Using machine learning to predict quantitative phenotypes from protein and nucleic acid 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 sequence and phenotype is essential to understanding the molecular mechanisms of evolution, and the design of proteins and genes with specific properties. However, it is difficult to describe the relationship between sequence and protein or organismal phenotypes, due to the complex relationship between sequence, protein folding and activity, and organismal physiology. Here, we use machine learning models trained on individual families of proteins or nucleic acids to predict the originating species' optimal growth temperatures or other quantitative phenotypes. Trained multilayer perceptrons (MLPs) outperformed linear regressions in predicting the originating species growth temperature from protein sequences, achieving a root mean squared error of 3.6 °C. Similar machine learning models were able to predict the binding affinity of mutant WW domain sequences, brightness of fluorescent proteins, and enzymatic activity of ribozymes. Notably, the trained models are protein or nucleic acid family specific and therefore useful in the design of biopolymers with particular properties. This method provides a new tool for the *in silico* prediction of quantitative biophysical and organismal phenotypes directly from sequence.

1 Introduction

The relationship between a protein or nucleic acid's sequence and the biochemical or organismal phenotype is central to the study of evolution, protein folding, and enzymatic activity. Further, knowing a protein's sequence and

temperature dependent properties is also clinically and industrially valuable. For example, mutations may alter a protein's sequence such that it has reduced expression (Tate, 2012) or enzymatic activity (Abrahamson and Grubb, 1994), or the organism exhibits an increased disease phenotype (Raimondo *et al.*, 2014). Accordingly, highly tailored high-throughput experimental methods have been designed to efficiently screen the correspondence of sequence and the phenotypes of protein stability (C. A. Sarkar *et al.*, 2008; Serrano-Vega *et al.*, 2008; Abdul-Hussein *et al.*, 2013) or abundance (Matreyek *et al.*, 2018).

The ability to predict phenotype from sequence *in silico* is clearly of value. However, the link between sequence and phenotype is often difficult to describe quantitatively, due to the large potential sequence space, the complexities of protein and nucleic acid folding, and unclear or unknown biophysical and physiological mechanisms. Therefore, many current quantitative methods require a densely sampled sequence space, a highly specific biophysical model of the process being studied, or sacrifice specificity for sufficient examples necessary for optimizing model parameters. In particular these limitations are seen in the computational method used to describe a protein's thermodynamic properties, where highly accurate models of protein folding free energy require a three-dimensional protein structure and a highly specific biophysical model (Yasuda *et al.*, 2017; Pucci *et al.*, 2016), extensive mutagenesis of the target protein family (Muk *et al.*, 2019), or are purely trained with many protein families and therefore

of limited specificity to any particular protein family (Montanucci *et al.*, 2008; Li and Fang, 2010; Gromiha and Suresh, 2008; Li *et al.*, 2019).

Fortunately for the study of protein stability, many protein families natural selection has already broadly sampled both sequence space and temperature space. Protein families often include homologs from 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 (Sauer *et al.*, 2015; Nguyen *et al.*, 2017; Perl *et al.*, 2000). Introducing corresponding mutations into model proteins often result in altered thermal stability or temperature dependent activity, reflecting the role of these amino acids in thermoadaptation. Therefore, the available homologus protein sequences and organismal growth temperatures provide a large dataset for analyzing temperature dependent protein properties, allowing novel methods of analysis.

Here we report a method to quantitatively predict a phenotype from a family of homologus sequences. The method, named protMLP, uses machine learning in the form of multilayer perceptron (MLP) models trained on individual family of proteins or nucleic acids. Notably, no assumptions are made about chemical, structural, epistatic, or thermodynamic effects, and a three-dimensional structure

is not used. We demonstrate the effectiveness of the method in predicting the originating organisms' growth temperatures (T_G) from protein sequences. Further, we show the correlation of predicted growth temperatures of the organism to the protein's experimentally determined melting temperature (T_M) or temperature of optimal activity (T_A). Thus, predicted organismal growth temperature ($\widehat{T}_G$) can serve as an easily calculable proxy for a protein's thermal stability and temperature dependent activity. Additionally, we demonstrate the generality of the protMLP method by predicting other quantitative phenotypes, such as binding affinity, fluorophore brightness, and catalytic activity, from the sequences of proteins and nucleic acids.

2 Methods

2.1 Sequence and organismal growth temperature collection

Domain sequence alignments were downloaded from the Pfam 32.0 database (Finn *et al.*, 2016) and used without modification unless otherwise noted. Reconstructed ancestral ADK sequences (Nguyen *et al.*, 2017) were combined with extant proteins from the Pfam alignment. CSP sequences identified in Pfam were extended at the C-terminus by one amino acid using the sequences in UniProtBK release 2018_04 (The UniProt Consortium, 2017). CSP and ADK sequences were then re-aligned in Promals3D (Pei *et al.*, 2008). Species' measured growth temperatures were collected from published sources (Sauer *et al.*, 2015; Engqvist, 2018) and BacDive (accessed August 14, 2019) (Söhngen *et al.*, 2016), averaging duplicate values of the same species. Growth temperatures

of uncharacterized species were predicted using previously published methods and proteomes (Li *et al.*, 2019). Proteins inducing gaps, annotated as fragments, or with residues outside the standard protein alphabet (ACDEFGHIKLMNPQRSTVWY-) were excluded.

2.2 Data division into training, validation, and test datasets

The method requires three datasets for training the regression model (training), preventing overfitting to the training dataset (validation), and final evaluation of model accuracy (test). Therefore, identical sequences were grouped and these groups were then randomly assigned into training (70%), validation (10%), or test (20%) datasets.

2.3 Sequence assignment, encoding, and balancing

For T_G regression, species assignments for each protein were collected from UniProtKB release 2018_04. Those sequences without a species assignment or without an originating species' T_G were excluded from further analysis. All CSP and ADK sequences with a characterized T_A and T_M were removed from the alignments; and used only for the comparison of T_G and $\widehat{T}_G$ to T_M and T_A.

All datasets were then one-hot encoded, with invariant positions of the one-hot training sequences removed from all alignments.

Training dataset were balanced by first calculating a histogram of training sequence phenotypes 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 bin until all bins had the same number of sequences as the most populous bin.

Phenotype-correlated positions were identified using the point-biserial correlation or a top-hat function fit to the relationship between phenotype and the one-hot encoded amino acid. If a threshold was provided, positions with a top-hat function Pearson correlation or point-biserial correlation less than the threshold were removed.

2.4 MLP training

The MLPs were trained using an identity or leaky ReLu activation function (Maas *et al.*, 2013). All regressions were trained using the Adam solver (Kingma and Ba, 2014), with the mean square error (MSE) as the loss function. Model checkpoints were saved at each epoch, if the current model validation dataset MSE was lower than the previous checkpoint. Training was stopped when the validation dataset MSE did not decrease for two consecutive training epochs.

2.5 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 and that the network be over-determined. Topologies were limited to between 1 and 5 hidden layers, in addition to the input and output layers (Fig. S1A). 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, mean squared error, and root mean square error were calculated and reported using the test dataset for the lowest validation MSE model.

2.6 Regression of deeply mutagenized sequences

Phenotypes and sequences or mutants were downloaded for WW domain (Fowler *et al.*, 2010; Esposito *et al.*, 2019), eqFP611 (Poelwijk *et al.*, 2019), guanine-inhibited ribozyme Lib-2 (Kobori *et al.*, 2017), and BRCA1 (Findlay *et al.*, 2014), with multiple sequence alignments generated *in silico*. For the guanine-inhibited ribozymes and BRCA1 RNA sequences, uracil (U) was replaced with thymine (T). The presence or absence of the regulating guanine in the guanine-inhibited ribozymes samples was encoded as an additional G or gap, respectively, at the 5' end of the sequence. Groups of identical sequences were assigned to training, validation, or test datasets as above. MLPs were then trained using the standard protMLP protocol, using the phenotypes as the target for regression.

All calculations used custom scripts written in Python with the Biopython (Cock *et al.*, 2009), Tensorflow (Abadi *et al.*, 2016), Keras (Chollet, 2015), NumPy (Oliphant, 2006), SciPy (Jones *et al.*, 2001), and Matplotlib (Hunter, 2007) libraries. Source code is available at: https://github.com/DavidBSauer/protMLP

3 Results

3.1 Construction of multi-layer perceptrons

Setting out, we aimed to devise a method to predict quantitative phenotypes from a protein or nucleic acid family's sequences. As a part of making this a general method, we also want to avoid an explicit structure or description of the forces underlying protein or nucleic acid folding or organismal physiology. 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 (Webb, 2018; LeCun *et al.*, 2015). Machine learning has been successfully applied to predicting a protein's fold from the sequence (AlQuraishi, 2019), the genotype of cancers from histopathology images (Coudray *et al.*, 2018), and the antimicrobial activity of a peptide sequence (Veltri *et al.*, 2018). Similarly, here we applied machine learning in the form of multilayer perceptrons (MLPs) to quantitatively predict the originating organism's growth temperature using protein 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 (Fig. S1A). 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. 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 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 (Hornik, 1991). 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 epistatic interactions. We therefore used MLPs with nodes that apply the non-linear, leaky rectifier activation function (rMLPs) (Fig. S1B).

As MLPs are mathematical models, the inputs are necessarily numerical. The inputs here are aligned amino acid or nucleic acid sequences from a particular homologus family. We therefore convert the aligned sequences to

sequences of Boolean variables (one-hot encoding) (Fig. S2), where one or zero indicates the respective presence or absence of a particular amino acid or nucleic acid. We further removed positions from all one-hot encoded sequences that were absolutely conserved in the one-hot encoded training sequences, as these would not contribute to the regression. Therefore, one-hot encoding preserves the chemical sequence of a biopolymer in a numerical sequence of ones and zeroes. Notably, one-hot encoding of the sequence does not contain a description of chemical or physical properties. This minimizes any assumptions of the relevant physical or chemical properties.

Regression methods, including machine learning, require the optimization of model parameters. Through “training”, the model weights and biases are iteratively refined using sequences with known regression target values, such as protein sequences and the associated originating organism’s optimal growth temperatures. However, it is essential to have mechanisms to avoid over-fitting and to independently evaluate accuracy (Zou *et al.*, 2019). Therefore, we used only 70% of the sequence-target pairs in the training process to refine the MLP weight and bias parameters. We used the remaining 30% of the sequence-target pairs for evaluating the regressions, assigning the pairs to validation (10%) and test (20%) datasets. Notably, sequences with 100% identity are placed in the same training, validation, or test datasets. Therefore the datasets have no sequences in common, while retaining the same distribution as the input sequence alignment. Only the training dataset is used for optimizing model

parameters. The validation dataset is used to avoid over-fitting by calculating the Mean Square Error (MSE) after each iteration of training, with training stopped when the MSE no longer decreases. Finally, the test dataset is used to calculate MLP accuracy. This allows for an evaluation of model generalizability, the accuracy of the model in predicting unseen sequences.

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 or nucleic acid family. Therefore, for each family we considered all possible MLP topologies with the restrictions that: the number of nodes in any hidden layer can range between two and twice the one-hot encoded sequence 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 (Vikhar, 2016), recombining and randomly permutating the lowest validation MSE topologies over multiple generations. Although, this method does not evenly sample the entire topology space, and may not find the optimal topology, this method is empirically time efficient in finding an optimized MLP topology for the prediction of T_G.

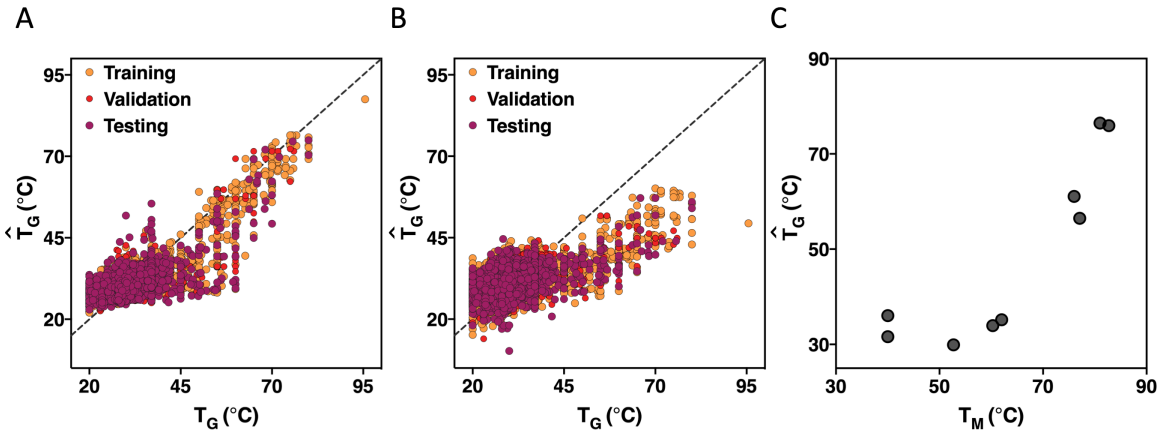

Figure 1. Organismal growth temperature can be predicted from the sequences of a protein family by using MLP regression. Regression of T_G using (A) the best rMLP or (B) linear regression model. The dotted line indicates perfect prediction. (C) Protein melting temperature versus predicted growth temperature of CSP homologs.

3.2 A trained MLP can predict organismal growth temperature from protein sequences

As an initial prototype for the prediction of a quantitative phenotype from a sequence, we examined the ability of a MLP to predict organismal growth temperature from protein sequences. Organismal growth temperature was selected as the regression target as T_Gs are measured for a large number of species (Sauer *et al.*, 2015). Similarly, we expected proteins to be a good input for such regression as many homologus protein sequences are known (Finn *et al.*, 2016), and, most importantly, often contain adaptations to particular thermal niches (Nguyen *et al.*, 2017; Perl *et al.*, 2000).

We initially studied the Cold Shock Protein (CSP) family. The CSP family's small protein size and strong conservation (Phadtare *et al.*, 1999) results in many available CSP 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. Homologus Cold Shock Protein sequences were collected from Pfam (Finn *et al.*, 2016), extended by one amino acid at the C-terminus, and aligned in Promals3D (Pei *et al.*, 2008). In total 34,254 homologus CSP sequences were identified with an available source organism growth temperature, with T_Gs measured 4.0 to 95.5 °C. All protein sequences were one-hot encoded and MLPs were trained using the described protMLP algorithm. Of the 10^{17} possible topologies, using the evolutionary algorithm 5,000 topologies were trained over 10 generations with 23,976 training sequences (70% of the total sequences). The 3,400 validation sequences (9.9% of the total sequences) were used to stop training, and to compare the accuracies of the various topologies. The trained MLP predicted the source organism growth temperature of 6,878 test sequences (20% of the total sequences) with a root mean squared error of 4.0 °C (Pearson correlation r = 0.70). Notably, this MLP clearly outperforms a linear regression trained with the same 23,976 training sequences (RMSE = 4.5 °C, r = 0.61), particularly in predicting T_G of proteins from thermophiles.

In examining the trained MLP results, we noted predicted T_G accuracy was poor for proteins from organisms with a growth temperature less than 20 °C. This is perhaps due to the rarity of these sequences, as they comprise only 1.2% of the training species-T_G pairs. 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. Finally, the mechanisms of adaptation from mesophilic to psychrophilic conditions may be distinct from the adaptations from thermophilc to mesophilic (Yang *et al.*, 2015). We therefore tested the accuracy of a MLP regression using only proteins from mesophiles and thermophiles (Fig. 1). Excluding protein sequences from organisms with a growth temperature less than 20 °C improved MLP accuracy (RMSE = 3.6 °C, r = 0.74) (Fig. 1A), again outperforming a linear regression (Fig. 1B). Therefore, we used only sequences from species with a T_G ≥20 °C in subsequent growth temperature studies.

3.3 Non-linearity is needed to predict growth temperatures

In examining the MLP topologies trained in the Cold Shock Protein regression, we found three distinct populations of model accuracy (Fig. S3A). A low accuracy population consists of topologies that converged to a single value (peak a). A second population is of topologies with accuracies similar to a linear regression (peak b). This set is unsurprising, as an rMLP can model a linear function. The final population of MLP models is more accurate than a linear regression (peak c), suggesting the rectified activation function is essential to regression accuracy.

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 MLPs using a rectified activation function (rMLPs), we trained MLPs of the same

topology with an identity activation function (iMLPs), where the activation function output is equal to the input. iMLPs 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. S3B). Notably, MLPs using rectified and identity activation functions have distinct distributions (Wilcoxon signed-rank test $p < 10^{-78}$). This confirms that the rectified activation function is essential to the improved prediction accuracy.

3.4 Predicted organismal growth temperature is correlated with experimentally determined melting temperatures of CSPs

We next explored if the rMLP predicted growth temperature would be of use in the biochemical characterization of proteins. Species' T_G is known to correlate with its proteins' thermal stability (Dehouck *et al.*, 2008) and temperature dependent activity (Engqvist, 2018). We therefore examined if the predicted organismal growth temperatures of CSP homologs were similarly correlated with measured protein melting temperatures (Fig. 1C). We found that cold shock protein homologs' predicted growth temperatures and measured melting temperatures to be directly correlated ($r = 0.87$) (Perl *et al.*, 2000; D'Auria *et al.*, 2010; Jin *et al.*, 2014; Lee *et al.*, 2013; Welker *et al.*, 1999; Phadtare *et al.*, 2003; Wassenberg *et al.*, 1999; Chatterjee *et al.*, 1993; Petrosian and Makhatadze,

2008; Phadtare, Inouye, *et al.*, 2002; Phadtare, Tyagi, *et al.*, 2002; Wunderlich *et al.*, 2005; Martin *et al.*, 2002; Garcia-Mira *et al.*, 2004; Mueller *et al.*, 2000). This is similar to the correlation between T_M and measured T_G for the same sequences (r = 0.85).

It is worth noting the evolutionary selection of proteins protecting from cold shock complicates the common theoretical relationship between T_M and T_G. In principle, to support organismal growth most proteins must be folded at the growth temperature. This imposes a theoretical bound on protein melting temperature such that $T_M > T_G$. Broad analyses of many proteins from several organisms support this theory (Dehouck *et al.*, 2008). However, cold shock temperatures are necessarily lower than organismal growth temperatures. As a consequence, CSP protein melting temperatures must only be greater than the cold shock temperature to support organismal growth. Therefore, the melting temperatures of Cold Shock Proteins could in principle be lower than the originating organism's growth temperature. However, we observe the CSP homolog's T_Ms are still greater than both rMLP predicted T_Gs and measured T_Gs of each of the originating species. This may indicate the temperature difference between organismal optimal growth and the physiological onset of CSP activity is generally small, or there are other functions of CSP homologs at the organisms' optimal growth temperatures (Keto-Timonen *et al.*, 2016).

3.5 Adenosine kinases' biochemical characteristics correlate with $\widehat{T}_G$

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), an extensively studied (Wolf-Watz *et al.*, 2004; Nguyen *et al.*, 2017; Schrank *et al.*, 2009) family that catalyzes the interconversion of adenosine nucleotides. As has been observed in other proteins (Engqvist, 2018), we found that the ADK homologs' temperatures of optimal activity (Nguyen *et al.*, 2017; Jeske *et al.*, 2019) were correlated with measured originating species' growth temperatures (r = 0.79). While there were too few ADK sequences to train an over-determined MLP, a linear regression was already highly accurate at predicting the originating species' growth temperature (RMSE = 3.9 °C, r = 0.86) (Fig. S4A). We found the $\widehat{T}_G$s for characterized ADK homologs and reconstructed ancestral sequences correlated with protein T_M (r=0.76) (Fig. S4B) and T_A (r = 0.63) (Fig. S4C), indicating that $\widehat{T}_G$ could serve as a proxy for T_A and T_M.

3.6 Thermophilic proteins have unique adaptations to temperature

In the CSP MLP regression, we noted that 98.5% of sequences are from mesophiles (Fig. S5A). This was unsurprising given the bias of the characterized organisms (Sauer *et al.*, 2015; Engqvist, 2018). While random sampling ensures the train, validation, and test datasets have the same distribution as the input

alignment, it was necessary to explore if the non-uniform distribution in phenotypes affected rMLP accuracy.

We first considered if sequences from mesophiles alone sufficiently sampled sequence space to accurately predict the T_G of homologs from thermophiles (Fig. S5B). 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 = 4.2 °C, r = 0.63), with a systematic under-prediction of proteins from thermophiles. Therefore, proteins from thermophiles likely contain amino acids at particular positions that are outside the sequence variation seen within CSPs from mesophiles.

We also investigated 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”. In balancing, a new training dataset was generated where rare thermophile sequences are over-sampled to be equal in proportion to the more common mesophiles (Fig. S5C). The validation and test datasets remained unchanged. Applying this new training dataset, the accuracy

of the rMLPs in predicting the unseen test dataset was slightly worse than without balancing (RMSE = 3.9 °C, r = 0.70) (Fig. S5D). As the number of unique sequences from thermophiles is much smaller than those from mesophiles, the oversampling of sequences from thermophiles may have led to over-fitting of inconsequential amino acids unique to these sequences.

Together, these results make clear that the relatively few (1.5%) sequences from thermophiles in the training dataset are necessary and sufficient for the prediction of optimal growth temperature of homologs from thermophiles. The non-uniform distribution of protein sequences and species T_Gs does not appear to have harmed regression accuracy, though accuracy may increase with more unique homologs from thermophiles having associated organismal growth temperatures.

3.7 Non-linear activation function MLPs are more accurate than linear regressions even with relatively few sequences

The ability of an 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, rMLPs and

iMLPs were trained as previously described (Fig. 2). It was not possible to build over-determined MLPs with 10% of the training sequences. However, the rectified activation function clearly outperformed an identity activation function at 20% of the training and validation sequences, or 4,761 and 686 sequences, respectively. This suggests that as few as 3.16 training sequences per one-hot encoded amino acid, or 24.9 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.

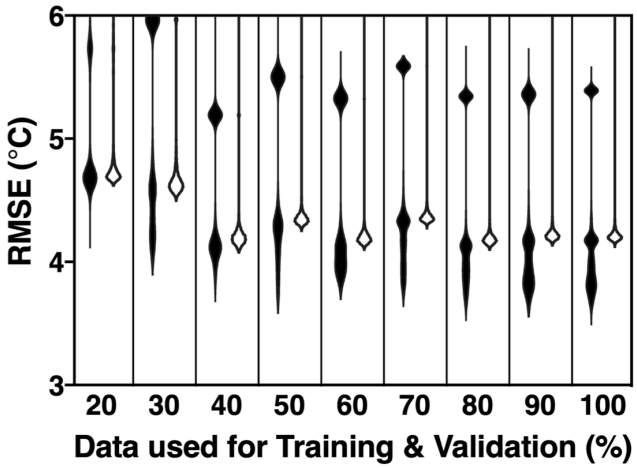

Figure 2. The proportion of non-linear MLP topologies outperforming equivalent topologies with a linear activation function increases with more training data. MLP accuracy trained using subsets of the training and validation sequences with either rectified (filled) or identity (unfilled) activation functions.

3.8 Particular amino acids are key to T_G prediction

When exploring possible MLP topologies, we only considered those topologies that were over-determined, with more training sequences than model

parameters. As the number of model parameters is dependent upon the one-hot encoded sequence length, we realized longer or less well conserved protein families may be precluded from analysis by the protMLP method. Fortunately, previous studies had indicated that only a small fraction of mutations to a protein's sequence alter protein stability (Abdul-Hussein *et al.*, 2013; Serrano-Vega *et al.*, 2008; Perl *et al.*, 2000). We therefore hypothesized that most 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 whether 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. S6A). Excluding those encoded amino acids with a correlation of less than 0.1 causes a loss in accuracy (RMSE = 4.0 °C, r = 0.68) while still outperforming a linear regression (RMSE = 4.8 °C, r = 0.47) using only 4.3% of the encoded protein sequence (Fig. S6B). We similarly used a fit top-hat function to identify amino acids with a second-order correlation to growth temperature (Fig. S6C). While only 26% of the encoded 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 3.7 °C using an rMLP (r = 0.73) (Fig. S6D).

These results confirm that only a subset of amino acids in the alignment of the CSP family are correlated with temperature and needed to predict the originating species' growth temperature. Therefore, using a similar point-biserial or top-hat correlation threshold would increase the effective data-to-parameter ratio, and allow for the over-determined regression of longer proteins. Alternatively, deeper and wider topologies could be used for shorter proteins, improving accuracy by accounting for more complex interactions in the sequence.

3.9 Non-linear MLPs in the regression of other protein families

We next examined if rMLPs could be used as a general method for predicting organismal growth temperature. Therefore, we trained MLP regression models to predict the originating species' T_G for other protein families (Table S1). These families included interaction and enzymatic domains, those targeted to various cellular localizations, and alpha-helical and beta-barrel membrane proteins. Of the 40 protein families examined, 25 had a sufficient number of sequences to train an over-determined MLP. The rMLP predicted growth temperature was correlated with measured growth temperature for all 25 families, and in each case outperformed a linear regression. Growth temperature predictions were consistent across all rMLP regressions with a test set species' T_G average standard deviation of 3.2 °C.

We also considered the effect of protein phylogeny on rMLP accuracy, particularly as the available sequences of a protein family may unevenly sample sequence space. We found only weak to no anti-correlation between maximum sequence identity and T_G prediction accuracy (r = -0.13 to -0.41). Therefore, the protMLP models are likely capturing features beyond sequence identity for the protein families.

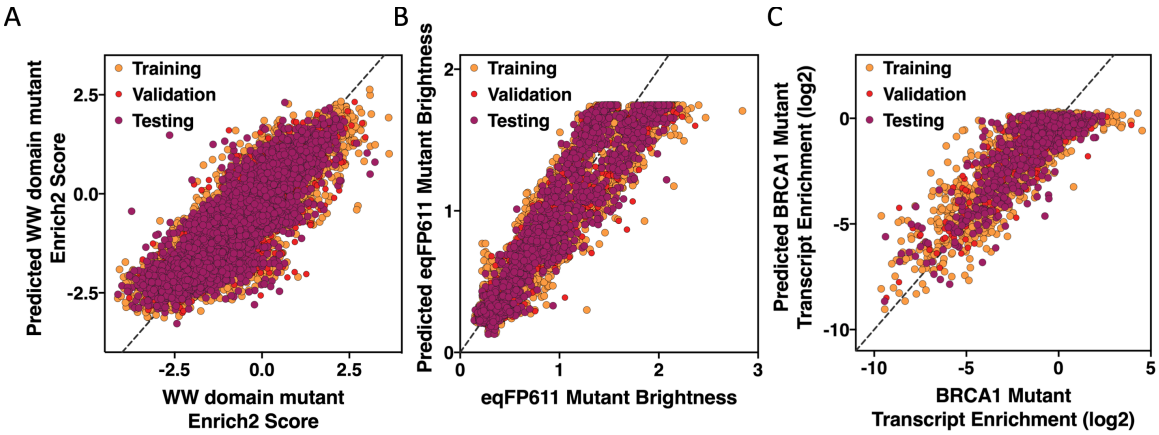

Figure 3. Non-linear MLPs can predict other phenotypes from protein and nucleic acid sequences. Predicting the (A) enrichment of WW domain mutants and (B) brightness of eqFP611 mutants and (C) transcript enrichment of mutant BRCA1 nucleic acid sequences using trained rMLP models. The dotted lines indicate perfect prediction.

3.10 MLPs can predict the phenotypes of proteins and nucleic acids

With the success of rMLPs to predict organismal growth temperature from a variety of protein families, we hypothesized that the same protMLP method could be applied for the regression of other quantitative phenotypic traits from the aligned sequences of other biopolymers.

We first examined the ability of protMLP to predict the phenotype of densely sampled sequence space by calculating regression on deeply mutagenized sequences of the human Yes Associated Protein 65 WW domain (Fowler *et al.*, 2010) and the fluorescent protein eqFP611 (Poelwijk *et al.*, 2019) (Fig. 3A and 3B). In both cases we found strong correlations between predictions and the measured phenotypes binding affinity (r = 0.856) and brightness (r = 0.962), respectively.

Noting that protMLP makes no assumptions of the physical or chemical characteristics of the individual monomers of the biopolymer, we next tested if nucleic acids can be analyzed by the same method. We therefore applied the protMLP method to mutant sequences of the BRCA1 gene (Findlay *et al.*, 2014) and an engineered guanine-inhibited ribozyme (Kobori *et al.*, 2017). protMLP model predictions strongly correlated with the measured transcript enrichment (r = 0.843) (Fig. 3C) or catalytic activity (r = 0.850) (Fig. S7B), respectively.

These results verify the generality of this method to predict a variety of quantitative phenotypes from proteins and nucleic acid sequences.

4 Discussion

The identification or design of biopolymers with particular biochemical or biophysical properties is often central to their study or for their use in industrial applications. However, the link between sequence and phenotype is often difficult to describe, due to the large potential sequence space and the difficulty of characterizing individual protein sequences.

Here, we successfully generated mathematical models to predict the various quantitative phenotypes from sequences. Growth temperatures could be predicted from protein sequences with a root mean squared error of 3.6 °C. These predicted T_Gs correlate with experimentally determined melting temperatures and temperatures of optimal activity. Similarly, binding affinity, fluorophore brightness, and ribozyme activity could be predicted from the sequences of biopolymers with strong correlation to the measured phenotypes. As phenotypically characterizing a sequence is typically tedious and time consuming, the protMLP method of predicting phenotype *in silico* from sequence offers significant advantages over experimental techniques.

4.1 Linear regressions are sufficient for some protein families

Some phenotypes can clearly be modeled as the linear combination of individual amino acid contributions, such as thermostability as seen in some membrane (Casim A. Sarkar *et al.*, 2008) and soluble proteins (Perl *et al.*, 2000), and prediction of T_G from ADK protein sequences. However, non-linear effects are clearly seen in the thermal stability of the Arc repressor (Brown and Sauer, 1999), the brightness of the eqFP611 fluorescent protein (Poelwijk *et al.*, 2019). Similarly, non-linear effects are seen here in the regression of organismal growth temperature from sequences. The varied success of linear regression models in predicting organismal growth temperature from sequence supports the hypothesis that the physical interactions that underlie thermoadaptation vary by

protein family (Petsko, 2001). As an rMLP can model a linear regression, in addition to more complex functions, the protMLP algorithm likely represents a general solution to describing the relationship between sequence and quantitative characteristics of the protein.

4.2 Protein families available for T_G analysis will increase

We recognize that construction of machine learning models for the regression of organismal growth temperatures is inherently limited by the number of sequence - T_G pairs available. For Cold Shock Proteins, 24.9 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 (Wang and Sauer, 2019; Li *et al.*, 2019), providing T_Gs for homologus proteins from species whose growth temperatures have not been experimentally determined. Using species' growth temperatures predicted by the Li *et al.* method

increased the number of training and validation sequences with an assigned T_G by 15%. Using these additional sequences and proteome predicted T_Gs in the protMLP method improved accuracy slightly (RMSE = 3.6 °C, r = 0.75) (Fig. S8).

4.3 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 single and double mutants of a CSP ortholog from *Bacillus subtillis* (BsCSP) (Fig. S9A). We found no correlation between mutant protein melting temperatures (Wunderlich *et al.*, 2005; Garcia-Mira *et al.*, 2004) and predicted organismal optimal growth temperatures calculated from the mutant proteins' sequences (r = -0.21). 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. S9B). This is in contrast to BsCSP mutants, with only one or two amino acid changes, exhibiting significantly altered melting temperature from wild type (Perl *et al.*, 2000; Wunderlich *et al.*, 2005; Garcia-Mira *et al.*, 2004). We therefore suspect that the available CSP sequences do not sufficiently sample sequence space to capture the effects of few or rare sequence differences. This hypothesis is supported by the accuracy of protMLP in predicting binding affinity of a deeply mutagenized WW domain (Fig. 3A). This dataset consists of only single, double, and triple

mutants, corresponding to 91-97% sequence identity. Therefore, this result demonstrates that rMLPs can accurately predict the effects of few mutations with sufficiently sampled sequence space.

4.4 Application to describing organismal phenotypes

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 (Wang and Sauer, 2019; Li *et al.*, 2019), protMLP calculates $\widehat{T}_G$ 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. It is possible to describe the thermal niche of no longer extant organisms by analyzing their reconstructed sequences using protMLP. Though the ecological niches of ancestral organisms have been inferred by the experimental characterization of reconstructed proteins' melting temperatures (Nguyen *et al.*, 2017), by predicting organismal growth temperature *in silico*, protMLP is faster and likely more accurate.

Acknowledgements

The authors thank David Fenyo for helpful discussion of this work, and Jennifer Marden for critical review of this manuscript.

Funding

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 Defence 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 Defence.

Conflict of Interest: none declared.

Abadi,M. *et al.* (2016) TensorFlow: A System for Large-Scale Machine Learning. In, *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*. USENIX Association, Savannah, GA, pp. 265–283.

Abdul-Hussein,S. *et al.* (2013) Thermostabilisation of the serotonin transporter in a cocaine-bound conformation. *J. Mol. Biol.*, **425**, 2198–2207.

Abrahamson,M. and Grubb,A. (1994) 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.*, **91**, 1416–1420.

AlQuraishi,M. (2019) AlphaFold at CASP13. *Bioinformatics*, btz422.

Brown,B.M. and Sauer,R.T. (1999) Tolerance of Arc repressor to multiple-alanine substitutions. *Proc. Natl. Acad. Sci.*, **96**, 1983–1988.

Chatterjee,S. *et al.* (1993) The Backbone Structure of the Major Cold-Shock Protein CS7.4 of Escherichia coli in Solution Includes Extensive β-Sheet Structure1. *J. Biochem. (Tokyo)*, **114**, 663–669.

Chollet,F. (2015) Keras. *GitHub*.

Cock,P.J.A. *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinforma. Oxf. Engl.*, **25**, 1422–1423.

Coudray,N. *et al.* (2018) Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nat. Med.*, **24**, 1559–1567.

D'Auria,G. *et al.* (2010) Dynamical properties of cold shock protein A from Mycobacterium tuberculosis. *Biochem. Biophys. Res. Commun.*, **402**, 693–698.

Dehouck,Y. *et al.* (2008) Revisiting the correlation between proteins' thermoresistance and organisms' thermophilicity. *Protein Eng. Des. Sel.*, **21**, 275–278.

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**.

Esposito,D. *et al.* (2019) An open-source platform to distribute and interpret data from multiplexed assays of variant effect. *bioRxiv*.

Findlay,G.M. *et al.* (2014) Saturation editing of genomic regions by multiplex homology-directed repair. *Nature*, **513**, 120–123.

Finn,R.D. *et al.* (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.*, **44**, D279-285.

Fowler,D.M. *et al.* (2010) High-resolution mapping of protein sequence-function relationships. *Nat. Methods*, **7**, 741–746.

Garcia-Mira,M.M. *et al.* (2004) The Folding Transition State of the Cold Shock Protein is Strongly Polarized. *J. Mol. Biol.*, **339**, 555–569.

Gromiha,M.M. and Suresh,M.X. (2008) Discrimination of mesophilic and thermophilic proteins using machine learning algorithms. *Proteins*, **70**, 1274–1279.

Hornik,K. (1991) Approximation capabilities of multilayer feedforward networks. *Neural Netw.*, **4**, 251–257.

Hunter,J.D. (2007) Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.*, **9**, 90–95.

Jeske,L. *et al.* (2019) BRENDA in 2019: a European ELIXIR core data resource. *Nucleic Acids Res.*, **47**, D542–D549.

Jin,B. *et al.* (2014) Structure and flexibility of the thermophilic cold-shock protein of Thermus aquaticus. *Biochem. Biophys. Res. Commun.*, **451**, 402–407.

Jones,E. *et al.* (2001) SciPy: Open source scientific tools for Python.

Keto-Timonen,R. *et al.* (2016) Cold Shock Proteins: A Minireview with Special Emphasis on Csp-family of Enteropathogenic Yersinia. *Front. Microbiol.*, **7**.

Kingma,D.P. and Ba,J. (2014) Adam: A Method for Stochastic Optimization.

Kobori,S. *et al.* (2017) Deep Sequencing Analysis of Aptazyme Variants Based on a Pistol Ribozyme. *ACS Synth. Biol.*, **6**, 1283–1288.

LeCun,Y. *et al.* (2015) Deep learning. *Nature*, **521**, 436–444.

Lee,J. *et al.* (2013) Structural and dynamic features of cold-shock proteins of Listeria monocytogenes, a psychrophilic bacterium. *Biochemistry*, **52**, 2492–2504.

Li,G. *et al.* (2019) Machine Learning Applied to Predicting Microorganism Growth Temperatures and Enzyme Catalytic Optima. *ACS Synth. Biol.*, **8**, 1411–1420.

Li,Y. and Fang,J. (2010) Distance-dependent statistical potentials for discriminating thermophilic and mesophilic proteins. *Biochem. Biophys. Res. Commun.*, **396**, 736–741.

Maas,A.L. *et al.* (2013) Rectifier nonlinearities improve neural network acoustic models. In, *in ICML Workshop on Deep Learning for Audio, Speech and Language Processing*.

Martin,A. *et al.* (2002) Origins of the High Stability of an in vitro-selected Cold-shock Protein. *J. Mol. Biol.*, **318**, 1341–1349.

Matreyek,K.A. *et al.* (2018) Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nat. Genet.*, **50**, 874–882.

Montanucci,L. *et al.* (2008) Predicting protein thermostability changes from sequence upon multiple mutations. *Bioinforma. Oxf. Engl.*, **24**, i190-195.

Mueller,U. *et al.* (2000) Thermal stability and atomic-resolution crystal structure of the Bacillus caldolyticus cold shock protein. *J. Mol. Biol.*, **297**, 975–988.

Muk,S. *et al.* (2019) Machine Learning for Prioritization of Thermostabilizing Mutations for G-protein Coupled Receptors. *bioRxiv*.

Nguyen,V. *et al.* (2017) Evolutionary drivers of thermoadaptation in enzyme catalysis. *Science*, **355**, 289–294.

Oliphant,T. (2006) NumPy: A guide to NumPy.

Pei,J. *et al.* (2008) PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic Acids Res.*, **36**, 2295–2300.

Perl,D. *et al.* (2000) Two exposed amino acid residues confer thermostability on a cold shock protein. *Nat. Struct. Biol.*, **7**, 380–383.

Petrosian,S.A. and Makhatadze,G.I. (2008) Contribution of proton linkage to the thermodynamic stability of the major cold-shock protein of Escherichia coli CspA. *Protein Sci.*, **9**, 387–394.

Petsko,G.A. (2001) Structural basis of thermostability in hyperthermophilic proteins, or 'there's more than one way to skin a cat'. *Methods Enzymol.*, **334**, 469–478.

Phadtare,S. *et al.* (1999) Cold-shock response and cold-shock proteins. *Curr. Opin. Microbiol.*, **2**, 175–180.

Phadtare,S. *et al.* (2003) CspB and CspL, thermostable cold-shock proteins from Thermotoga maritima. *Genes Cells Devoted Mol. Cell. Mech.*, **8**, 801–810.

Phadtare,S., Inouye,M., *et al.* (2002) The Nucleic Acid Melting Activity of *Escherichia coli* CspE Is Critical for Transcription Antitermination and Cold Acclimation of Cells. *J. Biol. Chem.*, **277**, 7239–7245.

Phadtare,S., Tyagi,S., *et al.* (2002) 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.*, **277**, 46706–46711.

Poelwijk,F.J. *et al.* (2019) Learning the pattern of epistasis linking genotype and phenotype in a protein. *Nat. Commun.*, **10**, 4213.

Pucci,F. *et al.* (2016) Predicting protein thermal stability changes upon point mutations using statistical potentials: Introducing HoTMuSiC. *Sci. Rep.*, **6**, 23257.

Raimondo,A. *et al.* (2014) Phenotypic severity of homozygous GCK mutations causing neonatal or childhood-onset diabetes is primarily mediated through effects on protein stability. *Hum. Mol. Genet.*, **23**, 6432–6440.

Sarkar,C. A. *et al.* (2008) Directed evolution of a G protein-coupled receptor for expression, stability, and binding selectivity. *Proc. Natl. Acad. Sci.*, **105**, 14808–14813.

Sarkar,Casim A. *et al.* (2008) Directed evolution of a G protein-coupled receptor for expression, stability, and binding selectivity. *Proc. Natl. Acad. Sci. U. S. A.*, **105**, 14808–14813.

Sauer,D.B. *et al.* (2015) Rapid Bioinformatic Identification of Thermostabilizing Mutations. *Biophys. J.*, **109**, 1420–1428.

Schrank,T.P. *et al.* (2009) Rational modulation of conformational fluctuations in adenylate kinase reveals a local unfolding mechanism for allostery and functional adaptation in proteins. *Proc. Natl. Acad. Sci.*, **106**, 16984–16989.

Serrano-Vega,M.J. *et al.* (2008) Conformational thermostabilization of the beta1-adrenergic receptor in a detergent-resistant form. *Proc. Natl. Acad. Sci. U. S. A.*, **105**, 877–882.

Söhngen,C. *et al.* (2016) BacDive--The Bacterial Diversity Metadatabase in 2016. *Nucleic Acids Res.*, **44**, D581-585.

Tate,C.G. (2012) A crystal clear solution for determining G-protein-coupled receptor structures. *Trends Biochem. Sci.*, **37**, 343–352.

The UniProt Consortium (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **45**, D158–D169.

Veltri,D. *et al.* (2018) Deep learning improves antimicrobial peptide recognition. *Bioinformatics*, **34**, 2740–2747.

Vikhar,P.A. (2016) Evolutionary algorithms: A critical review and its future prospects. In, *2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC)*. IEEE, Jalgaon, India, pp. 261–265.

Wang,D.-N. and Sauer,D.B. (2019) Predicting the optimal growth temperatures of prokaryotes using only genome derived features.

Wassenberg,D. *et al.* (1999) Thermodynamics of the unfolding of the cold-shock protein from Thermotoga maritima 1 1Edited by A. R. Fersht. *J. Mol. Biol.*, **289**, 187–193.

Webb,S. (2018) Deep learning for biology. *Nature*, **554**, 555–557.

Welker,C. *et al.* (1999) Cloning, overexpression, purification, and physicochemical characterization of a cold shock protein homolog from the hyperthermophilic bacterium Thermotoga maritima. *Protein Sci. Publ. Protein Soc.*, **8**, 394–403.

Wolf-Watz,M. *et al.* (2004) Linkage between dynamics and catalysis in a thermophilic-mesophilic enzyme pair. *Nat. Struct. Mol. Biol.*, **11**, 945–949.

Wunderlich,M. *et al.* (2005) Stabilization of the Cold Shock Protein CspB from Bacillus subtilis by Evolutionary Optimization of Coulombic Interactions. *J. Mol. Biol.*, **347**, 1063–1076.

Yang,L.-L. *et al.* (2015) Low Temperature Adaptation Is Not the Opposite Process of High Temperature Adaptation in Terms of Changes in Amino Acid Composition. *Genome Biol. Evol.*, **7**, 3426–3433.

Yasuda,S. *et al.* (2017) 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*, **121**, 6341–6350.

Zou,J. *et al.* (2019) A primer on deep learning in genomics. *Nat. Genet.*, **51**, 12–18.