Decoding Protein Stability - a Data-Driven Approach for Predictive Modeling of Enhanced Thermostability

Preet Shah, Maximilian Vassen, Marik Müller, Tobias Kreusel

University Heidelberg, Molecular Biotechnology

Seminar Application of Bioinformatical Methods

Maximilian Fidlin, Benedict Wolf, Dr. Jan Mathony, Prof. Dr. Dominik Niopek

08.07.2024

Abstract

Extending protein thermostability has significant applications in research and industry, such as extending shelf life, optimising enzyme catalysis, or improving growth of organisms at higher temperatures. This project utilised an extensive thermal proteome profiling dataset of mesophile and thermophile organisms to comprehensively analyse how various factors and structural motifs of proteins influence thermostability and to identify parameters for improving thermal stability. These findings were then used in a regression analysis, yielding a model that accurately predicts protein thermostability from protein sequence ($r^2 = 0.72$). Using this model, essential proteins that presumably act as a bottleneck, restricting the maximum growth temperature of an organism, were identified and classified according to their function. In a second step, a mutational screen was applied to these proteins, predicting beneficial mutations for increased thermostability with minimal structural changes. To facilitate automated use, a user friendly python class was developed combining SPARC (**S**equence-based **P**rotein **A**ttribute-de**R**ived (melting point) **C**alculator) and ThERMOS (**Th**ermal **E**nhancement by **R**apid **M**utational **O**ptimization **S**creen) as the main features to predict mutations enhancing thermostability while maintaining structural integrity.

Table of contents

Introduction

Protein thermostability has many implications ranging from improved protein stability to faster enzyme catalysis or the possibility to use proteins in high temperature processes such as 3D printing. Evolutionary approaches are often hampered by the inability of organisms to endure higher temperatures during the selection process and are also time consuming (Speck et al., 2012), making an in-silico approach as presented here even more relevant. This project was based on data from a publication that identified the melting temperatures (T_m) of 48,000 proteins across 13 species (Jarzab et al., 2020), in which Thermal Proteome Profiling (TPP) was used to identify the T_m in whole cells and cell lysates. TPP, a method initially established to study protein-drug interactions in the context of cancer drugs (Savitski et al., 2014), assumes protein unfolding leading to precipitation as an indicator for thermal stability. Accordingly, proteins are heated, the precipitate removed by ultracentrifugation, digested with trypsin, marked by stable isotopes, subjected to LC-MS/MS and quantified. The temperature, at which 50 % of the proteins were precipitated, was defined as the melting point of each protein. Jarzab et al. have already established the use of TPP as a method useful in drug discovery, but have also highlighted that their investigation will be useful for further studies.

While the authors have shown high intra- and interlab accuracy, there are some method-based shortcomings that should be considered for any results generated from the data set. For one, the method uses protein precipitation as an indicator of protein denaturation. Additionally, the cell-lysate data analysed proteins outside their native environment, which could result in the loss of significant factors that normally stabilise or destabilise the protein, thus falsifying the results. To minimise such errors, the presented work focuses on prokaryotic proteins with melting point measurements from lysate samples. Lysate data were chosen over whole cells, due to a larger number of unique proteins and organisms in the data set as well as to investigate structural factors and motifs of thermostability without influence from cellular factors.

Previously, various factors affecting the temperature stability of a protein have been identified. It has been reported that a high content of hydrophobic amino acids significantly affects thermostabilty by preventing denaturation through so-called hydrophobic cores (Vieille & Zeikus, 2001).

On the level of secondary structure, it was reported that protein rigidity also contributes to a higher melting point. In this regard, more and longer α-helices as well as β-sheets positively contribute to thermostability (Vihinen, 1987). However, this theory of rigidity versus flexibility has not yet been fully answered, and there are indicators that both rigidity and flexibility play important roles in protein thermostability (Karshikoff et al., 2015).

Lastly, tertiary and quaternary structure also play an essential role not only in thermal stability but also protein function. Hydrophobic patches, disulfide bonds, hydrogen bonds and protein packing were shown to contribute significantly to thermostability (Kumar et al., 2001; Niu et al., 2016; Vieille & Zeikus, 2001).

Beyond factors that arise directly from protein structure, there are also further contributors to protein thermostability, such as post-translational-modifications (Johnson, 2009; Shental-Bechor & Levy, 2008). These factors could not, or were not determined in above mentioned studies.

The central aim of the presented project was to provide a comprehensive study of factors affecting the melting temperature of proteins to an extent that has not been investigated previously. Accordingly, next to considering a range of factors previously determined to have significant impact on thermostability, we included new structural factors and motifs. Based on this analysis, which factored in over 300 structural factors of thermostability, a model was built to predict the melting point of a protein based on the amino acid sequence. Importantly our model predicted melting points with comparable or better accuracy and speed than previously described tools (Abramson et al., 2024; Baek et al., 2021; Jumper et al., 2021). Further, we designed an *in-silico* mutagenesis tool to predict beneficial mutations for improved protein thermostabilty, with minimal structural changes.

Additionally, this project identified key proteins and pathways in organisms that likely are the limiting factor in their temperature tolerance. This was done by looking at individual proteins in the entire modified data set, as well as analysing gene ontology groups in order to find common biological processes that are key factors in the thermo-sensitivity of the respective organisms.

Combining the outcomes of these analyses, has tremendous potential in identifying key components for raising temperature tolerance and optimal growth temperatures of whole organisms, with relevant applications in industry and research.

4 Materials and Methods

Dataset handling and curation was done with pandas and all dimension reduction and regression calculations were performed using the SciPy (scipy.stats) model, while data visualisation was done with the seaborn and matplotlib packages (Hunter, 2007; McKinney, 2010; Virtanen et al., 2020; Waskom, 2021). For coding related tasks ChatGPT-4o was used(OpenAI, 2024)

4.1 Data set construction

The initial dataset contains the measured fold changes at different temperatures for 34,000 proteins in various prokaryotic cells and lysates as well as in several human and mouse cell lines (Jarzab et al., 2020). The dataset consists of the experimentally determined melting point (Tm) of each protein based on their precipitation midpoint as well as Protein IDs associated with the UniProt database, the protein length and its amino acid sequence. The initial data was extended by incorporating further data provided Gene Ontology classification. Additional information was acquired from the UniProt (The UniProt Consortium, 2023) database, namely extended Gene Ontology classifications, AlphaFold and PDB IDs.

4.2 Primary and secondary structure analysis

Firstly, the dataset was filtered to only encompass data for unique proteins in prokaryotic organisms from measurements of cell lysates, which ultimately provided a total of 6,500 samples. For these proteins sequence information, such as the relative contents of each 20 canonical amino acids and their pairwise combinations, as well as the relative content of hydrophobic, polar, charged amino acids, were calculated (Table S1). This was then further enriched with structural features using the public software S4pred (Moffat & Jones, 2021), a tool, which uses machine learning algorithms and evolutionary information derived from multiple sequence alignments to predict protein secondary structures (mainly α-helices and β-sheets)from the primary sequence.

4.3 Tertiary structure analysis

Tertiary structure analysis was performed for all proteins with either available crystal structures or by using the AlphaFold prediction algorithm (Jumper et al., 2021).

Salt bridges were defined as an interaction between an acidic carboxyl-group in the side chain of amino acids glutamate and aspartate (atoms OD and OE respectively) and a basic nitrogenous group of arginine, lysine or histidine (NH, NZ and NE or ND respectively). An atom distance matrix was calculated between all potential salt bridge forming atoms, and salt bridge formation was assumed if the distance was within 4.0 Angstrom (Ferruz et al., 2021).

Hydrogen bond networks were derived from predicted hydrogen atom positions of amino acid side chains at pH 7 with the help of molecular force field calculations using AMBER as well as the PDB2PQR package (Dolinsky et al., 2004). The presence of hydrogen bonds was then defined for an angle $100° < \vartheta < 180°$ between donor, acceptor and hydrogen atom (Table S2) and a distance d < 3.5 Å between the acceptor and donor atom (Tan et al., 2021).

Hydrophobic clusters were predicted within a protein, if the van der Waals radii, enlarged by a water molecule radius (1.4 Å), were intersecting each other (Bondi, 1964; Ferruz et al., 2021). Overlap volume was calculated between each pair using the equation 1, where r_2 and r_1 are the enlarged atom radii, and d the distance between both atom centres. Hydrophobic clusters are defined for all atoms that are connected by these interactions.

$$V_{intersect} = \frac{\pi \, (r_2 + r_1 - d)^2 \times (d^2 + 2dr_1 - 3r_1^2 + 2dr_2 + 6 \, r_1 r_2 - 3r_2^2)}{12d}$$ ***Equation 1***

Solvent accessible surface area (SASA) was used to calculate the protein surface size. The calculation was done with the Biopython package (Cock et al., 2009). Shortly, the protein is probed with a ball with a radius of 1.4 Å (representing a solvent molecule), and the accessible surface is calculated (Cock et al., 2009; Shrake & Rupley, 1973).

4.4 PCA and regression analysis

All aforementioned thermostability-influencing features were filtered based on a significant Pearson's correlation with regards to T_m higher than 0.2, whereas the significance of the correlation was computed using the standard Pearson's product moment correlation with a significance value α = 0.05. This resulted in 300 descriptive variables, which were converted to principal components (PCs), where the first 25 PCs were kept. Following this, a PCA biplot was constructed to ascertain novel combinations of features that positively contribute to the thermostability of proteins.

The above-mentioned PCs were utilised as descriptive variables in a Gradient Boosting Regression model to predict the T_m of other proteins, Gradient Boosting Regression is a machine-learning technique that creates an ensemble of initially weak models, typically decision trees, in a sequential manner; here each new model is generated by focusing on reducing the residual errors of the combined ensemble from previous iterations, which is done by adjusting the model parameters in the direction that most reduces the overall error, as computed by the gradient of the loss function. To substantiate this model's significance, it was compared to a standard linear regression model using an F-test.

4.5 Essential proteins

For each protein from the curated lysate dataset, the temperature value T_{90}, where 90% of the protein precipitated, was calculated according to the melting curves obtained from the raw data, For this the algebraic equation was solved (Equation 2), where the parameters a, b and plateau for each protein were incorporated in the extracted melting curve data.

$$f(T) = \frac{1 - plateau}{1 + e^{-(\frac{a}{T} - b)}} + plateau \qquad \textbf{\textit{Equation 2}}$$

Essential proteins, for each prokaryotic species were consequently defined as those proteins, where the calculated T_{90} value was in a range of 1.5 °C around the maximal growth temperature T_{max} for that species, which was acquired from published data (Leuenberger et al., 2017; Ohtani et al., 2010; Šimunović et al., 2022; Yakimov et al., 2003). Furthermore, the identified proteins were grouped based on their gene ontology, and further analysed to deduce important biological processes they are involved in.

4.6 Protein mutation screen

For *in silico* mutational screening, amino acids were divided into mutable and non-mutable amino acids depending on their involvement in tertiary structures, whereas those involved in secondary structures are only allowed to mutate into structure maintaining amino acids (Wang & Jardetzky, 2002). Based on these criteria, a list of possible substitutions for each amino acid has been taken from (Pechmann & Frydman, 2014) (Table S3) and adjusted to account for the non-mutable amino acids. Characteristics for the strongest positively and negatively correlating features were calculated for the top 10th percentile and considered as ideal values. The effectiveness of the substitution was based on the deviation from these ideal values.

During the mutation process, first random substitutions were generated within the given constraints to create a larger starting pool, from which the top ten scoring and ten most deviating sequences were used for further rational screening.

Rational screening iterated through each mutable position and attempted to optimise the strongest deviating feature. This was done for a user-defined number of cycles or until the top deviating feature remained constant, after which the five best performing variants were passed on. In the final step, the predicted melting point based on our regression model (SPARC) was calculated and the variant with the strongest increase selected as the final candidate. The results were tested for normal distribution using the scipy.stats.normaltest based on D'Agostino and Pearson (R. B. D'Agostino, 1971; R. D'Agostino & Pearson, 1973; Virtanen et al., 2020).

The results of the mutation screen were analysed by predicting the 3D structure of the wild-type and mutant protein using AlphaFold 3 (Abramson et al., 2024) and using PyMOL (Schrödinger, LLC, 2015) to analyse structural deviations and possible effects of the mutations.

5 Results

5.1 General correlations

For the aforementioned primary, secondary and tertiary structure elements of all proteins, the Pearson correlation with melting points was calculated. The 25 features with the highest and lowest correlations respectively are shown in Fig. 1
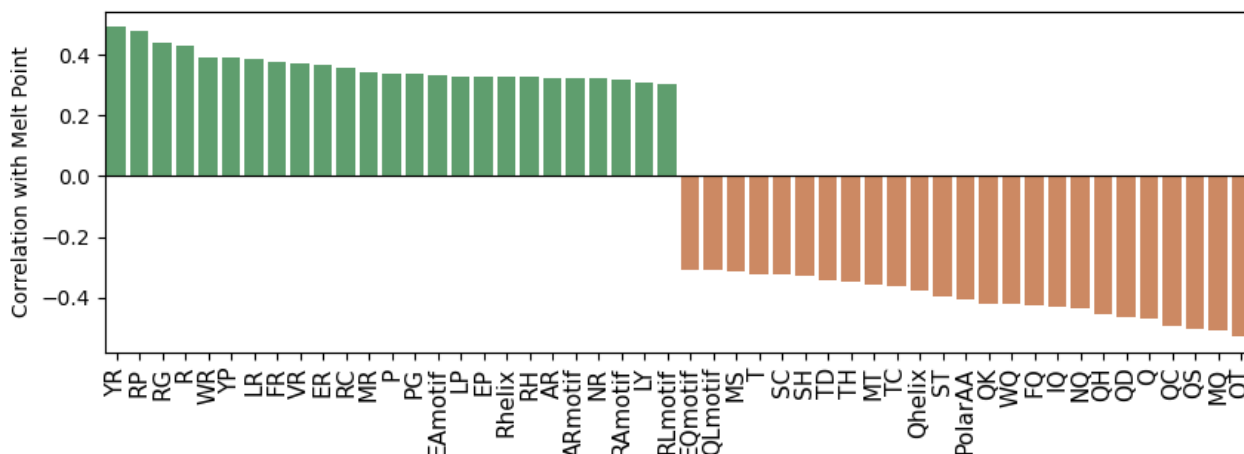

Figure 1: Top 25 features with the highest and lowest correlation with melting point.
The Pearson correlation was calculated for all primary, secondary and tertiary structure elements. The features with the strongest positive and negative correlations are depicted. Features with two amino acids describe the relative combined content, motif describes the relative content of the respective amino acids, helix describes the relative content of the amino acid within all helices of a protein, and the "PolarAA" describes the relative content of polar amino acids within a protein. All correlations were tested using Pearson's product moment correlation test, with p < 0.05.

This analysis revealed that the feature with the highest correlation was the combined relative tyrosine and arginine content (correlation of 0.49), and many other features with high correlations included arginine as well, indicating its importance for thermostability. Another important feature identified the "EAmotif" (correlation of 0.33) which describes the amount of glutamic acid and alanine dipeptides in protein sequences. Notably the highest correlating secondary structure element was the "Rhelix" (correlation of 0.33) which is the arginine content within all α-helices.

The feature with the strongest negative correlation was the relative glutamine and threonine content (correlation of -0.52). Another important observation was that the relative content of polar amino acids had a high negative correlation of -0.41. Notably, the twelve features with the highest negative correlation all included glutamine. Other prominent amino acids impacting T_m were serine, threonine and cysteine. Overall primary and secondary structure features showed stronger correlations with melting points, compared to tertiary structure elements, with the most notable being the relative hydrophobic cluster length with a correlation of 0.12 and the protein surface area with -0.11.

5.2 PCA analysis

After computation of the PCA, PC1 was able to explain 15% and PC2 5% of the total variance explained by the 300 features. To now corroborate new combinations of thermostability-influencing features, the following PCA biplot of PC2 plotted against PC1 was constructed, where proteins with high, medium and low stability were highlighted (Fig. 2).
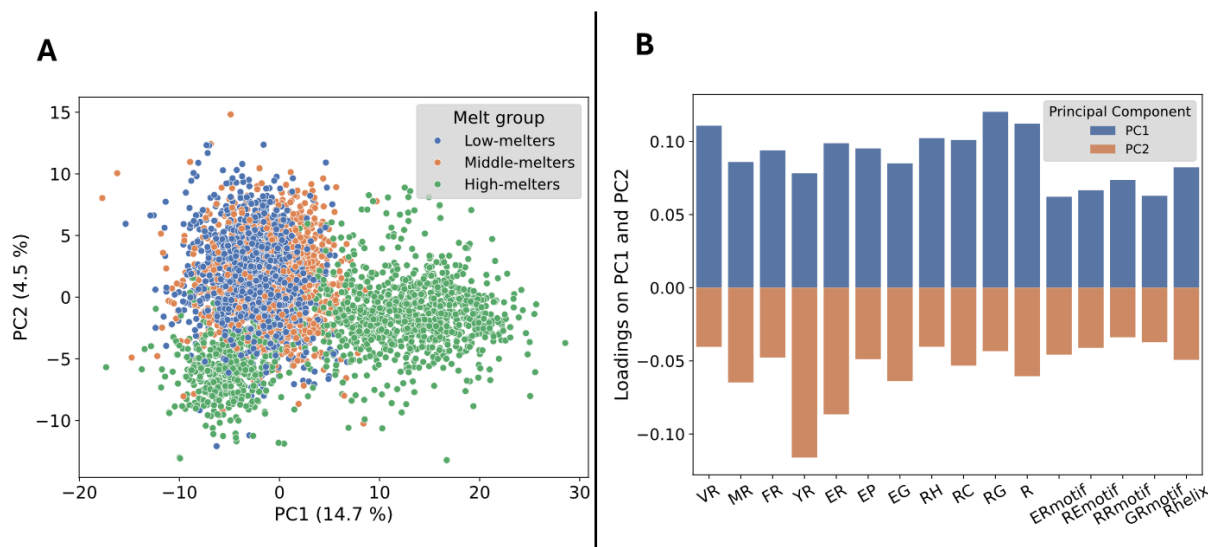
Figure 2: PCA biplot of PC1 and PC2 (A). Features with positive PC1 and negative PC2 loadings (B).
(A) Absolute values of PC2 were plotted against PC1 for proteins belonging to the following melt groups: proteins with high thermostability (top 20% of T_m values) in green, with medium thermostability (40-60%) in orange and low thermostability (bottom 20 %) in blue; (B) The value of the loadings for features positively contributing to PC1 (blue) and negatively contributing to PC2 (orange) are depicted in a bar plot.

Fig. 2A visually elucidates the formed clusters of proteins with high, medium, and low thermostability, namely those proteins that display positive values of PC1 (between 10 and 20) and negative values of PC2 (between -5 and -10). As a consequence, proteins having high melting points must comprise features that have a positive contribution to PC1 and a negative contribution to PC2: Analysis of the respective loadings of both PCs revealed the following set of features, which together are responsible for higher thermostability of proteins (Figure 2B). Here it becomes apparent that this list predominantly consists of arginine-associated features, like relative arginine content in the protein and relative arginine content in α-helices, suggesting that arginine-rich proteins generally have a higher thermostability.

The first 25 PCs, which altogether explain 56% of the total variance, were implemented in the following regression analysis.

5.3 Regression analysis

Several commonly used regression models, including linear, ridge, lasso, random forest and gradient boosting regression were trained and tested. For this, 80% of the PCA transformed dataset were randomly selected for the training set while the remaining 20% were used for the test set (a seed of 1 was used for random selection to ensure reproducibility). The performance of the different models were compared by their mean square error and their r^2. The random forest and gradient boosting regression performed best, however the gradient boosting regressor model was used in the end as it performed better, especially for higher melting points above 62 °C.

The final model revealed an r^2-value of 0.72, a root mean square error of 8.3 °C and a pearson correlation coefficient of 0.85, calculated using 10 different random dataset splits as training/test split. To confirm the significance of the model, an f-test was performed. As the f-test is not directly applicable to more complex models such as the gradient boosting regression, a linear model was fitted to the training set first. A f-statistic of 382.66 and a p-value of 1.1×10^{-6} were calculated indicating its significance. The gradient boosting model was significantly better, yielding an f-statistic of 45.8 and a p-value of 1.1×10^{-6}, and thus confirming the legitimacy of the model.

The predicted temperature dependent root mean square error (RMSE) of the model was further investigated. For this, proteins were binned by their predicted melting points into groups of 3 °C from 38 to 92 °C. Fig. 3 shows the RMSE for each category and the number of proteins in each bin. A sixth degree polynomial function was fitted onto the RMSE values, resulting in a function that could be used to approximate the error of the predicted melting point.
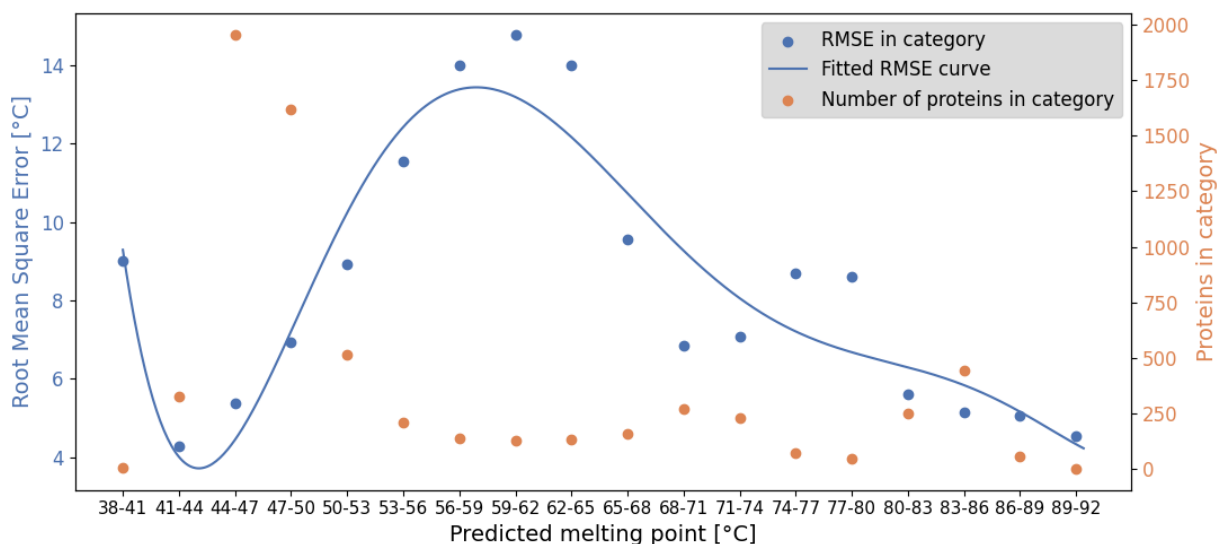
Figure 3: Temperature dependent RMSE of predicted melting points and number of proteins.
Proteins were grouped by dividing into 3 °C bins and the RMSE of each group was calculated (blue dots). A sextic polynomial function was fitted onto the binned Tm (blue curve). Orange dots show the number of proteins in each melting point group.

5.4 Essential proteins

Essential proteins are defined as proteins which exhibit 90 % denaturation at the maximum growth temperature ($\pm 1.5\ °C$), thus likely being one factor limiting their thermostability.

The computation of essential proteins revealed 221 unique essential proteins among the organisms of *E.coli*, *B. subtilis*, *P. torridus* and *T. thermophilus*, while no essential proteins were found for *O. antarctica*. As essential proteins from *P. torridus* and *T. thermophilus* only accounted for 5% of the total amount, these organisms were not further analysed due to the lack of data. After grouping the remaining proteins, key biological processes could be identified (Fig 4).
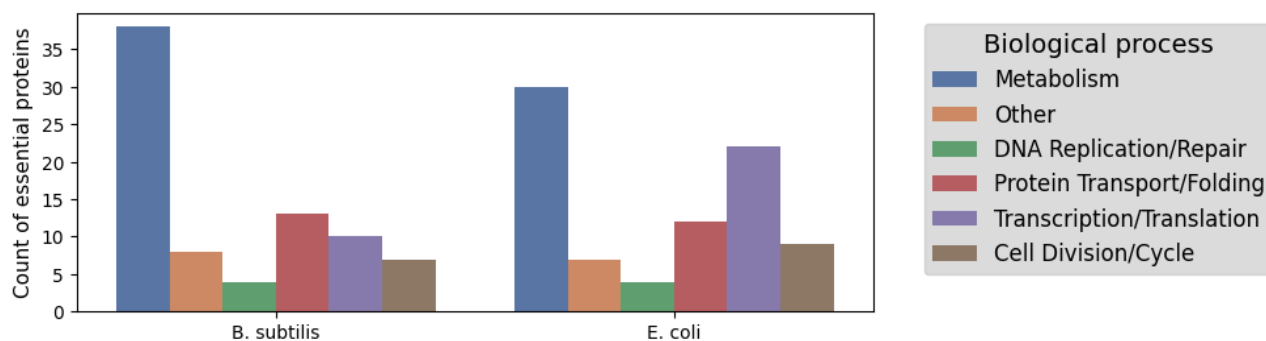

Figure 4: Essential proteins and their biological function for B.subtilis and E.coli.
For B. subtilis and E.coli the number of essential proteins for specific biological process categories are shown in a bar plot.

It becomes clear that a multitude of the essential proteins in both organisms are responsible for metabolic processes and synthesis of essential metabolites. Examples are the CTP synthase and thymidylate synthase, which are required in nucleotide synthesis, or glucose-6-phosphate dehydrogenase and triosephosphate isomerase, which play a key role in glycolysis and the pentose phosphate pathway, respectively. The following groups contained proteins involved in transcriptional and translational processes and protein folding or transport, where proteins related to transcription and translation displayed a higher count in *E. coli*. Among these were e.g. proteins associated with the 50S large ribosomal subunit, lysyl-tRNA synthetase, as well as the RNA polymerase subunit. Some examples for proteins involved in protein transport and folding were glycerol-3-phosphate transporter or the HSP-70 cofactor.

5.5 *In-silico* mutational screen

From the previously determined essential proteins, all proteins with an available 3D structure were used for *in-silico* mutagenesis. From 115 tested proteins with available protein structures, 103 showed an increase in the predicted melting point after our mutation approach.

For each organism the measured wild type (WT) melting point, the predicted WT melting point and the predicted melting point of the best mutant were (Fig. 5). B. subtilis and E. coli showed a significant difference, both when comparing measured melting point with WT predicted ($p = 2.05 \times 10^{-12}$ and $p = 1.67 \times 10^{-7}$, respectively) and when comparing WT and mutated melting point ($p = 3.70 \times 10^{-6}$ and $p = 9.59 \times 10^{-4}$, respectively). T. thermophilus and P. torridus, showed no significant difference for neither the measured vs. predicted nor the WT predicted vs. mutated predicted melting point ($p \geq 0.25$).
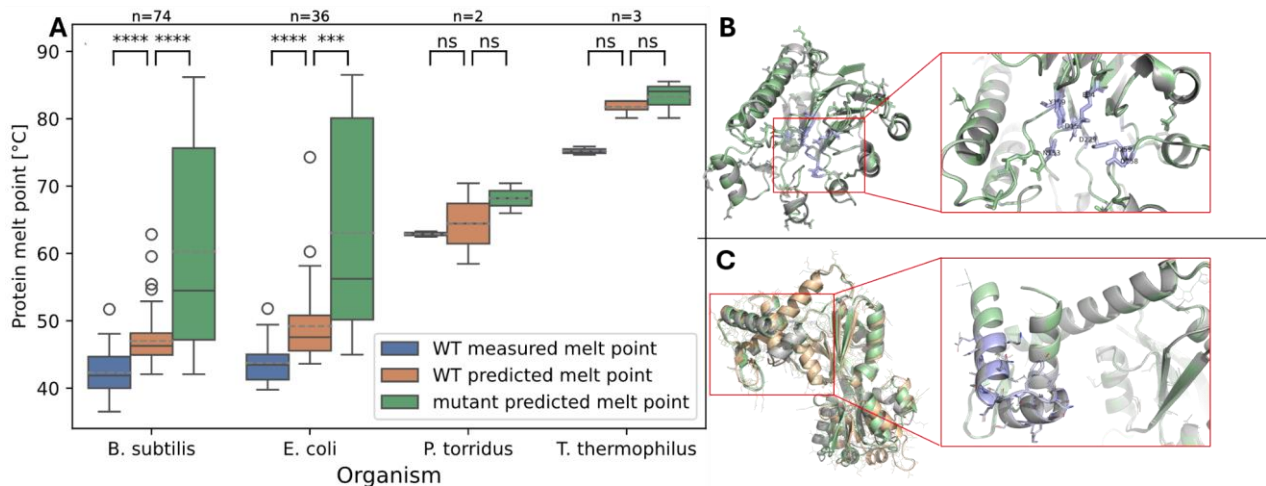

Figure 5: Melting point and structural analysis of mutated proteins.
(A) Proteins are categorised by organism, and the measured WT T_m (blue), predicted WT T_m (orange) and predicted mutated T_m (green) are shown. The dotted grey line represents the mean of each group. A normal distribution based on the temperature dependent error was used and random values drawn 1000 times, resulting averaged values were taken to calculate the significance using a Wilcoxon signed rank test.. *p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001.
(B) Predicted structures of the exodeoxyribonuclease III (UniProt: P09030)
(C) Predicted structure of the HTH-type transcriptional regulator GltR (UniProt: P94501).
The wild-type protein is shown in grey, and the optimised mutatand in green. Panel C shows in addition to the final optimised mutant (green) also the mutant before applying ThERMless in orange. For both proteins, the inset on the right side shows a detailed view of the active centre, determined by UniProt annotations. The purple sticks indicate amino acids involved in binding and catalytic activity. All shown 3D structures were predicted using the AlphaFold 3 web server.

The protein with the highest predicted increase in the melting point (46 °C to 87 °C, ΔTm = 41 °C, 32 mutations) was chosen for further analysis to evaluate the impact of the mutations on the protein structure. The chosen protein, exodeoxyribonuclease III (UniProt ID: P09030) is a nuclease involved in the DNA repair process.

The mutation at position 229 was reverted so that all amino acids involved in the catalysis remained unchanged, resulting in a predicted ΔTm of 38 °C. The structural alignment showed minimal deviation from the wild type structure with an RMSD of 0.198 Å, more importantly, both the active site accessibility and orientation of the catalytic active amino acid positions remained unchanged (Fig. 5B).

Additionally, the HTH (helix-turn-helix) type transcriptional regulator GltR from B. subtilis (P94501) was analysed as an example for a protein with a high percentage of predicted mutations, namely 70 out of 924 amino acids were mutated. The wild type protein has a predicted melting point of 46 °C (45 °C measured), which was improved to 84 °C (or 83 °C, when keeping residues in the active centre fixed) in the predicted mutant. After reducing the number of mutations with ThERMless (see section 5.6) by 17, the optimised mutant had a predicted melting point of 77 °C (ΔTm = 31 °C) As shown in Fig. 5 C the first, heavily mutated structure (shown in orange) exhibited a large deviation from the wild type protein (RMSD of 1.693 Å). After reverting 17 mutations, the structural deviation was significantly reduced (RMSD of 0.523 Å), with nearly unchanged structural position of the amino acids in the active centre (Fig. 5C).

5.6 Developed software

The SPARC (**S**equence-based **P**rotein **A**ttribute-de**R**ived (melting point) **C**alculator) function was developed to first calculate all Tm relevant parameters from the primary and secondary structure. These values were then transformed into the PCs used by the regression model which then predicts the melting point of the input protein. The function yields as output a list containing the predicted melting point and all computed protein features. For fast mutational analysis we created ThERMOS (**Th**ermal **E**nhancement by **R**apid **M**utational **O**ptimization **S**creen). ThERMOS performs site-saturation mutagenesis on free amino acids (as described in 4.6), optimising for the strongest correlating features using an internal scoring function. The function takes a PDB file as an input and creates a list of possible mutations for enhanced thermal stability as well as the predicted melting point increase using SPARC. A third function ThERMless can be used to remove mutations with minimal impact to reduce 3D structure deviations.

The three main functions SPARC, ThERMOS and ThERMless were concatenated into the Protein class, which takes an amino acid sequence and optionally a PDB file as input. Upon creating an instance of the class, all features and the predicted melting point of the protein are calculated and assigned as attributes of the object. ThERMOS and ThERMless can then be called as methods to get a mutated amino acid sequence with increased melting point, each adding its outputs to the attributes of the object.

All code is available on GitHub.

6 Discussion and Outlook

After analysing a variety of structural factors and motifs, most of our findings were consistent with published data. Arginine had a strong positive correlation as a single amino acid, but also in motifs, helices, and dipeptide content (16 out of the top 22 features) (Fig. 1, 2). Glutamine showed a strong negative correlation, being present in the top 13 out of 22 features (Fig. 1, 2). Both findings match previous publications (Facchiano et al., 1998; Kumar et al., 2001).
Identified secondary factors, specifically the number of helices did not match previous reports from literature. Comparing α-helix count to melting point, a correlation of -0.11 was calculated contradictory to a positive correlation identified by Vihinen, 1987 (data not shown). Furthermore the influence of α-helices on protein rigidity and flexibility and its subsequent effect on thermal stability is not yet fully understood (Karshikoff et al., 2015).

The contribution of factors related to secondary structure, specifically the number of α-helices, to thermostability is more controversial.
Correlating the number of α-helices to melting points, we observed a negative correlation of -0.11 (data not shown). Published data show very different results ranging from no correlation between the number of α-helices and thermostability (Facchiano et al., 1998) to a positive correlation (Vihinen, 1987). One possible explanation is the not yet fully understood influence of α-helices on protein rigidity and flexibility and its subsequent effect on thermostability (Karshikoff et al., 2015). Importantly, our observations show that the number of arginine residues in α-helices positively correlates with the melting point (0.35, Fig. 2), indicating that α-helices with high arginine content contribute to stability, overriding the otherwise rather neutral or even slightly destabilising effect of helices. This observation suggests that not the mere number of helices is important for thermostability, but rather their amino acid composition, which is in accordance with literature describing decreased flexibility and increased hydrophobicity, such as stabilising Arg substitutions in helices as main stabilising principles (Menéndez-Arias & Argosf, 1989; Petukhov et al., 1997).

The found tertiary structural motifs were consistent with the literature (Kumar et al., 2001; Niu et al., 2016; Vieille & Zeikus, 2001). However, the identified motifs generally showed low direct or inverse correlations (data not shown). The length of hydrophobic clusters had a positive correlation (0.12) while the number of hydrophobic clusters correlated negatively, leading to the assumption that few but larger hydrophobic patches are the most beneficial for a higher thermostability. The relative amount of hydrogen bonds correlated with a value of 0.07, thus showing no relevant correlation to melting point (Data not shown).
Lastly, protein packing also correlates to thermostability, which is in accordance with literature (Vieille & Zeikus, 2001). Here a negative correlation of -0.11 could be identified (data not shown), revealing that a more densely packed and thus a smaller protein exhibits a higher melting point, which was also previously described in literature (Kumar et al., 2000). Surprisingly, our findings showed only a minor impact of tertiary elements on protein stability, which differs from many results found in literature (Ahmed et al., 2022; Gromiha et al., 2013; Vogt & Argos, 1997).

After testing different approaches for regression analysis, a gradient boosting model performed best. The model error heavily depends on the predicted melting point and is also partially inversely proportional to the number of proteins in the category (Fig. 3). It especially shows a very low error of less than 5 °C in the range of 42 to 46 °C where the amount of proteins is highest, compared to an error of over 10 °C in the range of 55 °C to 65 °C with substantially less proteins. Towards the highest melting points, the error becomes very small again with only slightly more proteins compared to the intermediate temperature range. This suggests that high melting proteins are easier to distinguish from low and intermediate melting proteins than proteins in the intermediate temperature ranges. The RMSE of the model could be improved by using a larger dataset than the 6558 proteins we used, especially more proteins in the melting point ranges where the RMSE was highest. During development of the model we faced several limitations. To only rely on the amino acid sequence as an input, we chose to omit any tertiary structure information which was previously used in our analyses. Although structural prediction tools are available, they require significant computing power, making them infeasible to run locally (Abramson et al., 2024; Baek et al., 2021; Jumper et al., 2021). Furthermore a machine learning approach for secondary structure prediction was chosen (S4pred, (Moffat & Jones, 2021), resulting in an error of 0.433 and 0.390 for α-helices and β-sheets, respectively (data not shown). A different model, only trained

on proteins with available crystal structure, thus not using S4pred, could achieve an R^2 and RMSE only negligibly higher than the model using S4pred showing the error is not relevant for the regression (data not shown).

Despite the error, the performance of the model is on par with other published models. Another model based on the same initial dataset (Jarzab et al., 2020) is ProTstab2 (Yang et al., 2022), which also uses a gradient boosting regressor model. However it has an RMSE of 9.1 °C, R^2 of 0.58 and Pearson correlation coefficient of 0.8. OUR SPARC model outperforms ProTstab2 in each of these metrics. The SCooP algorithm (Pucci et al., 2017), trained on a small data set of 22 proteins, showed a worse Pearson correlation compared to SPARC (0.72 vs 0.86). A model that is clearly superior to SPARC is ProtSTABp (Jung et al., 2023), which uses a transformer-based protein language model combined with further deep learning techniques with higher complexity compared to our approach (R^2 of 0.8 and RMSE of 4.3 °C). Another state-of-the art tool is DeepTM (M. Li et al., 2023) which uses a deep learning algorithm (R^2 of 0.69, RMSE of 7.11 °C , correlation coefficient of 0.83) and is comparable to the results obtained from SPARCs. Another advantage of SPARC is its low operating time of around 15 seconds for a protein with 1000 amino acids,while other more complex machine learning approaches tend to have much longer processing times.

The analysis of essential proteins showed that most of them have biological functions related to metabolism or transcription/translation processes. These results were expected as such processes are crucial for cellular homeostasis. Proteins involved in cell cycle and growth activity were not as prominent as previously thought. This suggests that the growth of *B. subtilis* and *E. coli* is inhibited less by cell cycle/growth proteins and more by metabolism/cellular energetics and protein biosynthesis. Furthermore, the maximum growth temperature heavily depends on experimental conditions. For example it has been shown that the maximum growth temperature of *E. coli* can increase, when the temperature is raised slowly (Guyot et al., 2014).

Understanding the structural basis of thermal stability remains a key challenge in protein engineering. Improving protein thermal stability is an important goal for applications in biotechnology and pharmaceuticals (Rahban et al., 2022).

Next to evolutionary based protein engineering *in silico* modelling has become a powerful tool (Bunzel et al., 2021; Walker et al., 2021). As of now only few tools exist focusing on thermal engineering. Most commonly the Gibbs free energy ΔG and its changes through mutation (ΔΔG) are used as a predictive measure for thermal stability. This was successfully implemented in various tools ,despite having an imperfect correlation with experimentally measured thermostability (Pucci et al., 2016; Zhou et al., 2023). Recently new advances have been made combining an energy-based, evolution-based and ancestral reconstruction-based as described by (Musil et al., 2024). Furthermore, promising results have been generated using convolutional networks (G. Li et al., 2024). The main drawback of all these approaches is the high computational cost, requiring web-server integration, while still taking multiple hours to days for average sized proteins (Musil et al., 2024), whereas our created model needs minimal computational power. Although the 3D structure is required as an input, the advances in the field of structure prediction tools as AlphaFold have enabled fast and accurate structure prediction (Abramson et al., 2024; Jumper et al., 2021).

The model ThERMOS combines random mutagenesis with rational mutations, first creating a large diversity and then iteratively improving the protein. During the mutation process all internal calculations solely rely on the amino acid sequence, using an internal scoring function as well as SPARC temperature predictions. There still are several drawbacks using this approach. Using an iterative approach, does not take possible synergistic effects, and long range interactions into account, as well as getting trapped in local optima. To minimise this a random mutational screen was introduced and the best performing proteins, and strongest sequence deviating proteins are chosen for rational analysis. Furthermore during rational mutation only the strongest positively and negatively correlating features are considered, using the averaged calculated values of the top ten percentile as ideal values. This not only limits features optimised for but can also create a larger bias towards our dataset and identified correlations. Since, in the end, mutational success is evaluated based on our regression model, errors and biases therein could further impact the accuracy and performance of ThERMOS. To show the potential of our approach all essential genes with available structure data were mutated and improvements analysed. Significant improvements could be achieved for the mesophilic organisms B. subtilis and E. coli (maximum growth temperature of 55 °C and 45 °C, respectively) (Fig 5 A). Exemplary structure analysis of mutated proteins showed minimal structural deviations with no changes in structural positions of amino acid in the active centre (Fig. 5 B,C). Furthermore, it was shown that utilising ThERMless, mutations

with minimal impact on thermal stability can be reverted yielding an overall better structural alignment (Fig. 5 C). Although possible mutations could not be predicted for every protein, 48 % of proteins from *E. coli* and *B. subtilis* showed an increase more than 10 °C (data not shown).

In the presented work, we identified novel features contributing to the thermostability of prokaryotic proteins. We developed a comprehensive software tool capable of, but not limited to, predicting the melting temperature with high accuracy, and providing a novel approach to predict mutations for thermostability with minimal structural changes. To test its capability we identified important proteins, limiting prokaryotic thermostability. Using the developed software tools we showed substantial improvement for mesophilic prokaryotes.

We believe our toolset will be a valuable addition for protein engineering. Due to the low computation time it can easily be applied to large data sets as a starting point for wet lab work as well as augmenting already software tools approaches.

8 References

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., Bodenstein, S. W., Evans, D. A., Hung, C.-C., O'Neill, M., Reiman, D., Tunyasuvunakool, K., Wu, Z., Žemgulytė, A., Arvaniti, E., … Jumper, J. M. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, *630*(8016), 493–500. https://doi.org/10.1038/s41586-024-07487-w

Ahmed, Z., Zulfiqar, H., Tang, L., & Lin, H. (2022). A Statistical Analysis of the Sequence and Structure of Thermophilic and Non-Thermophilic Proteins. *International Journal of Molecular Sciences*, *23*(17), 10116. https://doi.org/10.3390/ijms231710116

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., Wang, J., Cong, Q., Kinch, L. N., Schaeffer, R. D., Millán, C., Park, H., Adams, C., Glassman, C. R., DeGiovanni, A., Pereira, J. H., Rodrigues, A. V., van Dijk, A. A., Ebrecht, A. C., … Baker, D. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, *373*(6557), 871–876. https://doi.org/10.1126/science.abj8754

Bondi, A. (1964). Van der Waals Volumes and Radii. *The Journal of Physical Chemistry*, *68*(3), 441–451. https://doi.org/10.1021/j100785a001

Bunzel, H. A., Anderson, J. L. R., & Mulholland, A. J. (2021). Designing better enzymes: Insights from directed evolution. *Current Opinion in Structural Biology*, *67*, 212–218. https://doi.org/10.1016/j.sbi.2020.12.015

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., & De Hoon, M. J. L. (2009). Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*(11), 1422–1423. https://doi.org/10.1093/bioinformatics/btp163

D'Agostino, R. B. (1971). An omnibus test of normality for moderate and large size samples. *Biometrika*, *58*(2), 341–348. https://doi.org/10.1093/biomet/58.2.341

D'Agostino, R., & Pearson, E. S. (1973). Tests for departure from normality. Empirical results for the distributions of b^2 and $\sqrt{b^1}$. *Biometrika*, *60*(3), 613–622. https://doi.org/10.1093/biomet/60.3.613

Dolinsky, T. J., Nielsen, J. E., McCammon, J. A., & Baker, N. A. (2004). PDB2PQR: An automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Research*, *32*(Web Server), W665–W667. https://doi.org/10.1093/nar/gkh381

Facchiano, A. M., Colonna, G., & Ragone, R. (1998). Helix stabilizing factors and stabilization of thermophilic proteins: An X-ray based study. *Protein Engineering Design and Selection*, *11*(9), 753–760. https://doi.org/10.1093/protein/11.9.753

Ferruz, N., Schmidt, S., & Höcker, B. (2021). ProteinTools: A toolkit to analyze protein structures. *Nucleic Acids Research*, *49*(W1), W559–W566. https://doi.org/10.1093/nar/gkab375

Gromiha, M. M., Pathak, M. C., Saraboji, K., Ortlund, E. A., & Gaucher, E. A. (2013). Hydrophobic environment is a key factor for the stability of thermophilic proteins. *Proteins*, *81*(4), 715–721. https://doi.org/10.1002/prot.24232

Guyot, S., Pottier, L., Hartmann, A., Ragon, M., Hauck Tiburski, J., Molin, P., Ferret, E., & Gervais, P. (2014). Extremely rapid acclimation of *Escherichia coli* to high temperature over a few generations of a fed-batch culture during slow warming. *MicrobiologyOpen*, *3*(1), 52–63. https://doi.org/10.1002/mbo3.146

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Jarzab, A., Kurzawa, N., Hopf, T., Moerch, M., Zecha, J., Leijten, N., Bian, Y., Musiol, E., Maschberger, M., Stoehr, G., Becher, I., Daly, C., Samaras, P., Mergner, J., Spanier, B., Angelov, A., Werner, T., Bantscheff, M., Wilhelm, M., … Kuster, B. (2020). Meltome atlas—Thermal proteome stability across the tree of life. *Nature Methods*, *17*(5), 495–503. https://doi.org/10.1038/s41592-020-0801-4

Johnson, L. N. (2009). The regulation of protein phosphorylation. *Biochemical Society Transactions*, *37*(4), 627–641. https://doi.org/10.1042/BST0370627

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Jung, F., Frey, K., Zimmer, D., & Mühlhaus, T. (2023). DeepSTABp: A Deep Learning Approach for the Prediction of Thermal Protein Stability. *International Journal of Molecular Sciences*, *24*(8), 7444. https://doi.org/10.3390/ijms24087444

Karshikoff, A., Nilsson, L., & Ladenstein, R. (2015). Rigidity versus flexibility: The dilemma of understanding protein thermal stability. *The FEBS Journal*, *282*(20), 3899–3917. https://doi.org/10.1111/febs.13343

Kumar, S., Tsai, C.-J., & Nussinov, R. (2000). Factors enhancing protein thermostability. *Protein Engineering, Design and Selection*, *13*(3), 179–191. https://doi.org/10.1093/protein/13.3.179

Kumar, S., Tsai, C.-J., & Nussinov, R. (2001). Thermodynamic Differences among Homologous Thermophilic and Mesophilic Proteins. *Biochemistry*, *40*(47), 14152–14165. https://doi.org/10.1021/bi0106383

Leuenberger, P., Ganscha, S., Kahraman, A., Cappelletti, V., Boersema, P. J., von Mering, C., Claassen, M., & Picotti, P. (2017). Cell-wide analysis of protein thermal unfolding reveals determinants of thermostability. *Science*, *355*(6327), eaai7825. https://doi.org/10.1126/science.aai7825

Li, G., Yao, S., & Fan, L. (2024). ProSTAGE: Predicting Effects of Mutations on Protein Stability by Using Protein Embeddings and Graph Convolutional Networks. *Journal of Chemical Information and Modeling*, *64*(2), 340–347. https://doi.org/10.1021/acs.jcim.3c01697

Li, M., Wang, H., Yang, Z., Zhang, L., & Zhu, Y. (2023). DeepTM: A deep learning algorithm for prediction of melting temperature of thermophilic proteins directly from sequences. *Computational and Structural Biotechnology Journal*, *21*, 5544–5560. https://doi.org/10.1016/j.csbj.2023.11.006

McKinney, W. (2010). *Data Structures for Statistical Computing in Python*. 56–61. https://doi.org/10.25080/Majora-92bf1922-00a

Menéndez-Arias, L., & Argosf, P. (1989). Engineering protein thermal stability: Sequence statistics point to residue substitutions in α-helices. *Journal of Molecular Biology*, *206*(2), 397–406. https://doi.org/10.1016/0022-2836(89)90488-9

Moffat, L., & Jones, D. T. (2021). Increasing the accuracy of single sequence prediction methods using a deep semi-supervised learning framework. *Bioinformatics*, *37*(21), 3744–3751. https://doi.org/10.1093/bioinformatics/btab491

Musil, M., Jezik, A., Horackova, J., Borko, S., Kabourek, P., Damborsky, J., & Bednar, D. (2024). FireProt 2.0: Web-based platform for the fully automated design of thermostable proteins. *Briefings in Bioinformatics*, *25*(1), bbad425. https://doi.org/10.1093/bib/bbad425

Niu, C., Zhu, L., Xu, X., & Li, Q. (2016). Rational Design of Disulfide Bonds Increases Thermostability of a Mesophilic 1,3-1,4-β-Glucanase from Bacillus terquilensis. *PLOS ONE*, *11*(4), e0154036. https://doi.org/10.1371/journal.pone.0154036

Ohtani, N., Tomita, M., & Itaya, M. (2010). An Extreme Thermophile, *Thermus thermophilus* , Is a Polyploid Bacterium. *Journal of Bacteriology*, *192*(20), 5499–5505. https://doi.org/10.1128/JB.00662-10

Open AI. (2024) *ChatGPT-4o* (July 07 version) [Large Language Model]

Pechmann, S., & Frydman, J. (2014). Interplay between Chaperones and Protein Disorder Promotes the Evolution of Protein Networks. *PLoS Computational Biology*, *10*(6), e1003674. https://doi.org/10.1371/journal.pcbi.1003674

Petukhov, M., Kil, Y., Kuramitsu, S., & Lanzov, V. (1997). Insights into thermal resistance of proteins from the intrinsic stability of their α-helices. *Proteins: Structure, Function, and Bioinformatics*, *29*(3), 309–320. https://doi.org/10.1002/(SICI)1097-0134(199711)29:3<309::AID-PROT5>3.0.CO;2-5

Pucci, F., Bourgeas, R., & Rooman, M. (2016). Predicting protein thermal stability changes upon point mutations using statistical potentials: Introducing HoTMuSiC. *Scientific Reports*, *6*(1), 23257. https://doi.org/10.1038/srep23257

Pucci, F., Kwasigroch, J. M., & Rooman, M. (2017). SCooP: An accurate and fast predictor of protein stability curves as a function of temperature. *Bioinformatics*, *33*(21), 3415–3422. https://doi.org/10.1093/bioinformatics/btx417

Rahban, M., Zolghadri, S., Salehi, N., Ahmad, F., Haertlé, T., Rezaei-Ghaleh, N., Sawyer, L., & Saboury, A. A. (2022). Thermal stability enhancement: Fundamental concepts of protein engineering strategies to manipulate the flexible structure. *International Journal of Biological Macromolecules*, *214*, 642–654. https://doi.org/10.1016/j.ijbiomac.2022.06.154

Savitski, M. M., Reinhard, F. B. M., Franken, H., Werner, T., Savitski, M. F., Eberhard, D., Molina, D. M., Jafari, R., Dovega, R. B., Klaeger, S., Kuster, B., Nordlund, P., Bantscheff, M., & Drewes, G. (2014). Tracking cancer drugs in living cells by thermal profiling of the proteome. *Science*, *346*(6205), 1255784. https://doi.org/10.1126/science.1255784

Schrödinger, LLC. (2015). *The PyMOL Molecular Graphics System, Version 1.8*.

Shental-Bechor, D., & Levy, Y. (2008). Effect of glycosylation on protein folding: A close look at thermodynamic

stabilization. *Proceedings of the National Academy of Sciences*, *105*(24), 8256–8261. https://doi.org/10.1073/pnas.0801340105

Shrake, A., & Rupley, J. A. (1973). Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *Journal of Molecular Biology*, *79*(2), 351–371. https://doi.org/10.1016/0022-2836(73)90011-9

Šimunović, K., Stefanic, P., Klančnik, A., Erega, A., Mandic Mulec, I., & Smole Možina, S. (2022). Bacillus subtilis PS-216 Antagonistic Activities against Campylobacter jejuni NCTC 11168 Are Modulated by Temperature, Oxygen, and Growth Medium. *Microorganisms*, *10*(2), Article 2. https://doi.org/10.3390/microorganisms10020289

Speck, J., Hecky, J., Tam, H.-K., Arndt, K. M., Einsle, O., & Müller, K. M. (2012). Exploring the Molecular Linkage of Protein Stability Traits for Enzyme Optimization by Iterative Truncation and Evolution. *Biochemistry*, *51*(24), 4850–4867. https://doi.org/10.1021/bi2018738

Tan, K. P., Singh, K., Hazra, A., & Madhusudhan, M. S. (2021). Peptide bond planarity constrains hydrogen bond geometry and influences secondary structure conformations. *Current Research in Structural Biology*, *3*, 1–8. https://doi.org/10.1016/j.crstbi.2020.11.002

The UniProt Consortium. (2023). UniProt: The Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*, *51*(D1), D523–D531. https://doi.org/10.1093/nar/gkac1052

Vieille, C., & Zeikus, G. J. (2001). Hyperthermophilic Enzymes: Sources, Uses, and Molecular Mechanisms for Thermostability. *Microbiology and Molecular Biology Reviews*, *65*(1), 1–43. https://doi.org/10.1128/MMBR.65.1.1-43.2001

Vihinen, M. (1987). Relationship of protein flexibility to thermostability. *Protein Engineering, Design and Selection*, *1*(6), 477–480. https://doi.org/10.1093/protein/1.6.477

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., Van Der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … Vázquez-Baeza, Y. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

Vogt, G., & Argos, P. (1997). Protein thermal stability: Hydrogen bonds or internal packing? *Folding and Design*, *2*, S40–S46. https://doi.org/10.1016/S1359-0278(97)00062-X

Walker, S. P., Yallapragada, V. V. B., & Tangney, M. (2021). Arming Yourself for The *In Silico* Protein Design Revolution. *Trends in Biotechnology*, *39*(7), 651–664. https://doi.org/10.1016/j.tibtech.2020.10.003

Wang, Y., & Jardetzky, O. (2002). Probability-based protein secondary structure identification using combined NMR chemical-shift data. *Protein Science*, *11*(4), 852–861. https://doi.org/10.1110/ps.3180102

Waskom, M. (2021). seaborn: Statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021

Yakimov, M. M., Giuliano, L., Gentile, G., Crisafi, E., Chernikova, T. N., Abraham, W.-R., Lünsdorf, H., Timmis, K. N., & Golyshin, P. N. (2003). Oleispira antarctica gen. Nov., sp. Nov., a novel hydrocarbonoclastic marine bacterium isolated from Antarctic coastal sea water. *International Journal of Systematic and Evolutionary Microbiology*, *53*(3), 779–785. https://doi.org/10.1099/ijs.0.02366-0

Yang, Y., Zhao, J., Zeng, L., & Vihinen, M. (2022). ProTstab2 for Prediction of Protein Thermal Stabilities. *International Journal of Molecular Sciences*, *23*(18), 10798. https://doi.org/10.3390/ijms231810798

Zhou, Y., Pan, Q., Pires, D. E. V., Rodrigues, C. H. M., & Ascher, D. B. (2023). DDMut: Predicting effects of mutations on protein stability using deep learning. *Nucleic Acids Research*, *51*(W1), W122–W128. https://doi.org/10.1093/nar/gkad472

9 Supplementary Material

Table S1 Amino acid categorization

Polar	Hydrophobic	Charged
N, Q, S, T, Y	A, V, I, L, M, F, W	R, H, K, D, E

Table S2 Amino acids and atoms involved in hydrogen bonds

amino acid	donor atom	acceptor atom
glutamine	NE2: HE21, HE22	OE1
glutamate	OE2: HE2	OE1, OE2
aspartate	OD2: HD2	OD1, OD2
asparagine	ND2: HD21, HD22	OD1
histidine	NE2: HE2 ND1: HD2	
lysine	NZ: HZ1, HZ2, HZ3	
arginine	NE: HE NH1: HH11, HH12 NH2: HH21, HH22	
serine	OG: HG1	OG
threonine	OG1: HG1	OG1
tryptophan	NE1: HE1	
tyorsine	OH: HH	

Table S3 Conservative mutations

amino acid	possible substitutions
A	D, E, G, S, T
C	G, R, S, W, Y
D	A, E, G, H, N, V, Y
E	A, D, G, K, Q, V
F	I, L, Y
G	A, C, D, E, R
H	D, L, N, P, Q, R, Y
I	D, L, M, N, V
K	E, M, N, Q, R, T
L	F, H, I, M, P, Q, R, V, W
M	I, K, L, R, T, V
N	D, H, I, K, S, T, Y
P	H, L, Q, R, S
Q	E, H, K, L, P, R
R	C, G, H, K, L, M, P, Q, T, W
S	A, C, N, P, T, W, Y
T	A, K, M, N, R, S
V	D, E, I, L, M
W	C, L, R, S
Y	C, D, F, H, N