Using machine learning to predict organismal growth temperatures from protein primary sequences

David B. Sauer[1]* and Da-Neng Wang[1]*

[1]Department of Cell Biology, The Helen L. and Martin S. Kimmel Center for Biology and Medicine, Skirball Institute of Biomolecular Medicine, New York University School of Medicine, New York, New York, United States of America

*Corresponding authors

E-mail: david.sauer@med.nyu.edu (D.B.S), da-neng.wang@med.nyu.edu (DN.W.)

Abstract

The link between a protein's primary sequence and its thermal stability and temperature dependent activity is central to an understanding of protein folding, stability, and evolution. However, the relationship between primary sequence and these biochemical properties can be difficult to quantify, due to the large sequence space and complexity of protein folding. Fortunately, evolution naturally explores both sequence space and temperature space through organismal adaptation to various thermal niches. Here, we use machine learning, in the form of multilayer perceptrons, to predict the originating species' optimal growth temperatures from a protein family's primary sequences. Trained machine learning models outperformed linear regressions in predicting the originating species growth temperature, achieving a root mean squared error of 3.34 °C. Notably, the models are protein family specific, and the predicted organismal growth temperatures are correlated with the proteins' temperatures for melting and optimal activity. Therefore, this method provides a new tool for quickly predicting an organism's optimal growth temperature *in silico*, which can serve as a convenient proxy for protein stability and temperature dependent activity.

Introduction

The relationship between a protein's primary sequence and its biochemical properties is central to the study of protein evolution, folding, and stability. Of particular interest are a protein's temperature dependent properties such as stability and enzymatic activity, as temperature represents the internal energy of a system. Increasing temperature leads to

greater protein flexibility, faster enzymatic kinetics, and eventual protein unfolding, while decreasing temperature reduces protein dynamics and lowers enzymatic activity. This relationship between a protein's primary sequence and temperature dependent properties is also valuable clinically and industrially. For example, proteins which exhibit increased thermal stability are also more resistant to denaturants [1,2] or detergents [3,4] and have longer *in vivo* half-lives [5]. In contrast, pathogenic alleles can code for proteins with decreased thermal stability, leading to reduced expression [6], loss of enzymatic activity [7], and increased disease phenotype [8].

While clearly valuable, it is currently difficult to quantatively describe the relationship between a protein's sequence and its thermal stability or temperature dependent enzymatic activity. Studying this relationship experimentally is difficult due to the large potential sequence space, which grows exponentially with protein length. Therefore, reported experimental methods typically sample only a limited portion of sequence space [3,9], or apply high-throughput techniques [10]. However, these methods still require significant labor, are optimized to identify single point mutations, sample non-native sequences, or are tailored to specific proteins.

Various computational methods have also had success in describing protein stability from sequence. If a three-dimensional protein structure is available, the free energy of the folded sequence can be calculated [11,12]. However, calculating a protein's potential energy in the absence of a structure is effectively equivalent to *de novo* protein

folding, and therefore limited by the vast possible conformational space that grows exponentially with sequence length. Comparative computational methods are available to describe protein stability using only the protein's primary sequence [13–15]. However, these methods are trained with many protein families, and therefore of limited specificity in describing the stability of a particular protein family.

Fortunately, for many protein families natural selection has already broadly sampled both sequence space and temperature space. Homologs belonging to many protein families can be found in organisms that grow at a wide range of temperatures. Organismal growth in each thermal niche places specific constraints on its proteins' sequences such that the proteins are folded and active under native conditions. Accordingly, studies comparing homologus proteins from species with distinct growth temperatures have identified sequence differences which correlate to the native thermal environment of the originating organisms [16–20]. Introducing corresponding mutations into model proteins often result in altered temperature dependent activity or thermal stability, reflecting the role of these amino acids in thermoadaptation. Therefore, the large number of available homologus protein sequences and experimentally determined organismal growth temperatures provides a large dataset for analyzing temperature dependent protein properties, enabling novel methods of analysis.

Here we report protMLP, a generalized method of quantatively predicting the originating organisms' growth temperatures (T_G) from the protein family's primary sequences.

Further, we demonstrate the correlation of this predicted growth temperature to a protein's experimentally determined melting temperature (T_M) or temperature of optimal activity (T_A). Notably, no assumptions are made about the chemical, structural, epistatic, or thermodynamic effects of any particular amino acid, and a protein structure is not used. Thus, predicted organismal growth temperature ($\hat{T}_G$) can serve as a convenient and easily calculable proxy for a protein's thermal stability and temperature dependent activity.

Results

Construction of multi-layer perceptrons

Setting out, we aimed to devise a method to predict organismal growth temperatures from a protein family's primary sequences. As a part of making the method generalizable, we also wanted to avoid an explicit protein structure or description of the forces underlying protein folding and thermostability. We therefore chose machine learning, which has been demonstrated to be particularly useful when the relationship between the input and output is complex or unknown [21,22]. Machine learning has been successful applied to predicting a protein's fold from the primary sequence [23], the genotype of cancers from histopathology images [24], and the antimicrobial activity of a peptide sequence [25]. Similarly, here we apply machine learning in the form of multilayer perceptrons (MLPs) to quantatively predict the originating organism's growth temperature using protein primary sequences.

Generally, a MLP is a form of artificial neural network, a mathematical construct modeled on the structure and behavior of biological neural networks. As with a biological neural network, individual units (nodes or neurons) each accept and process input signals before producing an output. In an MLP these nodes are arranged into layers, termed “hidden layers”, with signals passed between consecutive layers, again mimicking the structure of biological neural networks (Fig. S1A). Starting from the input layer, the value of each node in the hidden layers is the result of an activation function applied to the weighted sum of the preceding layer’s nodes plus a layer specific bias value. The output is then the weighted sum of the final hidden layer and an additional bias value.

The activation function of a MLP node is typically non-linear, mimicking the threshold potential and non-linear response of biological neurons. Central to its application here, MLPs with nodes which apply a non-linear activation function can act as universal approximators [26]. Therefore, we reasoned a sufficiently complex non-linear MLP could describe non-linear interactions, such as electrostatics and van der Waal’s contacts. Further, a non-linear MLP can model logical operators, such as AND and OR, and therefore could likely capture a protein’s epistatic interactions [27–29]. We therefore trained MLPs with nodes that applied the non-linear, leaky rectifier activation function (rMLPs) (Fig. S1B).

As MLPs are mathematical models, the inputs are necessarily numerical. The inputs here are amino acid sequences from a particular protein family. We converted the aligned protein sequences to sequences of Boolean variables (one-hot encoding), where one or zero indicates the presence or absence of a particular amino acid at each position, respectively (Fig. S2). We further removed one-hot encoded amino acids that were absolutely conserved, as these would not contribute to the regression. Therefore, one-hot encoding preserves the chemical sequence of a polypeptide in a numerical sequence of ones and zeroes. Notably, one-hot encoding does not contain a description of the chemical or physical properties of the amino acid. This minimizes any assumptions as to the relevant properties of each side chain, which may be important for regression accuracy as apparently minor changes in side-chain chemistry have been show to result in large changes in a protein's folding and function [30].

For accurate prediction it is necessary to optimize the weight and bias parameters of the MLP. Through a machine learning process termed "training" these values are iteratively refined using homologus protein sequences with known originating organisms' optimal growth temperatures. However, it is essential to have mechanisms to avoid over-fitting and to independently evaluate accuracy [32]. Therefore, we used only 70% of the sequence-T_G pairs in the training process to refine the MLP weight and bias parameters. We used the remaining 30% of the sequence-T_G pairs for evaluating the regressions, assigning the pairs to test (20%) and validation (10%) datasets. We used the validation dataset to avoid over-fitting by calculating the Mean Square Error (MSE)

between true and predicted growth temperatures after each iteration of training, with training stopping when the MSE no longer decreases. The test dataset is then used to calculate final MLP accuracy.

The MLP's optimal number of nodes, and their arrangement into layers - collectively the MLP's "topology" - are not known *a priori*, and are likely specific to the protein family selected. Therefore, for each protein family we considered all possible MLP topologies with the restrictions that: the number of nodes in any hidden layer could range between two and twice the one-hot encoded protein length, the network can have at-most 5 hidden layers, and the network must be over-determined. The number of possible topologies is very large, up to $(2L - 1)^5$, where L is the one-hot encoded multiple sequence alignment length. We therefore applied an evolutionary algorithm to optimize the MLP topology [31]. This consisted of training 500 randomly selected topologies for 10 generations, recombining and randomly permutating the 100 lowest validation MSE topologies of each generation. This method does not completely or evenly sample the entire topology space, and therefore may not find the optimal topology. However, empirically this method is very time efficient finding an optimized MLP topology for the prediction of T_G.

A trained MLP can predict organismal growth temperature from a protein's primary sequence

As an initial prototype for organismal growth temperature prediction we used the Cold Shock Protein (CSP) family of proteins that bind and stabilize nucleic acids. The small protein size and strong conservation across organisms with different ecologies [33] results in many available sequences relative to the protein length from species with a wide range of growth temperatures. This made the CSP family an ideal case study for regression of organismal growth temperature from protein sequence.

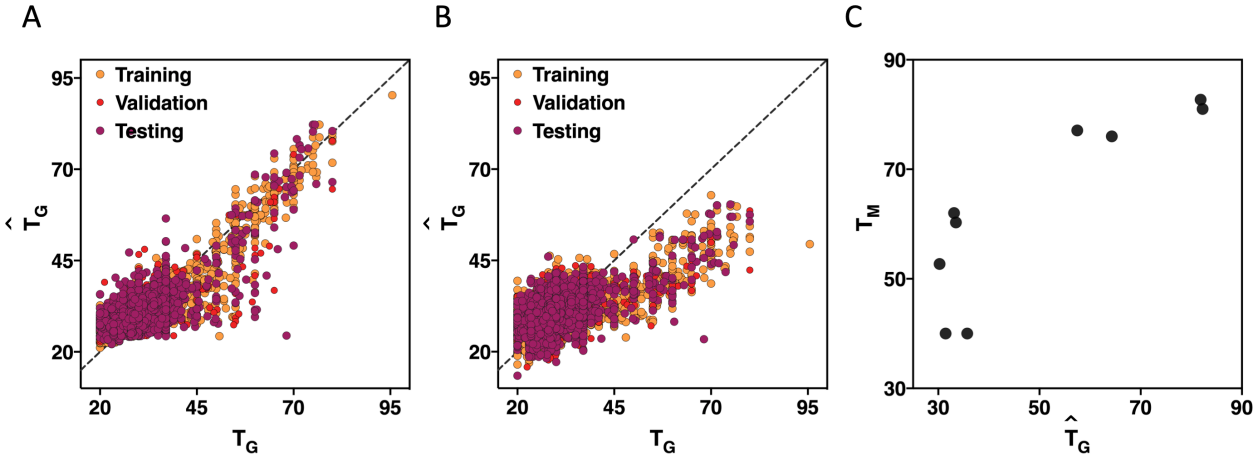

Figure 1. Organismal growth temperature can be predicted by using MLP regression from the primary sequences of thermophiles and mesophiles. Regression of T_G using (A) the best rMLP or (B) linear regression model. C) Predicted organismal growth temperature ($\hat{T}_G$) versus reported T_M for cold shock protein homologs.

Homologus Cold Shock Protein sequences were collected from Pfam [34], extended by one amino based on the results of Perl et al. [19], and aligned in Promals3D [35]. In total 34,068 homologus CSP sequences were identified that had an available source organism growth temperature, with T_Gs measured 4 to 95.5 °C. All protein sequences

were one-hot encoded and rMLPs were trained using the described protMLP algorithm. Of the 10^{17} possible topologies, using the evolutionary algorithm over 10 generations 5000 topologies were trained with 23,759 training sequences. The trained rMLP predicted the source organism growth temperature of 6839 un-seen test sequences, with a root mean squared error of 3.69 °C (r = 0.783) (Fig. S3A).

Notably, this rMLP clearly outperformed a linear regression trained with the same 23,759 training sequences (RMSE = 4.32 °C, r = 0.685), particularly in predicting T_G of proteins from thermophiles (Fig. S3B). However, accuracy in T_G prediction using proteins from psychrophiles ($T_G < 20$ °C) was poor. This is perhaps due to the rarity of these sequences, comprising only 1% of the species-T_G pairs, or differences in the adaptive mechanisms to psychrophilic conditions [36]. Additionally, training minimizes the squared error, which may lead to preferential optimization of sequences from thermophiles due to the positive skew of the T_G distribution (Fig. S3C). Excluding protein sequences from psychrophiles further improved regression accuracy (RMSE = 3.34 °C, r = 0.810) (Fig. 1A), and again outperformed a linear regression (Fig. 1B). This T_G range of ≥20 °C was therefore used in all subsequent studies.

A non-linear activation function is necessary to predict organismal growth temperature

In examining the rMLP topologies trained in the Cold Shock Protein regression, we found three distinct populations of model accuracy (Fig 2A). A low accuracy population

is of topologies that converged to a single constant value (peak a). A second population is of networks with accuracies similar to a linear regression (peak b). This set is unsurprising, as an rMLP can model a linear function. The final population consists of MLP models that are more accurate than a linear regression (peak c), suggesting that a non-linear activation function is essential in increasing regression accuracy.

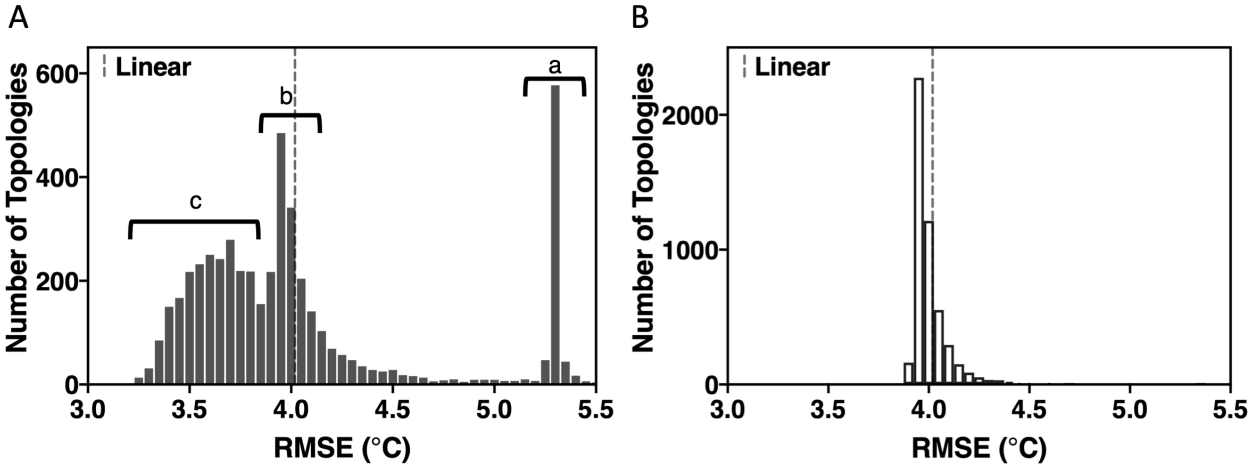

Figure 2. The non-linear activation function is essential for MLP accuracy. Accuracy in predicting the validation dataset for all trained MLPs, using either (A) a rectified or (B) identity activation function. The accuracy of a linear regression is indicated by the dotted line.

However, as multiple MLPs with many parameters are trained, it was necessary to ensure that the improved accuracy of MLP regressions was not due to over-fitting or cherry-picking. Therefore, concurrent with the training of rMLPs for the Cold Shock Protein regression, we trained MLPs of the same topology with an identity activation function, where the activation function output is equal to the input. MLPs with an identity

activation function are mathematically equivalent to linear regressions but fit the same number of parameters as MLPs with a rectified activation function for the same topology. As expected, the accuracy of these MLP regressions with an identity activation function is similar to the linear regression (Fig. 2B). Notably, MLPs using rectified and identity activation functions have distinct distributions (Wilcoxon signed-rank test $p < 10^{-14}$). This confirms that the rectified activation function is essential to the improved prediction accuracy.

We also considered the possibility that protein phylogeny might present benefits and challenges to this analysis, particularly as the collected homologus sequences may include both orthologs and paralogs. Homologs with similar organismal growth temperatures, including paralogs, allow for the identification of T_G relevant amino acids based on sequence conservation [16]. However, sequence identical homologs with similar T_Gs may lead to an over-estimation of MLP accuracy when randomly assigning individual sequences to the training, test, and validation datasets. Therefore it was necessary to examine the effect of sequence similarity on prediction accuracy. We found only a weak ($r = -0.301$) effect of sequence identity on T_G prediction accuracy (Fig. S4A). Further addressing the issue, we generated new training, test, and validation datasets for the Cold Shock Proteins, placing identical sequences into the same dataset. Training MLPs as previously described, the best rMLP predicted the test dataset with a root mean squared error of 3.79 °C ($r = 0.717$) (Fig. S4B). The non-linear MLPs were again more accurate than a linear regression (4.22 °C, $r = 0.624$) and MLPs

trained with an identity activation function (Fig. S4C). These results indicate that sequence identity does not confound the application of non-linear MLPs to predicting organismal growth temperature from a protein sequence.

Non-linear MLPs are necessary for accurate regression of other protein families

We next set out to examine how general rMLPs could be as a method of predicting organismal growth temperature. We therefore trained new regression models of other protein families, using the protMLP algorithm to train MLPs to predict the originating species' T_G from homologus sequences of each family. Examining the Thioredoxin, [2Fe-2S] Ferrodoxin, and MarR families, rMLPs notably outperformed linear regressions in predicting the originating species optimal growth conditions from the primary sequences of homologus proteins (Fig. S5). Notably, the species' growth temperatures predicted using different protein families are strongly correlated, with pairwise Pearson correlation coefficients ranging from 0.761 to 0.848 (pairwise RMSD 2.63 °C to 3.59 °C).

Predicted organismal growth temperature is correlated with experimentally determined melting temperatures of the protein

In order to study the possible application of the protMLP method to thermal stability of the protein, we examined if the predicted organismal growth temperatures of CSP homologs correlate with measured protein melting temperatures. We found

characterized cold shock protein homologs' predicted growth temperatures and measured melting temperatures to be directly correlated (r = 0.860) (Fig. 1C) [19,37–50]. Melting temperatures of Cold Shock Proteins might be expected to be lower than proteins expressed under native growth conditions, as cold shock temperatures are inherently lower than the optimal growth temperatures. However, we observed the CSP homolog's T_Ms are still greater than both rMLP predicted T_Gs and measured T_Gs of the each originating species. This may indicate the temperature difference between organismal optimal growth and the physiological onset of CSP activity is generally small. This could also reflect other functions of CSP homologs at the organisms' optimal growth temperatures [51].

Predicted organismal growth conditions generally correlate with biochemical characteristics of the proteins

We next further examined if the protMLP predicted organismal growth temperature correlated with stability or activity of the protein. We therefore applied the protMLP method to Adenosine Kinases (ADK), a highly conserved protein family that catalyzes the interconversion of adenosine nucleotides. ADK stability and temperature dependent enzymatic activity have been extensively studied [17,18,52]. While there are too few ADK sequences to train an over-determined MLP, a linear regression is already highly accurate at predicting the originating species' growth temperature (RMSE = 3.78 °C, r = 0.836) (Fig. 3A). Furthermore, we found a strong correlation between the calculated $\hat{T}_G$s for characterized ADK homologs and reconstructed ancestral sequences and protein

melting temperatures (r=0.787) (Fig. 3B) and temperatures of optimal enzymatic activity (r = 0.650) (Fig. 3C) [17,53].

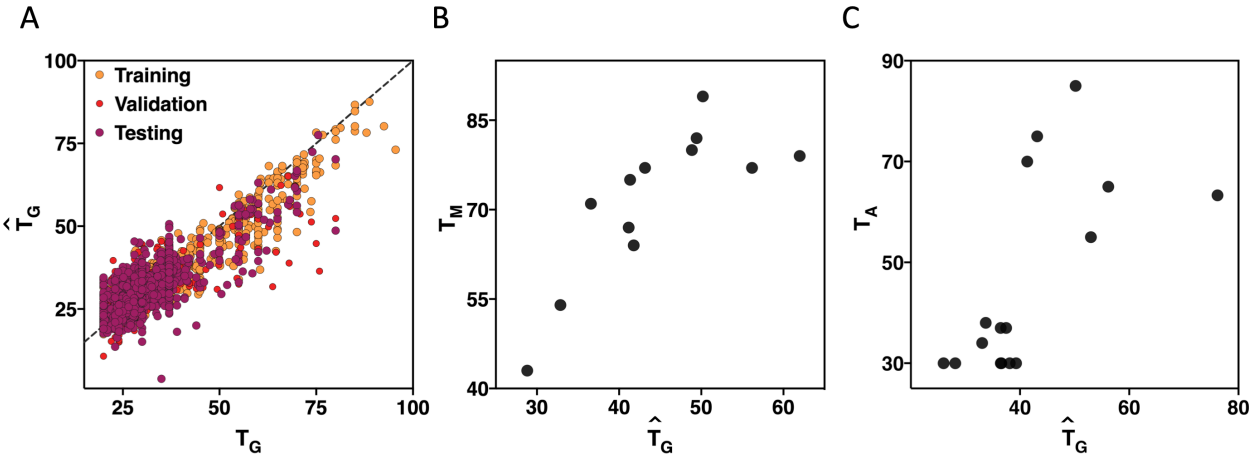

Figure 3. T_Gs predicted from linear regression of ADK sequences correlate with biochemical characteristics. A) Linear regression of T_G from ADK sequences. B) Predicted T_G versus reported T_M for ADK homologs. C) Predicted T_G versus reported T_A for ADK homologs.

The relatively few sequences from thermophiles are necessary but sufficient

In examining sequences used for the CSP MLP regression, we noted that 98.5% of sequences are from mesophiles (Fig. S6A). This was unsurprising given the bias of the characterized and sequenced organisms [16]. However, it was therefore necessary to ensure that this skew in sequence-growth temperature pairs did not confound the rMLP training.

We first examined if sequences from mesophiles alone sufficiently sampled sequence space to accurately predict the T_G of homologs from thermophiles. If successful, this would indicate that the thermoadaptive sequence differences between homologs from mesophiles and thermophiles are contained within the sequence space sampled by mesophiles alone. However, limiting the training and validation datasets to only Cold Shock Protein homologs from mesophiles reduced regression accuracy (RMSE = 4.32 °C, r = 0.643), with a clear systematic under-prediction of proteins from extremophiles (Fig. S6B). Therefore, proteins from thermophiles likely contain amino acid sequences that are outside the sequence variation seen within CSP homologs from mesophiles.

We also examined if the non-uniform distribution of organismal growth temperatures in the training dataset hindered the accuracy of the regression. This would be possible if, during training, the optimization of MLP weights and biases was dominated by the small but numerous differences in T_G among the protein sequences from mesophiles. We therefore calculated rMLPs for the Cold Shock Protein family after “balancing” the training dataset by artificially over-sampling sequences from thermophiles (Fig. S6C), while validation and test datasets remained unchanged. The accuracy of the MLPs in predicting the unseen test dataset was slightly worse than without balancing (RMSE = 3.74 °C, r = 0.775) (Fig. S6D). As the number of unique sequences from thermophiles is much smaller than those from mesophiles, the oversampling of the sequences from thermophiles may have lead to over-fitting of inconsequential amino acids unique to these sequences.

Together, these results make clear that the presence of relatively few (1.47%) sequences from thermophiles in the training dataset are necessary and sufficient for the prediction of optimal growth temperature of homologs from thermophiles. The bias of the available protein sequences and species T_Gs does not appear to have deleteriously harmed regression accuracy, though accuracy may increase with more unique homologs from thermophiles with an associated organismal growth temperature.

Non-linearity regressions improve T_G prediction accuracy even with fewer sequences

The ability of a rMLP to model increasingly complex functions is dependent upon increased network depth and width. However, as network topology is required to be over-determined, network complexity is limited by the number of training sequence – organismal growth temperature pairs. To examine how regression accuracy scales with the number of sequences, we generated smaller Cold Shock Protein training and validation datasets by random sampling. With the test set for evaluating regression accuracy remaining unchanged, linear regression and MLPs were trained as previously (Fig. 4). It was not possible to build an over-determined MLP with 10% of the training sequences. However, the rectified activation function clearly outperformed an identity activation function at 20% of the training and validations sequences, or 4,691 and 690 sequences, respectively. This suggests that as few as 3.15 training sequences per one-hot encoded amino acid, or 24.6 sequences per column of the multiple sequence

alignment, are sufficient to capture non-linear effects on the relationship between protein sequence and organismal optimal growth temperature.

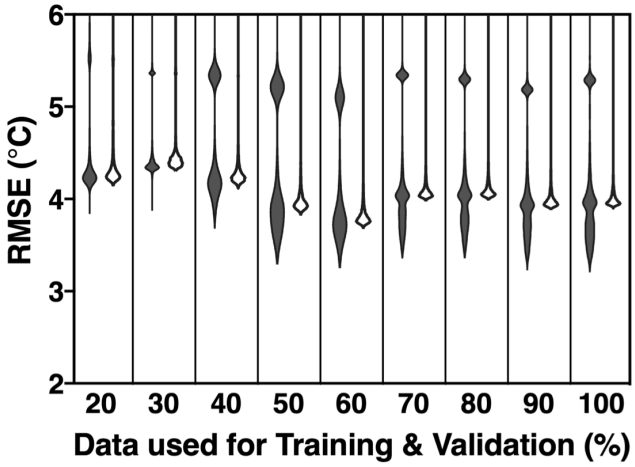

Figure 4. The proportion of non-linear MLP topologies outperforming equivalent topologies with a linear activation function increases with more training data. MLP accuracy trained using subsets of the training and validation sequences with either rectified (filled) or identity (unfilled) activation functions.

Particular amino acids are key to organismal growth temperature prediction

In requiring the MLPs to be over-determined, we realized this could preclude longer or less well conserved protein families from analysis. Fortunately, previous studies had indicated that only a small fraction of mutations to a protein's primary sequence alter protein stability [3,9,19,54]. We hypothesize that most primary sequence differences were neutral to thermoadaptation, analogous to passenger mutations. Therefore most one-hot encoded amino acids would not contribute to the accuracy of the regression,

while potentially adding noise to the regression and decreasing the maximum complexity of the topologies examined. To test this hypothesis, we examined the correlation of each one-hot encoded position with T_G and if excluding un-correlated amino acids would improve regression accuracy.

We identified first-order correlation between amino acid presence or absence and the originating species' growth temperature using the point-biserial correlation coefficient (Fig. S7A). Excluding those encoded amino acids with a correlation less than 0.1, we achieved similar accuracy as before (RMSE = 3.75 °C, r = 0.761) while using only 3.93% of the encoded protein sequence (Fig. S7B). We similarly used a fit top-hat function to identify amino acids with a second-order correlation to growth temperature (Fig. S7C). While only 25.4% of the amino acids had a maximum correlation to a top-hat function of greater than 0.1, these amino acids could predict growth temperature with a root mean squared error of 3.40 °C using an rMLP (r = 0.805) (Fig. S7D).

These results confirm that only a subset of amino acids in the sequence is needed to accurately predict the originating species' growth temperature. Therefore, using only the most T_G correlated amino acids would allow for the regression of longer proteins. Alternatively, deeper and wider topologies could be examined on shorter proteins, potentially improving accuracy by accounting for more complex interactions in the primary sequence.

Discussion

The design or identification of thermoadapted proteins is often central to their study or for their use in industrial applications. However, the study of protein thermal stability or temperature dependent activity is challenged by the large potential sequence space and the difficulty of characterizing individual protein sequences.

Here, we successfully generated mathematical models to predict the originating species' optimal growth temperature from a protein's primary sequence. Growth temperatures could be predicted with a root mean squared error of 3.34 °C, and required as few as 24.6 sequences per column of the multiple sequence alignment. These predicted T_Gs correlate with experimentally determined melting temperatures and temperatures of optimal activity. Therefore, this method allows for the rapid evaluation of protein sequences *in silico*, with the predicted values expected to correlate with protein thermal stability and temperature dependent activity.

Linear regressions are sufficient for some protein families

The linear contribution of particular amino acids to thermostability is seen in some membrane [55] and soluble proteins [19], including the ADK family (Fig. 2). However, non-linear effects are clearly central to thermal stability of the Arc repressor [27] and in the prediction of organismal growth temperatures for many protein families seen here (Fig. 1A and Fig. S5). The varied success of linear regression models in predicting organismal growth temperature from primary sequence supports the hypothesis that the

physical interactions that underlie thermoadaptation vary by protein family [56]. As a rMLP can model a linear regression, in addition to more complex functions, the protMLP algorithm likely represents a general solution to describing the relation between primary sequence and quantitative characteristics of the protein.

Protein families available for analysis will increase

We recognize that construction of machine learning models is inherently limited by the number of sequence - organismal growth temperature pairs available. For Cold Shock Proteins, 24.6 sequences per column of the multiple sequence alignment were sufficient for non-linear MLPs to outperform a linear regression. The number of homologus sequences available for training is likely to increase as more organisms are sequenced. However, any new homologus sequences are only useful if they have an associated organismal growth temperature. Notably, with the CSP family examined here, 48% of the sequences were discarded due to an unknown organismal growth temperature. Further, the number of protein sequences with an unknown T_G will likely increase as uncharacterized and unknown organisms are sequenced through metagenomics. Fortunately, computation methods are available to predict organismal growth temperatures from the genomic sequences of uncharacterized organisms [57,58], providing T_Gs for homologus proteins from species whose growth temperatures have not been experimentally determined.

Single mutant accuracy requires densely sampled sequence space

In principle, a trained rMLP should be sensitive to the effects of a single or few amino acid differences, such as experimentally generated point mutations. We therefore examined the correlation of rMLP predicted growth temperatures to the measured melting temperatures for mutants of a CSP ortholog from *Bacillus subtillis* (BsCSP). We found no correlation between mutant protein melting temperatures [47] and predicted organismal optimal growth temperatures calculated from the mutant proteins' sequences (r = -0.134) (Fig. S8A). Comparing BsCSP to the training sequences, we noted that homologs with high sequence identity to BsCSP come from organisms with T_Gs similar to *Bacillus subtillis* (Fig. S8B). This is in contrast to BsCSP mutants, with only one or two amino acid changes, exhibiting significantly altered melting temperature from wild type [19,47]. We therefore suspect that the available CSP sequences do not sufficiently sample sequence space to capture the effects of few or rare primary sequence differences.

To verify this hypothesis we applied the protMLP method to the densely sampled sequence space of the deeply mutagenized WW domain from the human Yes Associated Protein 65 [59]. Rather than reporting organismal growth temperatures, the study describes protein enrichment upon binding to a target peptide. Nevertheless, both dataset consists of pairs of protein sequences and numerical values. Therefore protMLP should be capable of predicting the enrichment scores for the WW domain sequences. Examining this, regressions were calculated as before, replacing organismal growth temperature with enrichment score as the regression target. Notably, a non-linear rMLP

can accurately predict the enrichment of mutant WW sequences upon binding a target peptide (RMSE = 0.575, r = 0.862) (Fig. 5). As with predicting organismal growth temperature, rMLPs significantly outperformed equivalent MLP topologies with an identity activation function (Wilcoxon signed-rank test $p < 10^{-99}$). Notably, this dataset consists of only single, double, and triple mutants, corresponding to 91-97% sequence identity. This result demonstrates that rMLPs can accurately predict the effects of few mutations with sufficiently sampled sequence space.

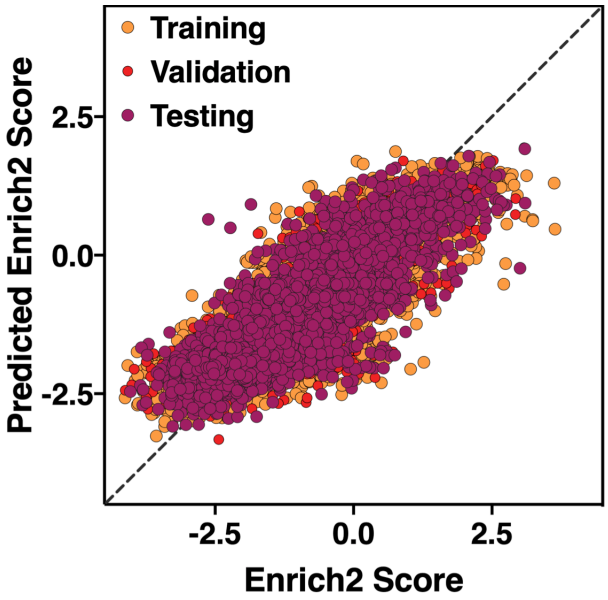

Figure 5. Non-linear MLPs can predict single and double mutant effects. Predicted enrichment for mutant WW domain sequences.

Application to other biochemical parameters and evolutionary biology

Though the thermoadaptation of proteins is the focus of this study, the accurate prediction of enrichment for mutant WW domains also validates the rMLP method as

readily applicable to other quantifiable characteristics of proteins. Additionally, quantifiable characteristics of nucleic acids are likely predictable using the same method.

Finally, we also note that the prediction of organismal ecological characteristics from protein sequences is itself valuable. While other methods are capable of predicting organismal growth temperatures [57,58], protMLP calculates $\hat{T}_G$ without requiring a complete genome or proteome sequence for the organism. This is particularly useful if the organism of interest no longer exists, such as ancestral organisms. By analyzing reconstructed ancestral sequences, protMLP could describe the thermal niche of no longer extant organisms. Though the ecological niches of ancestral organisms have been inferred from the reconstructed proteins' melting temperatures [17,60,61], by predicting organismal growth temperature *in silico*, protMLP is faster and likely more accurate.

Materials and Methods

Sequence and organismal growth temperature collection and encoding

Species' T_G values were collected from Sauer and Wang (2015) [16], Engqvist (2018) [62], and BacDive (accessed March 14, 2019) [63], and averaging values of the same species. Domain sequence alignments were downloaded from the Pfam 32.0 database [34] and used without modification unless otherwise noted. Reconstructed ancestral ADK sequences from Nguyen et al. [17] were combined with extant proteins from the

Pfam alignment. CSP sequences identified in Pfam were extended by one amino acid using the sequences in UniProtBK release 2018_04 [64]. CSP and ADK sequences were then re-aligned in Promals3D [35]. All CSP and ADK sequences with characterized T_A and T_M values were removed from the alignments prior to division into training, test, and validation datasets; and used only for the comparison of $\hat{T}_G$ to T_M and T_A.

Species assignment for each protein was collected from UniProtKB release 2018_04 [64]. Gap inducing proteins, proteins annotated as fragments, or proteins without an originating species' T_G were excluded from analysis. Proteins were randomly assigned individually into training (70%), validation (10%), or test (20%) datasets. The amino acid sequences were then one-hot encoded, and amino acids which were absent or absolutely conserved in the training sequences were removed from all alignments.

Balancing Training Sequences

Training data was balanced by first calculating a histogram of training sequence T_Gs with 20 bins. In addition to all the sequences in the original alignment, sequences were added to the alignment by random selection with replacement from each T_G bin until all bins had the same number of sequences as the most populous bin.

Amino acid correlation with T_G

From the one-hot encoded training sequences, the point-biserial correlation coefficient was calculated between T_G and the presence or absence of a particular amino acid. Alternatively, top-hat function was fit to the presence or absence of a one-hot encoded amino acid versus T_G by systematically screening hat widths and centers. If a threshold was provided, those positions with a Pearson correlation coefficient of fit top-hat function or point-biserial correlation coefficient less than the threshold were removed.

MLP training

The MLPs were trained using an identity or leaky ReLu activation function (alpha = 0.01) [65]. All regressions were trained with the training dataset using the Adam solver [66], with the mean square error (MSE) as the loss function. Training was stopped when the validation dataset MSE did not decrease for two consecutive training epochs.

Topology generation and search

MLP architectures were built systematically, requiring only that the first layer have at most twice as many nodes as the input layer, all subsequent layers have less than or equal to as many nodes as the previous layer, and that the network be over-determined. Topologies were limited to 5 or fewer hidden layers. Of the potential topologies, 500 were randomly selected and trained each generation for 10 generations. After each generation, the top scoring 20% of the topologies (based on the MSE of the validation dataset) were recombined and mutated, and used as input for the following generation. Recombining topologies consisted of joining two topologies at a random layer chosen

from each. Topologies were mutated by randomly changing the number of nodes in a randomly chosen layer. Finally, the Pearson correlation coefficient and root mean square error was calculated and reported using the test dataset for the best trained model of the last generation. For comparing inter-family species growth temperature prediction consistency, $\hat{T}_G$s from the test sets were averaged by species and then compared pairwise by protein family.

Regression of WW enrichment

WW domain mutations and Enrich2 scores were downloaded from MaveDB [59,67]. Mutant sequences were generated *in silico*, and regressions calculated as previously described, using Enrich2 scores as the regression target.

All calculations used custom scripts written in Python with the Biopython [68], Tensorflow [69], Keras [70], NumPy [71], SciPy [72], and Matplotlib [73] libraries. Source code is available at https://github.com/DavidBSauer/protMLP

Acknowledgements

The authors thank David Fenyo for helpful discussion of this work, and Jennifer Marden for critical review of this manuscript. This work was supported by the National Institutes of Health (R01-GM121994 and R01-NS108151 to D-N.W). D.B.S. was supported in part by a Postdoctoral Fellowship (PF-17-135-01) from the American Cancer Society and by the Office of the Assistant Secretary of Defense for Health Affairs, through the Peer

Reviewed Cancer Research Program under Award No. W81XH-16-1-0153. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the Department of Defense.

1. Owusu-Apenten R, Cowan D. Correlation between microbial protein thermostability and resistance to denaturation in aqueous:organic solvent two-phase systems. Enzyme Microb Technol. 1989;11: 568–574.

2. Cowan DA. Thermophilic proteins: stability and function in aqueous and organic solvents. Comp Biochem Physiol A Physiol. 1997;118: 429–438.

3. Serrano-Vega MJ, Magnani F, Shibata Y, Tate CG. Conformational thermostabilization of the beta1-adrenergic receptor in a detergent-resistant form. Proc Natl Acad Sci U S A. 2008;105: 877–882. doi:10.1073/pnas.0711253105

4. Zhou Y, Bowie JU. Building a thermostable membrane protein. J Biol Chem. 2000;275: 6975–6979.

5. Gao D, Narasimhan DL, Macdonald J, Brim R, Ko M-C, Landry DW, et al. Thermostable Variants of Cocaine Esterase for Long-Time Protection against Cocaine Toxicity. Mol Pharmacol. 2009;75: 318–323. doi:10.1124/mol.108.049486

6. Tate CG. A crystal clear solution for determining G-protein-coupled receptor structures. Trends Biochem Sci. 2012;37: 343–352. doi:10.1016/j.tibs.2012.06.003

7. Abrahamson M, Grubb A. Increased body temperature accelerates aggregation of the Leu-68-->Gln mutant cystatin C, the amyloid-forming protein in hereditary cystatin C amyloid angiopathy. Proc Natl Acad Sci. 1994;91: 1416–1420. doi:10.1073/pnas.91.4.1416

8. Raimondo A, Chakera AJ, Thomsen SK, Colclough K, Barrett A, De Franco E, et al. Phenotypic severity of homozygous GCK mutations causing neonatal or childhood-onset diabetes is primarily mediated through effects on protein stability. Hum Mol Genet. 2014;23: 6432–6440. doi:10.1093/hmg/ddu360

9. Abdul-Hussein S, Andréll J, Tate CG. Thermostabilisation of the serotonin transporter in a cocaine-bound conformation. J Mol Biol. 2013;425: 2198–2207. doi:10.1016/j.jmb.2013.03.025

10. Sarkar CA, Dodevski I, Kenig M, Dudli S, Mohr A, Hermans E, et al. Directed evolution of a G protein-coupled receptor for expression, stability, and binding selectivity. Proc Natl Acad Sci. 2008;105: 14808–14813. doi:10.1073/pnas.0803103105

11. Yasuda S, Kajiwara Y, Toyoda Y, Morimoto K, Suno R, Iwata S, et al. Hot-Spot Residues to be Mutated Common in G Protein-Coupled Receptors of Class A: Identification of Thermostabilizing Mutations Followed by Determination of Three-Dimensional Structures for Two Example Receptors. J Phys Chem B. 2017;121: 6341–6350. doi:10.1021/acs.jpcb.7b02997

12. Pucci F, Bourgeas R, Rooman M. Predicting protein thermal stability changes upon point mutations using statistical potentials: Introducing HoTMuSiC. Sci Rep. 2016;6: 23257. doi:10.1038/srep23257

13. Montanucci L, Fariselli P, Martelli PL, Casadio R. Predicting protein thermostability changes from sequence upon multiple mutations. Bioinforma Oxf Engl. 2008;24: i190-195. doi:10.1093/bioinformatics/btn166

14. Gromiha MM, Suresh MX. Discrimination of mesophilic and thermophilic proteins using machine learning algorithms. Proteins. 2008;70: 1274–1279. doi:10.1002/prot.21616

15. Li Y, Middaugh CR, Fang J. A novel scoring function for discriminating hyperthermophilic and mesophilic proteins with application to predicting relative thermostability of protein mutants. BMC Bioinformatics. 2010;11: 62. doi:10.1186/1471-2105-11-62

16. Sauer DB, Karpowich NK, Song JM, Wang D-N. Rapid Bioinformatic Identification of Thermostabilizing Mutations. Biophys J. 2015;109: 1420–1428. doi:10.1016/j.bpj.2015.07.026

17. Nguyen V, Wilson C, Hoemberger M, Stiller JB, Agafonov RV, Kutter S, et al. Evolutionary drivers of thermoadaptation in enzyme catalysis. Science. 2017;355: 289–294. doi:10.1126/science.aah3717

18. Wolf-Watz M, Thai V, Henzler-Wildman K, Hadjipavlou G, Eisenmesser EZ, Kern D. Linkage between dynamics and catalysis in a thermophilic-mesophilic enzyme pair. Nat Struct Mol Biol. 2004;11: 945–949. doi:10.1038/nsmb821

19. Perl D, Mueller U, Heinemann U, Schmid FX. Two exposed amino acid residues confer thermostability on a cold shock protein. Nat Struct Biol. 2000;7: 380–383. doi:10.1038/75151

20. Szilágyi A, Závodszky P. Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: results of a comprehensive survey. Structure. 2000;8: 493–504. doi:10.1016/S0969-2126(00)00133-7

21. Webb S. Deep learning for biology. Nature. 2018;554: 555–557. doi:10.1038/d41586-018-02174-z

22. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521: 436–444. doi:10.1038/nature14539

23. AlQuraishi M. AlphaFold at CASP13. Valencia A, editor. Bioinformatics. 2019; btz422. doi:10.1093/bioinformatics/btz422

24. Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyö D, et al. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. Nat Med. 2018;24: 1559–1567. doi:10.1038/s41591-018-0177-5

25. Veltri D, Kamath U, Shehu A. Deep learning improves antimicrobial peptide recognition. Hancock J, editor. Bioinformatics. 2018;34: 2740–2747. doi:10.1093/bioinformatics/bty179

26. Hornik K. Approximation capabilities of multilayer feedforward networks. Neural Netw. 1991;4: 251–257. doi:10.1016/0893-6080(91)90009-T

27. Brown BM, Sauer RT. Tolerance of Arc repressor to multiple-alanine substitutions. Proc Natl Acad Sci. 1999;96: 1983–1988. doi:10.1073/pnas.96.5.1983

28. Starr TN, Thornton JW. Epistasis in protein evolution. Protein Sci Publ Protein Soc. 2016;25: 1204–1218. doi:10.1002/pro.2897

29. Ghosh S, Bierig T, Lee S, Jana S, Löhle A, Schnapp G, et al. Engineering Salt Bridge Networks between Transmembrane Helices Confers Thermostability in G-Protein-Coupled Receptors. J Chem Theory Comput. 2018;14: 6574–6585. doi:10.1021/acs.jctc.8b00602

30. Sauer DB, Zeng W, Raghunathan S, Jiang Y. Protein interactions central to stabilizing the K+ channel selectivity filter in a four-sited configuration for selective K+ permeation. Proc Natl Acad Sci. 2011;108: 16634–16639. doi:10.1073/pnas.1111688108

31. Vikhar PA. Evolutionary algorithms: A critical review and its future prospects. 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC). Jalgaon, India: IEEE; 2016. pp. 261–265. doi:10.1109/ICGTSPICC.2016.7955308

32. Zou J, Huss M, Abid A, Mohammadi P, Torkamani A, Telenti A. A primer on deep learning in genomics. Nat Genet. 2019;51: 12–18. doi:10.1038/s41588-018-0295-5

33. Phadtare S, Alsina J, Inouye M. Cold-shock response and cold-shock proteins. Curr Opin Microbiol. 1999;2: 175–180. doi:10.1016/S1369-5274(99)80031-9

34. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 2016;44: D279-285. doi:10.1093/nar/gkv1344

35. Pei J, Kim B-H, Grishin NV. PROMALS3D: a tool for multiple protein sequence and structure alignments. Nucleic Acids Res. 2008;36: 2295–2300. doi:10.1093/nar/gkn072

36. Yang L-L, Tang S-K, Huang Y, Zhi X-Y. Low Temperature Adaptation Is Not the Opposite Process of High Temperature Adaptation in Terms of Changes in Amino Acid Composition. Genome Biol Evol. 2015;7: 3426–3433. doi:10.1093/gbe/evv232

37. D'Auria G, Esposito C, Falcigno L, Calvanese L, Iaccarino E, Ruggiero A, et al. Dynamical properties of cold shock protein A from *Mycobacterium tuberculosis*. Biochem Biophys Res Commun. 2010;402: 693–698. doi:10.1016/j.bbrc.2010.10.086

38. Jin B, Jeong K-W, Kim Y. Structure and flexibility of the thermophilic cold-shock protein of *Thermus aquaticus*. Biochem Biophys Res Commun. 2014;451: 402–407. doi:10.1016/j.bbrc.2014.07.127

39. Lee J, Jeong K-W, Jin B, Ryu K-S, Kim E-H, Ahn J-H, et al. Structural and dynamic features of cold-shock proteins of *Listeria monocytogenes*, a psychrophilic bacterium. Biochemistry. 2013;52: 2492–2504. doi:10.1021/bi301641b

40. Welker C, Böhm G, Schurig H, Jaenicke R. Cloning, overexpression, purification, and physicochemical characterization of a cold shock protein homolog from the hyperthermophilic bacterium *Thermotoga maritima*. Protein Sci Publ Protein Soc. 1999;8: 394–403. doi:10.1110/ps.8.2.394

41. Phadtare S, Hwang J, Severinov K, Inouye M. CspB and CspL, thermostable cold-shock proteins from *Thermotoga maritima*. Genes Cells Devoted Mol Cell Mech. 2003;8: 801–810.

42. Wassenberg D, Welker C, Jaenicke R. Thermodynamics of the unfolding of the cold-shock protein from *Thermotoga maritima* 1 1Edited by A. R. Fersht. J Mol Biol. 1999;289: 187–193. doi:10.1006/jmbi.1999.2772

43. Chatterjee S, Jiang W, Emerson SD, Inouye M. The Backbone Structure of the Major Cold-Shock Protein CS7.4 of *Escherichia coli* in Solution Includes Extensive β-Sheet Structure1. J Biochem (Tokyo). 1993;114: 663–669. doi:10.1093/oxfordjournals.jbchem.a124234

44. Petrosian SA, Makhatadze GI. Contribution of proton linkage to the thermodynamic stability of the major cold-shock protein of *Escherichia coli* CspA. Protein Sci. 2008;9: 387–394. doi:10.1110/ps.9.2.387

45. Phadtare S, Inouye M, Severinov K. The Nucleic Acid Melting Activity of *Escherichia coli* CspE Is Critical for Transcription Antitermination and Cold Acclimation of Cells. J Biol Chem. 2002;277: 7239–7245. doi:10.1074/jbc.M111496200

46. Phadtare S, Tyagi S, Inouye M, Severinov K. Three Amino Acids in *Escherichia coli* CspE Surface-exposed Aromatic Patch Are Critical for Nucleic Acid Melting Activity Leading to Transcription Antitermination and Cold Acclimation of Cells. J Biol Chem. 2002;277: 46706–46711. doi:10.1074/jbc.M208118200

47. Wunderlich M, Martin A, Schmid FX. Stabilization of the Cold Shock Protein CspB from *Bacillus subtilis* by Evolutionary Optimization of Coulombic Interactions. J Mol Biol. 2005;347: 1063–1076. doi:10.1016/j.jmb.2005.02.014

48. Martin A, Kather I, Schmid FX. Origins of the High Stability of an in vitro-selected Cold-shock Protein. J Mol Biol. 2002;318: 1341–1349. doi:10.1016/S0022-2836(02)00243-7

49. Garcia-Mira MM, Boehringer D, Schmid FX. The Folding Transition State of the Cold Shock Protein is Strongly Polarized. J Mol Biol. 2004;339: 555–569. doi:10.1016/j.jmb.2004.04.011

50. Mueller U, Perl D, Schmid FX, Heinemann U. Thermal stability and atomic-resolution crystal structure of the *Bacillus caldolyticus* cold shock protein. J Mol Biol. 2000;297: 975–988. doi:10.1006/jmbi.2000.3602

51. Keto-Timonen R, Hietala N, Palonen E, Hakakorpi A, Lindström M, Korkeala H. Cold Shock Proteins: A Minireview with Special Emphasis on Csp-family of Enteropathogenic *Yersinia*. Front Microbiol. 2016;7. doi:10.3389/fmicb.2016.01151

52. Schrank TP, Bolen DW, Hilser VJ. Rational modulation of conformational fluctuations in adenylate kinase reveals a local unfolding mechanism for allostery and functional adaptation in proteins. Proc Natl Acad Sci. 2009;106: 16984–16989. doi:10.1073/pnas.0906510106

53. Jeske L, Placzek S, Schomburg I, Chang A, Schomburg D. BRENDA in 2019: a European ELIXIR core data resource. Nucleic Acids Res. 2019;47: D542–D549. doi:10.1093/nar/gky1048

54. Milla ME, Brown BM, Sauer RT. Protein stability effects of a complete set of alanine substitutions in Arc repressor. Nat Struct Biol. 1994;1: 518.

55. Sarkar CA, Dodevski I, Kenig M, Dudli S, Mohr A, Hermans E, et al. Directed evolution of a G protein-coupled receptor for expression, stability, and binding selectivity. Proc Natl Acad Sci U S A. 2008;105: 14808–14813. doi:10.1073/pnas.0803103105

56. Petsko GA. Structural basis of thermostability in hyperthermophilic proteins, or "there's more than one way to skin a cat." Methods Enzymol. 2001;334: 469–478.

57. Wang D-N, Sauer DB. Predicting the optimal growth temperatures of prokaryotes using only genome derived features. 2019; doi:10.1093/bioinformatics/btz059

58. Li G, Rabe KS, Nielsen J, Engqvist MK. Machine learning applied to predicting microorganism growth temperatures and enzyme catalytic optima: Supplementary information. bioRxiv. 2019; doi:10.1101/522342

59. Fowler DM, Araya CL, Fleishman SJ, Kellogg EH, Stephany JJ, Baker D, et al. High-resolution mapping of protein sequence-function relationships. Nat Methods. 2010;7: 741–746. doi:10.1038/nmeth.1492

60. Wheeler LC, Lim SA, Marqusee S, Harms MJ. The thermostability and specificity of ancient proteins. Curr Opin Struct Biol. 2016;38: 37–43. doi:10.1016/j.sbi.2016.05.015

61. Gaucher EA, Govindarajan S, Ganesh OK. Palaeotemperature trend for Precambrian life inferred from resurrected proteins. Nature. 2008;451: 704–707. doi:10.1038/nature06510

62. Engqvist MKM. Correlating enzyme annotations with a large set of microbial growth temperatures reveals metabolic adaptations to growth at diverse temperatures. BMC Microbiol. 2018;18. doi:10.1186/s12866-018-1320-7

63. Söhngen C, Podstawka A, Bunk B, Gleim D, Vetcininova A, Reimer LC, et al. BacDive--The Bacterial Diversity Metadatabase in 2016. Nucleic Acids Res. 2016;44: D581-585. doi:10.1093/nar/gkv983

64. The UniProt Consortium. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2017;45: D158–D169. doi:10.1093/nar/gkw1099

65. Maas AL, Hannun AY, Ng AY. Rectifier nonlinearities improve neural network acoustic models. in ICML Workshop on Deep Learning for Audio, Speech and Language Processing. 2013.

66. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. 2014.

67. Esposito D, Weile J, Shendure J, Starita LM, Papenfuss AT, Roth FP, et al. An open-source platform to distribute and interpret data from multiplexed assays of variant effect. bioRxiv. 2019; doi:10.1101/555797

68. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinforma Oxf Engl. 2009;25: 1422–1423. doi:10.1093/bioinformatics/btp163

69. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al. TensorFlow: A System for Large-Scale Machine Learning. 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16). Savannah, GA: USENIX Association; 2016. pp. 265–283. Available: https://www.usenix.org/conference/osdi16/technical-sessions/presentation/abadi

70. Chollet F. keras [Internet]. GitHub; 2015. Available: https://github.com/fchollet/keras

71. Oliphant T. NumPy: A guide to NumPy [Internet]. 2006. Available: http://www.numpy.org/

72. Jones E, Oliphant T, Peterson P, others. SciPy: Open source scientific tools for Python [Internet]. 2001. Available: http://www.scipy.org/

73. Hunter JD. Matplotlib: A 2D Graphics Environment. Comput Sci Eng. 2007;9: 90–95. doi:10.1109/MCSE.2007.55