/s41467-017-01408-4].
21. Engqvist, M. K. M. Growth Temperatures For 21,498 Microorganisms. (2018) [doi: 10.5281/ZENODO.1175609].
22. Leinonen, R., Diez, F. G., Binns, D., Fleischmann, W., Lopez, R., and Apweiler, R. (2004) UniProt archive. *Bioinformatics,* **20**(17), 3236–3237 [PubMed: 15044231] [doi: 10.1093/bioinformatics/bth191].
23. Nguyen, L. T., Macaluso, N. C., Pizzano, B. L. M., Cash, M. N., Spacek, J., Karasek, J., Miller, M. R., Lednicky, J. A., Dinglasan, R. R., Salemi, M., and Jain, P. K. (2022) A thermostable Cas12b from Brevibacillusleverages one-pot discrimination of SARS-CoV-2 variants of concern. *EBioMedicine,* **77**, 103926 [PubMed: 35290826] [PubMed Central: PMC8917962] [doi: 10.1016/j.ebiom.2022.103926].
24. Pudžiuvelytė, I., Olechnovič, K., Godliauskaite, E., Sermokas, K., Urbaitis, T., Gasiunas, G., and Kazlauskas, D. TemStaPro Datasets. (2023) [doi: 10.5281/ZENODO.7743638].
25. Sasnauskas, G., Tamulaitiene, G., Druteika, G., Carabias, A., Silanskas, A., Kazlauskas, D., Venclovas, C., Montoya, G., Karvelis, T., and Siksnys, V. (2023) TnpB structure reveals minimal functional core of Cas12 nuclease family. *Nature,* **616**(7956), 384–389 [PubMed: 37020015] [doi: 10.1038/s41586-023-05826-x].
26. Eddy, S. R. (2011) Accelerated Profile HMM Searches. *PLoS Comput Biol,* **7**(10), e1002195 [PubMed: 22039361] [PubMed Central: PMC3197634] [doi: 10.1371/journal.pcbi.1002195].
27. Kavuri, N. R., Ramasamy, M., Qi, Y., and Mandadi, K. (2022) Applications of CRISPR/Cas13-Based RNA Editing in Plants. *Cells,* **11**(17), 2665 [PubMed: 36078073] [PubMed Central: PMC9454418] [doi: 10.3390/cells11172665].
28. Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., Comeau, D. C., Connor, R., Funk, K., Kelly, C., Kim, S., Madej, T., Marchler-Bauer, A., Lanczycki, C., Lathrop, S., Lu, Z., Thibaud-Nissen, F., Murphy, T., Phan, L., Skripchenko, Y., Tse, T., Wang, J., Williams, R., Trawick, B. W., Pruitt, K. D., and Sherry, S. T. (2022) Database resources of the national center for biotechnology information. *Nucleic Acids Res,* **50**(D1), D20–D26 [PubMed: 34850941] [PubMed Central: PMC8728269] [doi: 10.1093/nar/gkab1112].
29. Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R., and Wu, C. H. (2007) UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics,* **23**(10), 1282–1288 [PubMed: 17379688] [doi: 10.1093/bioinformatics/btm098].
30. Richardson, L., Allen, B., Baldi, G., Beracochea, M., Bileschi, M. L., Burdett, T., Burgin, J., Caballero-Pérez, J., Cochrane, G., Colwell, L. J., Curtis, T., Escobar-Zepeda, A., Gurbich, T. A., Kale, V., Korobeynikov, A., Raj, S., Rogers, A. B., Sakharova, E., Sanchez, S., Wilkinson, D. J., and Finn, R. D. (2023) MGnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Res,* **51**(D1), D753–D759 [PubMed: 36477304] [PubMed Central: PMC9825492] [doi: 10.1093/nar/gkac1080].
31. Camargo, A. P., Nayfach, S., Chen, I.-M. A., Palaniappan, K., Ratner, A., Chu, K., Ritter, S. J., Reddy, T. B. K., Mukherjee, S., Schulz, F., Call, L., Neches, R. Y., Woyke, T., Ivanova, N. N., Eloe-Fadrosh, E. A., Kyrpides, N. C., and Roux, S. (2023) IMG/VR v4: an expanded database of uncultivated virus genomes within a framework of extensive functional, taxonomic, and ecological metadata. *Nucleic Acids Res,* **51**(D1), D733–D743 [PubMed: 36399502] [PubMed Central: PMC9825611] [doi: 10.1093/nar/gkac1037].
32. Vig, J. and Belinkov, Y. (2019) Analyzing the Structure of Attention in a Transformer Language Model. [doi: 10.48550/ARXIV.1906.04284].

33. Vig, J., Madani, A., Varshney, L. R., Xiong, C., Socher, R., and Rajani, N. F. (2020) BERTology Meets Biology: Interpreting Attention in Protein Language Models. [doi: 10.48550/ARXIV.2006.15222].
34. Kingma, D. P. and Ba, J. Adam: A Method for Stochastic Optimization. (2017) arXiv:1412.6980 [cs].
35. Urbaitis, T., Gasiunas, G., Young, J. K., Hou, Z., Paulraj, S., Godliauskaite, E., Juskeviciene, M. M., Stitilyte, M., Jasnauskaite, M., Mabuchi, M., Robb, G. B., and Siksnys, V. (2022) A new family of CRISPR-type V nucleases with C-rich PAM recognition. *EMBO Rep,* **23**(12), e55481 [PubMed: 36268581] [PubMed Central: PMC9724661] [doi: 10.15252/embr.202255481].
36. Karvelis, T., Druteika, G., Bigelyte, G., Budre, K., Zedaveinyte, R., Silanskas, A., Kazlauskas, D., Venclovas, C., and Siksnys, V. (2021) Transposon-associated TnpB is a programmable RNA-guided DNA endonuclease. *Nature,* **599**(7886), 692–696 [PubMed: 34619744] [PubMed Central: PMC8612924] [doi: 10.1038/s41586-021-04058-1].
37. Khan, S. and Sallard, E. (2023) Current and Prospective Applications of CRISPR-Cas12a in Pluricellular Organisms. *Mol Biotechnol,* **65**(2), 196–205 [PubMed: 35939208] [PubMed Central: PMC9841005] [doi: 10.1007/s12033-022-00538-5].
38. Yang, H., Gao, P., Rajashankar, K. R., and Patel, D. J. (2016) PAM-Dependent Target DNA Recognition and Cleavage by C2c1 CRISPR-Cas Endonuclease. *Cell,* **167**(7), 1814–1828.e12 [PubMed: 27984729] [PubMed Central: PMC5278635] [doi: 10.1016/j.cell.2016.11.053].
39. Altae-Tran, H., Kannan, S., Demircioglu, F. E., Oshiro, R., Nety, S. P., McKay, L. J., Dlakić, M., Inskeep, W. P., Makarova, K. S., Macrae, R. K., Koonin, E. V., and Zhang, F. (2021) The widespread IS200/IS605 transposon family encodes diverse programmable RNA-guided endonucleases. *Science,* **374**(6563), 57–65 [PubMed: 34591643] [PubMed Central: PMC8929163] [doi: 10.1126/science.abj6856].
40. Mougiakos, I., Mohanraju, P., Bosma, E. F., Vrouwe, V., Finger Bou, M., Naduthodi, M. I. S., Gussak, A., Brinkman, R. B. L., van Kranenburg, R., and van der Oost, J. (2017) Characterizing a thermostable Cas9 for bacterial genome editing and silencing. *Nat Commun,* **8**(1), 1647 [PubMed: 29162801] [PubMed Central: PMC5698299] [doi: 10.1038/s41467-017-01591-4].