Received: 28 March 2022 | Revised: 2 September 2022 | Accepted: 15 October 2022

DOI: 10.1002/pro.4480

METHODS AND APPLICATIONS

Learning deep representations of enzyme thermal adaptation

Gang Li[1] | Filip Buric[1] | Jan Zrimec[1,2] | Sandra Viknander[1] | Jens Nielsen[1,3] | Aleksej Zelezniak[1,4,5] | Martin K. M. Engqvist[1,6]

[1]Department of Biology and Biological Engineering, Chalmers University of Technology, Gothenburg, Sweden

[2]Department of Biotechnology and Systems Biology, National Institute of Biology, Ljubljana, Slovenia

[3]BioInnovation Institute, Copenhagen N, Denmark

[4]Life Sciences Centre, Institute of Biotechnology, Vilnius University, Vilnius, Lithuania

[5]Randall Centre for Cell & Molecular Biophysics, King's College London, New Hunt's House, Guy's Campus, SE1 1UL, London, UK

[6]Enginzyme AB, Stockholm, Sweden

Correspondence

Martin K. M. Engqvist, Department of Biology and Biological Engineering, Chalmers University of Technology, Kemivägen 10, SE-41296 Gothenburg, Sweden.

Email: martin.engqvist@chalmers.se

Funding information

H2020 Marie Skłodowska-Curie Actions, Grant/Award Number: 722287; Javna Agencija za Raziskovalno Dejavnost RS, Grant/Award Number: J2-3060; Knut and Alice Wallenberg Foundation; Marius Jakulis Jason Foundation; Novo Nordisk Fonden, Grant/Award Number: NNF10CC1016517; Public Scholarship, Development, Disability and Maintenance Fund of the Republic of Slovenia, Grant/Award Number: 11013-9/2021-2; Science for Life Laboratory; Vetenskapsrådet, Grant/Award Number: 2019-05356; SciLifeLab

Review Editor: Nir Ben-Tal

Abstract

Temperature is a fundamental environmental factor that shapes the evolution of organisms. Learning thermal determinants of protein sequences in evolution thus has profound significance for basic biology, drug discovery, and protein engineering. Here, we use a data set of over 3 million BRENDA enzymes labeled with optimal growth temperatures (OGTs) of their source organisms to train a deep neural network model (DeepET). The protein-temperature representations learned by DeepET provide a temperature-related statistical summary of protein sequences and capture structural properties that affect thermal stability. For prediction of enzyme optimal catalytic temperatures and protein melting temperatures via a transfer learning approach, our DeepET model outperforms classical regression models trained on rationally designed features and other deep-learning-based representations. DeepET thus holds promise for understanding enzyme thermal adaptation and guiding the engineering of thermostable enzymes.

KEYWORDS

bioinformatics, deep neural networks, enzyme catalytic temperatures, optimal growth temperatures, protein thermostability, transfer learning

Gang Li and Filip Buric contributed equally to this study.

1 | INTRODUCTION

Nature has spent billions of years adapting organisms to various thermal niches, where environmental temperatures range from below −10°C to over +110°C.[1] Since a genome contains all the information required for building and maintaining an organism, the thermal adaptation strategies found in nature are inherently encoded in genomes. In the past decades, much effort has been made to uncover and understand such intrinsic strategies at various levels that include DNA, RNA, proteins, and metabolic pathways.[2,3] Unsurprisingly, most thermal adaptation strategies are clearly reflected at the protein level,[2] since proteins are involved in almost all cellular functions and are the most temperature sensitive out of all macromolecules.[4–6] Understanding temperature effects on proteins is also fundamental to basic biology,[5,7,8] drug discovery,[9] and protein engineering.[10] A large portion of studies have thus focused on the temperature effects on protein folding[5,7,11,12] and biological functions[8,13,14] as well as the combined effects at the systems level.[15–18] Despite this, it remains unclear how the effects of temperature on a protein are determined by its amino acid sequence.

Although there are many factors that were found to contribute to the thermosensitivity of proteins, including protein length,[12] amino acid compositions and properties[19–21] as well as structural properties,[22–24] these factors are found to be only weak determinants of the protein thermal properties, such as their unfolding behaviors[18,25] and optimal catalytic temperature points.[26] We hypothesize that by extracting patterns from protein sequences that are related to protein thermal adaptation, we can not only further our understanding of enzyme thermal adaptation, but also provide a rich feature set for many enzyme-related machine learning (ML) applications.

To this end, in the present study, we apply deep learning to uncover the protein sequence-encoded thermal determinants and learn a predictive representation of enzyme thermal adaptation. A few recent studies have applied a similar approach, but with a more limited scope. Gado et al.[27] trained a model to predict the optimal catalytic temperature of engineered thermophilic enzymes, Min et al.[28] sought to identify heat shock proteins, Yu et al.[29] predicted thermostability of collagen proteins, the model by Yang et al.[30] was used to predict protein melting temperatures, and Zhang et al.[31] used a deep model to identify thermophilic homologs of a given chitinase from a large pool of metagenomic data. Complementing these studies, we leverage a large data set of 3 million enzymes across a wide range of organisms and train a deep neural network model to capture sequence features that are predictive of thermostability. To give more value to the research community, we make the model available for download.

2 | RESULTS

2.1 | Learning representations of enzyme thermal adaptation

With the assumption that all proteins from an organism should be functional at its optimal growth temperature (OGT), we previously obtained a data set with 6.5 million enzymes labeled with OGT based on their source organisms.[32] Here, we removed similar and low-quality sequences and generated a data set with 3 million enzymes from bacteria, eukaryotes, and archaea (Figure 1a,b, Section 4), which we shall refer to as the OGT data set. For modeling, we chose the residual neural network architecture,[33] which has been successfully applied on protein function annotation.[34] After optimization (Section 4, Figures S1–S3), the resulting model contained only 1 residual block with 512 filters (kernels) (Figure 1c). For model training, one-hot encoded enzyme sequences were used as input and OGT values as output, after which the model could explain ~60% of the variance in the hold out data set (Pearson's $r = 0.77$, p value < 1e−16, Figure 1d). We refer to this model as DeepET hereafter. In DeepET, the network components preceding the Flatten layer can be considered as a feature extractor (Figure 1c), while the last dense layers can be considered as a regressor on top of the above feature extractor. Therefore the values in the Flatten layer (20,480 in total) form a temperature-related representation of input protein sequences (Figure 1c).

The considerable data imbalance (in value distribution) that is present in the OGT data set (Figure 1a,b) was addressed during hyperparameter optimization by subsampling 10,000 values such that a uniform distribution of values across 5° bins covered the entire range (Figure S2b and Section 4). Other rebalancing methods are possible, as illustrated in Zhang et al.[31] and Gado et al.,[27] but the potential of model improvement has to be weighed against the method's computational cost. For example, Gado et al.[27] performed a combination of data resampling and ensemble learning for their regression task to predict thermophiles (i.e., sparse data). DeepET was trained on a much larger data set and the measures taken in the aforementioned work would not have been practical here.

2.2 | Transfer learning improves the prediction of protein thermal properties

We next demonstrated the application of DeepET in a transfer learning approach.[35] In transfer learning, a model pre-trained on a large (source) data set, such as DeepET, is re-purposed to another similar (target) problem from the same or a related domain with a smaller

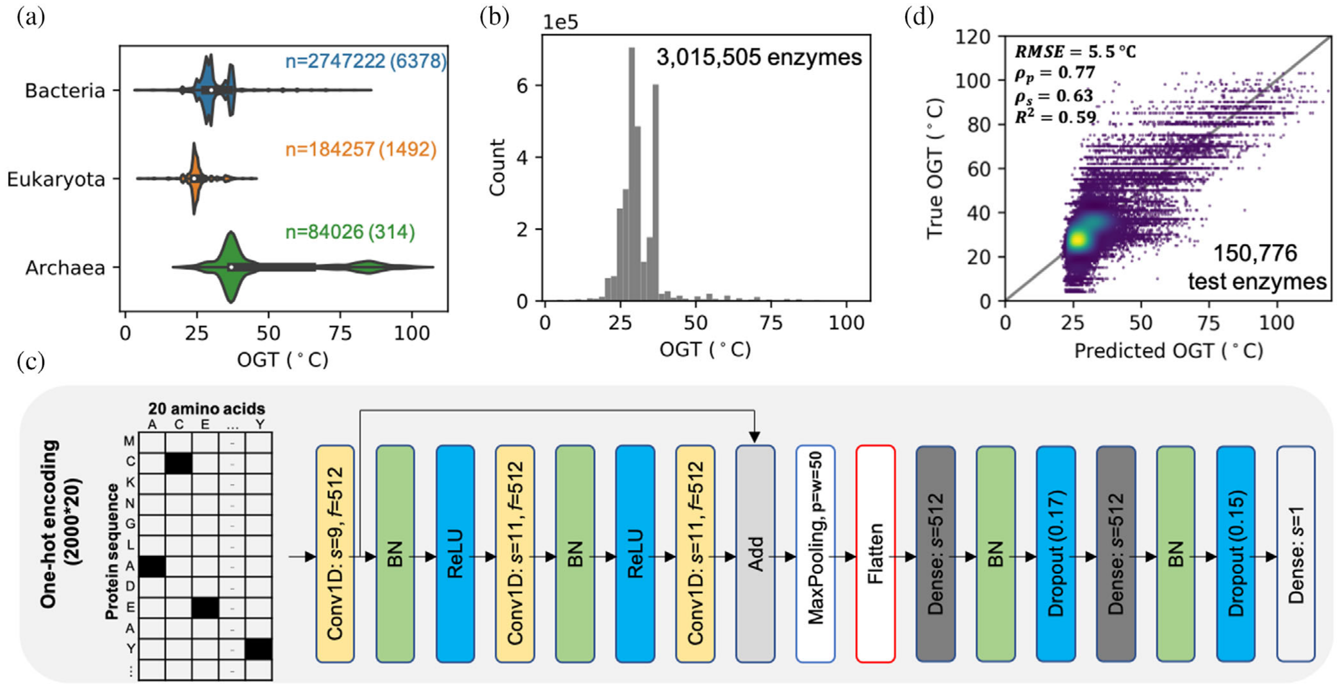

FIGURE 1 Learning representations of enzyme thermal adaptation with DeepET. (a) OGT distribution of enzymes from three domains, where *n* indicates the number of enzymes from each domain and the number in the parentheses is the number of species where those enzymes are from. (b) The OGT distribution of all enzymes in the training data set. (c) Optimized architecture of deep neural networks used in this study, where *s* indicates the filter (kernel) size of convolutional layers and number of nodes for dense layers; *f* indicates the number of filters (kernels); *p* and *w* in the max pooling layer indicate the pool size and the length of the stride, respectively; the floating-point number in a dropout layer indicates the dropout ratio; BN denotes Batch Normalization. The convolutional layers have padding set to "same," while the max pooling has it set to "valid" (no padding). Layers use the ReLU activation function. (d) Comparison between predicted and true OGT values of enzymes in the hold out data set. RMSE, root mean squared error; ρ_p, *Pearson*'s correlation coefficient; ρ_s, *Spearman*'s correlation coefficient; R^2, coefficient of determination. OGT, optimal growth temperature

amount of training samples, by (a) further training and thus fine-tuning certain layers or (b) resetting their weights and training them from scratch.[36,37] This is particularly useful for biological data sets since (a) large numbers of biological samples are expensive to collect and (b) the capacity of classical ML models like random forest are usually limited by the availability of relevant features.[26] Here, we chose to predict two critical temperature-related features of proteins: enzyme optimal catalytic temperatures (T_{opt}), at which the specific activity is maximized, and melting temperatures (T_{m}), at which there is a 50% possibility that a protein is in a denatured state. For this, two small data sets were collected from literature, one for enzyme T_{opt} with 1,902 samples, which we shall refer to as the TOPT data set (Figure S4a)[26] and another for protein T_{m} with 2,506 samples, which we shall refer to as the TM data set (Figure S4b)[5] (Section 4).

To compare against the deep learning approach, two feature sets were also extracted for classical regression models (Section 4): (a) *iFeatures*,[38] which contains 5,494 protein sequence features, such as amino acid composition and autocorrelation properties; and (b) *UniRep*,[39] a multiplicative long-/short-term-memory recurrent neural network (mLSTM) based representation (5,700 features) of protein sequences, which was trained by its authors on ~24 million protein sequences via unsupervised learning. UniRep provides "generic" high-dimensional features based on learning from sequence alone. This type of features, which generally capture protein physicochemical, structural, and evolutionary properties, can be obtained from other models, such as SeqVec,[40] TAPE,[41] or ESM.[42] Among these, TAPE proved technically problematic due to its large memory requirements, while ESM was only published in the late stages of our study. Furthermore, UniRep features enable more distinct sequence clustering than SeqVec, and was thus chosen as a viable benchmark to test our transfer learning approach against.

The performance of the two iFeature and UniRep feature sets was tested with six regression models (Figure 2a,b: three best models shown). The deep transfer learning procedure included: (a) training the model shown in Figure 1c *from scratch* (randomly initialized weights); (b) testing the performance of pre-trained DeepET without any fine-tuning steps (Figure 2a,b: *FrozenAll*); (c) freezing

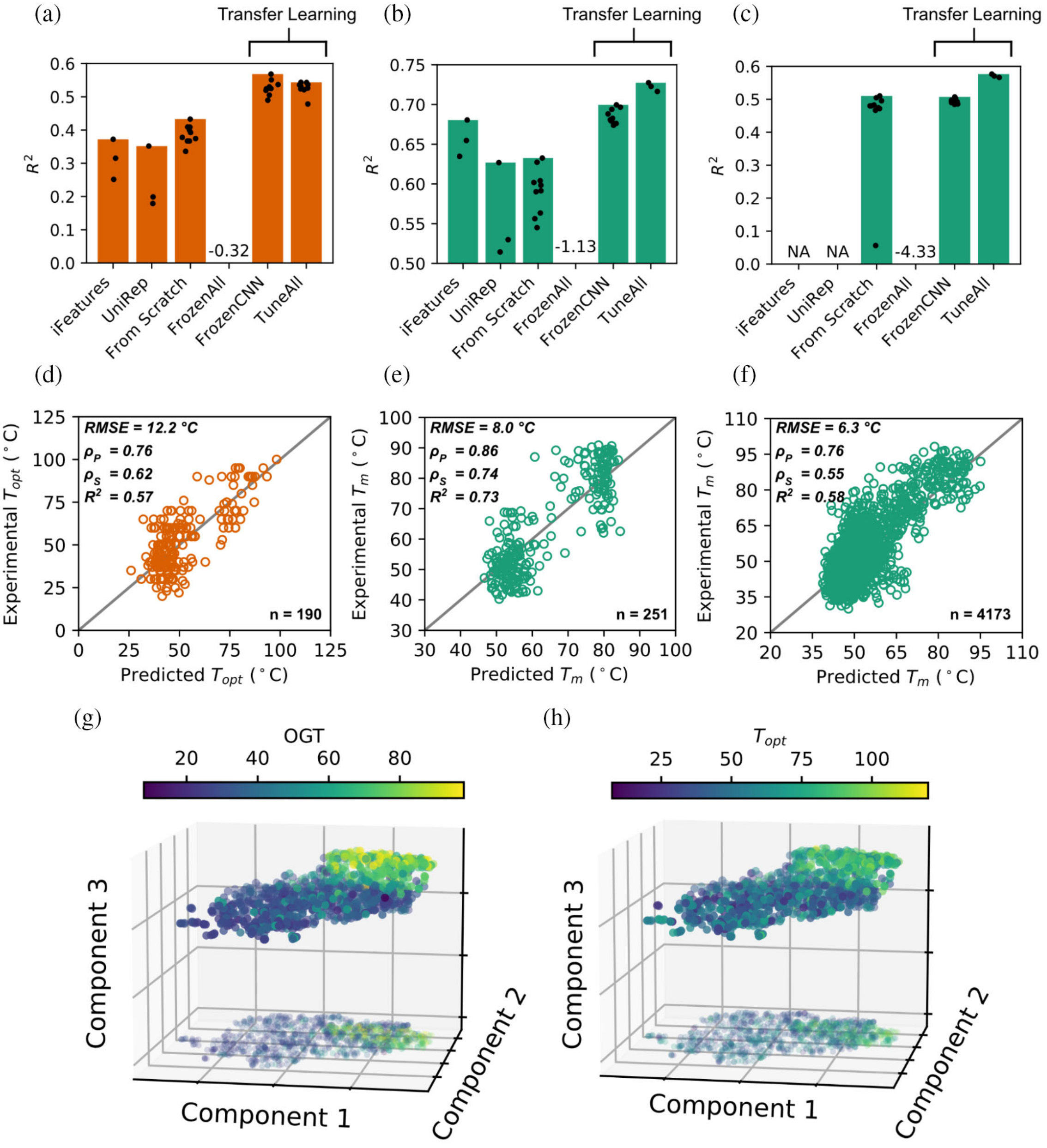

FIGURE 2 Transfer learning improves the prediction of protein thermal properties. (a–c) R^2 scores of different modeling approaches on hold out data sets representing 10% of their respective whole data sets: (a) TOPT data set: 190 enzyme optimal catalytic temperatures obtained from ref. 26, (b) TM data set: 251 protein melting temperatures obtained from ref. 5 and (c) MELT data set: 4,173 protein melting temperatures obtained from ref. 7 (details in Section 4). Bars indicate the maximal R^2 score, while the black dots indicate scores from repeated training. *iFeatures*: the performance of three best classical regression models using features extracted by iFeature.[38] *UniRep*: the performance of three best classical regression models using features extracted by UniRep.[39] *From Scratch*, the model shown in Figure 1c, was trained from scratch (repeated for 10 times). *FronzenAll*, the pre-trained model, was used without any further tuning for prediction. *FrozenCNN*, frozen all layers before Flatten (Figure 1c) and fine-tuned dense layers (repeated 10 times). *TuneAll*, fine-tuned all layers in the pre-trained OGT model (repeated three times). For (a) TOPT, the performance increase from *iFeatures* and *UniRep* regression models to the DeepET transfer learning models is significant (Welch's *t*-test *p* value < .05), as is the case between *From Scratch* and the transfer learning models (*p* value < 1e−9). For (b) TM, *TuneAll* is significantly better than the *iFeatures* model (*p* value < .05), the highest among the classical models. For (c) MELT, *TuneAll* is significantly better than *From Scratch* (*p* value < .05) and *FrozenCNN* (*p* value < 1e−4). (d–f) Comparison between predicted and experimental $T_{\mathrm{opt}}/T_{\mathrm{m}}$ in hold out data sets, matching (a–c), respectively. Results of the best model with the highest test R^2 score in (a–c) are shown, namely, (d) FrozenCNN on the TOPT data set, (e) TuneAll on the TM data set, and (f) TuneAll on the MELT data set. *RMSE* = root mean squared error; ρ_p = *Pearson*'s correlation coefficient; ρ_s = *Spearman*'s correlation coefficient; R^2 = coefficient of determination. (g, h) UMAP projections to three components (with shadows at the bottom of the scatter plots) of 1,040 TOPT enzyme sequence embeddings produced by the feature extractor section of DeepET (i.e., the output of the Flatten layer), colored by (g) OGT and (h) T_{opt}

convolution layers and fine-tuning the last two dense layers (*FrozenCNN*); (d) fine-tuning all layers in DeepET (*TuneAll*). All models considered were evaluated on the same data sets. In contrast to methodologically similar recent work in heat shock protein classification,[28] we wished to distinguish between the domain-specific representations learned by DeepET (and performance thereof) and the generic representations provided by models such as UniRep, by composing the latter with classical regression models. In contrast, in the heat shock protein study by Min et al., generic representations are composed with a CNN sequence classifier.

For the tasks of predicting T_{opt} and T_{m} (Figure 2a,b), DeepET showed superior performance over all other tested strategies when fine-tuning all of its layers (see Section 4). For prediction of enzyme T_{opt} (Figure 2a), the best model with an R^2 of .57 on the hold out data set was achieved by simply fine-tuning the last two dense layers (Figure 2a,d). This performance is over 50% higher than with the best classical regression models trained on iFeatures or UniRep, and over 30% higher than with the best deep learning model trained from scratch (Figure 2a). The previous best enzyme T_{opt} prediction model with an R^2 of 0.61 on the hold out data set[26] was achieved by using amino acid compositions together with OGT as input features. The application of this model is thus limited to native enzymes from microorganisms with known OGT. On the other hand, with DeepET, the new T_{opt} model can in principle be applied to any enzyme regardless of organismal sources. For the prediction of T_{m} (Figure 2b), melting temperatures of proteins from three microorganisms (*Escherichia coli*, *Saccharomyces cerevisiae*, and *Thermus thermophilus*) were used.[5] The best model with an R^2 of .73 on the hold out data set was achieved by fine-tuning all layers in DeepET. The performance is 7% higher than with the best model trained on iFeatures, 16% higher than with UniRep, and 15% higher than with the model trained from scratch (Figure 2b,e).

The two TOPT and TM data sets (Figure 2a,b) are small and comprise only a few thousand samples, which is why they benefited from the transfer learning approach.[35,43] Transfer learning may not provide the same benefit for large data sets. Therefore, in the third task, to test whether our DeepET network also delivers superior performance for big data sets, we used 41,725 proteins with known melting temperatures from Meltome[7] (Figure S4c), which we shall refer to here as the MELT data set. Due to the size of this data set, we could only test the performance of the various deep learning approaches (Figure 2c), since classical models become inefficient to train and optimize.[44] Surprisingly, fine-tuning all layers still outperform the model trained from scratch (Figure 2c,f: 13% improvement in R^2). The recent CNN-based architecture CARP[30] achieved a Spearman correlation of .54 on the Meltome data set, which is very similar (allowing for the differences in data partitioning) to the performance of DeepET correlation of .55. As a model that is pre-trained on sequences in an unsupervised way, CARP is another source of generic representations that may be used for downstream tasks, such as protein thermostability. While the Meltome performance was lower than the Transformer-based ESM[42] architecture, it shows the competitiveness of CNNs in terms of computational cost and thus accessibility to the average lab. In spite of the data imbalance in the OGT source data set, we saw less performance degradation for higher temperature proteins in the target tasks outlined here, compared to the source task. A recent study on bias mitigation for transfer learning illustrated that in some situations reducing bias is more impactful if done for the target task.[45] While we have not performed bias mitigation here, the TOPT and TM data sets are less biased than OGT (Figure S4a,b), a fact that may explain the strong performance of DeepET at high temperatures.

Our results demonstrate that the representations learned by DeepET (values in the Flatten layers, Figure 1c) were predictive in all the above three data sets (Figure 2a–c, FrozenCNN). For the task of predicting enzyme T_{opt} (Figure 2a), fine-tuning dense layers achieved similar performance as fine-tuning all layers (Welch's *t*-test *p* value = .97), meaning that features in the Flatten layer of DeepET are already a collection of informative descriptors for enzyme T_{opt}. For protein melting temperatures, although the Flatten layer contains informative descriptors for this task, fine-tuning all layers in DeepET showed even better results (Welch's *t*-test, Figure 2b: *p* value = 4e−5; Figure 2c: *p* value = 3e−9, see also Figure S5 for a pairwise assessment of performance difference significance).

To visualize how well these representations sort proteins by thermal adaptation, we took 1,040 sequences from the TOPT set (those that simultaneously had OGT and T_{opt} values), collected their DeepET representations as the output from the Flatten layer (a 20,480-dimensional vector for each sequence), and performed a UMAP [46] (non-linear) projection with 3 components (see Section 4), coloring each point according to its corresponding sequence OGT (Figure 2g) and T_{opt} value (Figure 2h), respectively. The structure (sorting) observed in the projection illustrates that the pre-trained layers of DeepET (unchanged in the FrozenCNN transfer learning model predicting T_{opt}) have learned to generally separate sequences by thermal adaptation (thermophiles appearing to have the best separation, while sequences with lower temperature values having poorer separation). That OGT and T_{opt} are correlated[3] is reflected in the rather good sorting of the latter values (Figure 2h), even without the fine-tuning of the dense output layers (FrozenCNN), illustrating the validity of transfer learning between these two value domains.

2.3 | Interpreting the sequence determinants of thermostability

Finally, we tested the learned predictive representation of enzyme thermal adaptation by querying the pre-trained DeepET model to identify the specific parts of the protein sequences that were most predictive of optimal catalytic temperature (T_{opt}). Given that the best performing T_{opt} model FrozenCNN was only a slight modification of the pre-trained network on OGT, we were interested to see what type of information the pre-trained model had learned towards predicting the protein-relevant T_{opt} value, in addition to the organism-relevant OGT on which it was trained. For this, we used a perturbation procedure to evaluate the *relevance* of each specific sequence position in relation to the predicted value. Namely, for each protein, we occluded sliding windows of five amino acids along its sequence and compared the predictions for all these occlusions with that of the original unoccluded sequence, thus producing a per-residue perturbation or *relevance profile* for each protein[47,48] (Section 4). The occlusion width was arbitrary, set to match short secondary structure lengths. However, different widths, up to 20 amino acids, produced no effective differences in results (Section 4). The occluded parts of the input protein sequences that yielded a significant deviation in prediction (exceeding ±2 standard deviations) from the original were regarded as the most relevant for T_{opt} prediction (Figure 3a). We then checked these relevance profiles against sequence-specific properties that might be salient for prediction: amino acid composition, secondary structure, and protein domains. Among possible features to explore in this way, these were the most straightforward and previously characterized.

Our perturbation approach was deemed an appropriate model-agnostic way to probe the model for learned patterns, both bypassing the problematic interpretation of the black box CNN inner values, as well as accounting for all residues in the sequence. In contrast, Yu et al.[29] relied on prior knowledge and mutated a few specific amino acid triplets (key collagen monomer constituents) along the length of the sequence in order to determine the most impactful position and substitution. Their results successfully recapitulated known collagen thermostability factors, but also contrasted some conclusions of in vivo studies.

In terms of amino acid composition, the most relevant towards T_{opt} DeepET prediction for mesophiles (OGT 20–45°C) were Met, Ser, Leu, Ala, Trp, Phe, His, Cys (of which 6/8 are hydrophobic), and conversely, the least relevant were Glu, Pro, Asp, Lys, Asn, Tyr, Gln (of which 6/7 are polar). For thermophiles (OGT > 45°C), the most relevant were Lys, Glu, Met, Ile, Leu, Arg, Phe, Tyr (5 hydrophobic, 3 polar) and the least relevant were Pro, Thr, Ser, Gln, Ala, Cys, His, Asn, Asp (of which 6/9 are polar, see Figure 3a and Section 4). These enrichments and depletions are in line with known results[49] supporting observations that the amounts of (uncharged polar) Cys, Gln, Ser, and Thr are less frequent in thermophiles compared to mesophiles and thus decrease with OGT, while Arg and Tyr increase. Perhaps unintuitively, we find here that Ala relevance decreases between mesophiles and thermophiles, although higher occurrence was noted in thermophiles.[50] There is also some agreement with Zeldovich et al.[21] in terms of amino acids whose fraction is most correlated with OGT, though we only observe Ile, Tyr, Arg, Glu, and Leu as enriched for our thermophile set. Overall, hydrophobic amino acids generally appear more informative towards prediction, while polar ones are least informative, for both mesophilic and thermophilic groups, which corroborates protein hydrophobicity as an indicator of thermal adaptation.[49] As a commonality between mesophiles and thermophiles, we found that Met, Leu, and Phe occurrence (in decreasing order of enrichment for both) to be determinant for predictions.

In terms of secondary structure (as predicted per-residue by DSSP from PDB files), the most relevant towards prediction for mesophiles were strands (E), 3–10 helices (G), bends (S), and coils (−) (Figure 3b). Least relevant were α-helices (H), π-helices (I), turns (T), and isolated β-bridge residues (B). For thermophiles, most relevant were α-helices (H) and turns (T). Thus, while for mesophiles all major secondary structure types are observed to factor towards prediction, only helices and turns are the most determinant for thermophiles. This is in line with the known increase in helical content with higher temperature adaptations,[51] due to its importance in stabilization. The increase in relevance of helices also fits with the enrichment of Arg and depletion of Cys, His, and Pro relevance for thermophile prediction observed here, as these are known to be favored and disfavored for helix formation, respectively.[51]

To assess whether certain protein domains are more salient for prediction, we measured the overlap of the relevance profiles with domains from the InterPro database,[52] as the fraction of domain positions covered by significant (absolute z-score > 2) perturbation values. Control consisted of measuring the coverage outside the

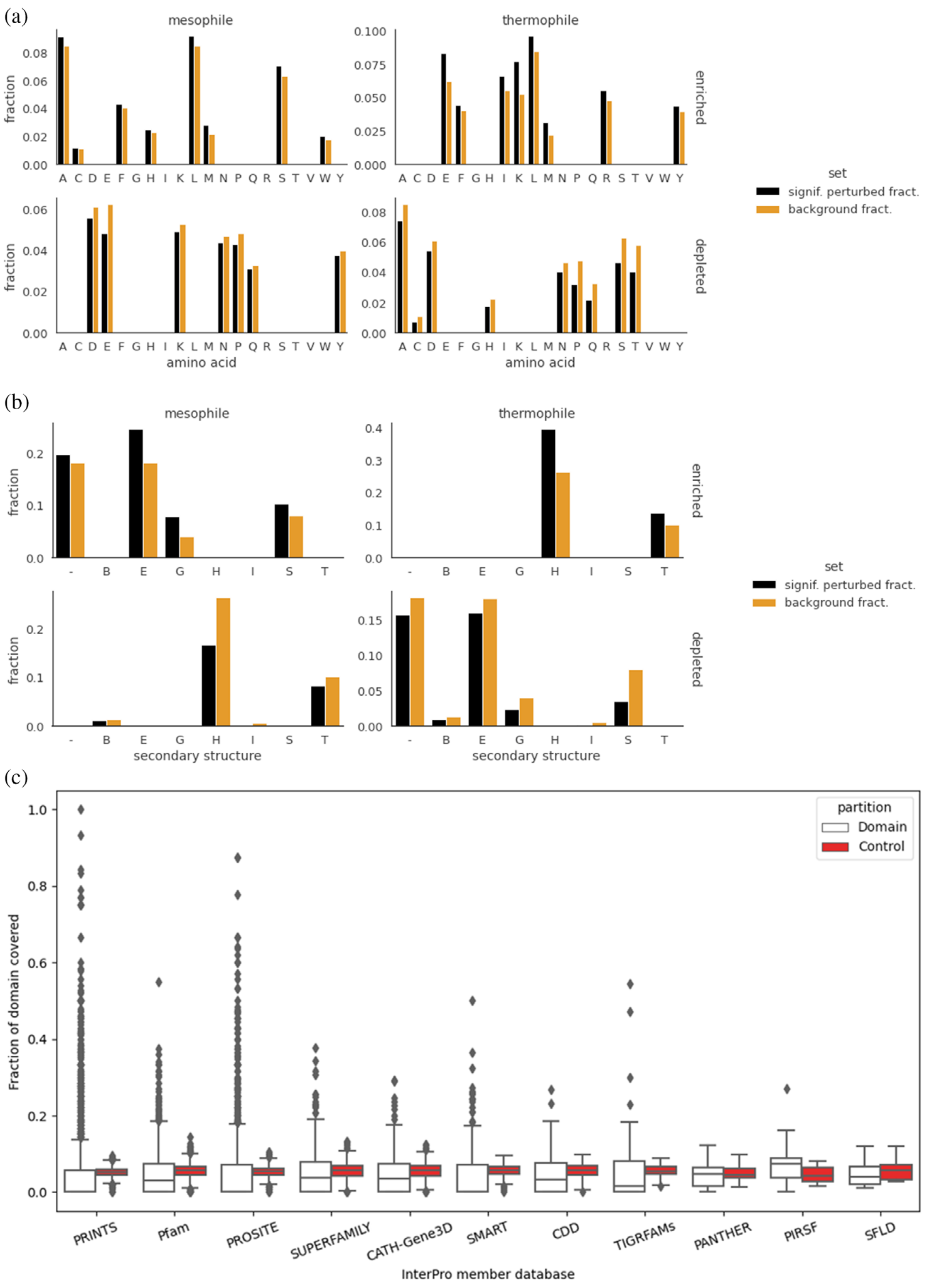

FIGURE 3 The determinants of thermostability. (a) Enriched and depleted amino acids in the perturbation profiles of mesophiles and thermophiles, showing the most relevant and least relevant amino acids, respectively, toward T_{opt} prediction. The fractions of amino acids found at significantly (absolute z-score > 2) perturbed positions are compared against the background amino acid distribution. (b) Enriched and depleted secondary structures (in DSSP notation) in the perturbation profiles of mesophiles and thermophiles, showing the most relevant and least relevant structures, respectively, towards T_{opt} prediction. The fractions of DSSP-annotated secondary structures found at significantly (absolute z-score > 2) perturbed positions are compared against the background amino acid distribution. In DSSP notation, B = isolated β-bridge residues, E = extended strands (parallel or antiparallel β-sheet), G = 3–10 helix, H = α-helix, I = π-helix, S = bend, T = turn, − = coil. (c) Fraction of protein domains covered by significantly perturbed positions, over InterPro domain databases (7,565 domains, 1,227 proteins). PRINTS, PROSITE, and Pfam member databases especially have many domains with high coverage. In total, 219 domains had a coverage of at least 30% across all member databases

domain, by dividing the outside sequence into windows with the same length as the domain, then taking the average coverage across these (Section 4). To ensure proper control, the search was limited to domains no longer than half of the protein sequence, a total of 7,565 domains across 1,227 proteins. While the median relevance profile coverage of domains did not greatly differ from control, the distributions of domain coverage fractions were quite wide and heavy-tailed across the InterPro member databases, with many domains clearly higher than control (Figure 3c). In total, 219 domains had a coverage of at least 30% (a cutoff chosen to be distinctly larger than any control). To get an overview of these domains, we took the GO terms associated with the protein domains (retrieved from InterPro) and produced GO slims using the Generic GO Subset (Section 4). While the GO slims for mesophilic enzymes spanned a diverse range of biological processes, including metabolic processes, stress response, protein transport, immune involvement, and cell adhesion, thermophile terms were limited to metabolic processes and response to stress (Tables S3 and S4). Thus, with increased temperature adaptation, the domains responsible for these latter functions are more determinative for the prediction of T_{opt}.

To validate the enriched sequence features thus detected as well as to estimate a lower bound of their predictivity, we trained linear and random forest models on amino acid composition, secondary structure composition, as well the combination of both. The highest performance (R^2 = 0.36, RMSE = 16) was obtained for a random forest model trained on the combination of both enriched factors (Figure S6). The models were trained and evaluated on the same train-test split of the TM data set as the deep models (Figure 2). That amino acid composition and secondary structure informs thermal adaptation was previously established [21,26,51] and the former was also quantified in this study using iFeature composition variables. Interestingly, for the relatively simple random forest models used in the validation, the performance on the subset of enriched amino acid composition was better than the entire set of amino acids (R^2 = .29 vs. .22, respectively). The gap between the performance of these compositional features and that of DeepET evidences the complex sequence relationships learned by our model.

3 | DISCUSSION

Here we presented DeepET, a deep learning model that learns temperature-related representations of protein sequences. We demonstrated that these representations are highly useful for the prediction of enzyme catalytic temperature optima and protein melting temperatures, by using a transfer learning approach. Our base model was trained to predict OGT with a high R^2 from 3 million enzyme sequences across all three domains of life. The model was then re-purposed via fine-tuning to predict optimal enzyme catalytic temperature, as well as melting temperature, both of which showed good performance. As the base DeepET model was trained from sequence alone, the transfer approach is more applicable than previous deep learning approaches for optimal catalytic temperature prediction,[26] which rely on extracted sequence features (amino acid composition) and OGT as input, the latter of which may not be available.

Further improvements to DeepET could be done in terms of data rebalancing to mitigate bias in the source or target data sets of the transfer learning setup, both at data level and at algorithmic level (assuming sufficient computational resources). Examples of such methods come from various fields, for example, the use of data augmentation and resampling for the in-domain transfer learning task of predicting lake chlorophyll concentration from satellite images (where water samples are sparse).[53] As discussed in recent work on bias mitigation in a transfer learning setting for large natural language processing models, care must be taken to avoid transferring bias to downstream tasks. Ideally, this could be largely handled in the upstream task, to provide a readily usable "off the shelf" model, lowering the effort threshold for applications. In some cases, upstream (source) mitigation is indeed sufficient, as shown in the work by Jin et al.,[54] though in other situations, it is the downstream mitigation that is most impactful, as illustrated by Steed et al.[45] This variation is perhaps not surprising, given the diversity of tasks and data, and further underlines the importance of such quality control measures in future iterations of models such as DeepET.

The good performance of DeepET on the transfer learning prediction tasks suggests that the representations indeed capture and provide a statistical summary of the enzyme thermal adaptation strategies from nature. To get insights into the sequence factors that are informative for the prediction of optimal catalytic temperature, we performed a perturbation analysis by exhaustively occluding sequences with a sliding window to measure the impact on the predicted value, then analyzing the properties of the most relevant sequence positions thus perturbed. We found a large overlap with known determinants of thermostability in terms of amino acid composition and secondary structure, both generally and when distinguishing between mesophiles and thermophiles. Moreover, the composition of enriched amino acids and secondary structures yielded modestly predictive random forest models, which is evidence that the

DeepET network has learned more complex relationships encoded in sequence, of which the features detected through the perturbation analysis were lower-dimensional projections. This gives greater confidence in the quality and general applicability of the learned DeepET features through transfer learning or data mining for sequence properties and patterns. Checking the relevant positions against protein domains, we saw that while the associated biological processes of domains present in mesophiles covered a wider range, domains thus found in thermophiles were limited to metabolic processes and response to stress, hinting at the adaptations of these enzymes for higher temperatures. As the field of ML interpretation expands in both theory and software availability, different avenues of feature identification may be pursued besides our black-box perturbation approach. A promising example is the integrated gradients method[55] (available currently only in the PyTorch-based package *captum*), as illustrated in the recent study by Kaminki et al.[56]

Given these recapitulations of known primary and secondary protein structure determinants of thermostability by DeepET's features, which were learned by the model from sequence alone, and the observed shift in model-relevant domains between mesophiles and thermophiles, the use of the DeepET is a promising avenue towards elucidating the physical mechanisms that convey enzymes resistance to extreme temperatures. Future work will therefore focus on further interpreting DeepET and its learned representations both using in silico analyses, and in a biological context, to deepen our understanding of enzyme thermal adaptation.

4 | MATERIALS AND METHODS

4.1 | The OGT data set with OGT-labeled enzyme sequences

About 6,270,107 enzyme sequences with unique Uniprot IDs were collected from the previous study.[32] After the removal of sequences that were (a) longer than 2,000; or (b) shorter than 100; or (c) with any non-standard amino acids, there were 6,141,006 enzyme sequences left. Then the cd-hit algorithm (−c 0.95, −T 20, −M 0, and other parameters as default)[57] was applied to cluster those sequences into 3,016,273 clusters. Only the representative sequence of each cluster was used for the next step, to keep the resulting sequences diverse. At last, 768 sequences were removed since they were present in the T_{opt} data set (see next section) by matching Uniprot IDs. In the end, a data set with 3,015,505 enzyme sequences from microorganisms with known OGTs was obtained. The data set was randomly split into training (2,864,729 enzymes) and test (150,776 enzymes) data sets based on a 95–5 ratio (see data distribution in Figure S9).

4.2 | The TOPT data set with enzyme optimal catalytic temperatures (T_{opt})

This data set was taken from Li et al.,[26] which contains 1,902 enzymes with known T_{opt} collected from BRENDA.[58] The data set was randomly split into training (1,712 enzymes) and test (190 enzymes) data sets based on a 90–10 ratio (Figure S9).

4.3 | The TM data set with protein melting temperatures (T_{m})

Leuenberger et al.[5] experimentally measured melting temperatures for more than 8,000 proteins from four species (*E. coli, S. cerevisiae, T. thermophilus*, and human cells). In this study, 2,506 proteins from three microorganisms (*E. coli*, *S. cerevisiae*, *T. thermophilus*) with experimentally measured T_{m} were obtained, after removal of ones with sequences that were (a) longer than 2,000; or (b) shorter than 100; or (c) with any non-standard amino acids. The data set was randomly split into training (2,255 enzymes) and test (251 enzymes) data sets based on a 90–10 ratio (Figure S9).

4.4 | The MELT data set with protein melting temperatures (T_{m})

The *Meltome* data set published by Jarzab et al.[7] contains melting temperatures for 48,000 proteins from 13 species, ranging from archaea to human. We first collected all T_{m} values from all 77 data sets and corresponding sequence IDs therein. Only proteins with an existing UniProt ID and protein sequence in the Uniprot database were considered. After removal of sequences that were (a) longer than 2,000; or (b) shorter than 100; or (c) with any non-standard amino acids, a data set with 41,725 proteins was obtained. For those proteins with multiple T_{m} values, the mean value was used. The data set thus processed was named MELT. The data set was randomly split into training (37,552 enzymes) and test (4,173 enzymes) data sets based on a 90–10 ratio (Figure S9).

4.5 | Deep neural networks

In the present study, we used Residual networks,[33] with the model architecture (Figure S1) similar to those that had been applied to protein functional annotation previously.[34] It contains 1–3 residual block(s) followed by two fully connected (FC) layers (Figure S1b). Batch Normalization[59] was applied after all layers; Weight dropout[60] was applied after FC layers and max-pooling[61] was applied after the last residual blocks. The Adam optimizer[62] with mean squared error loss function and ReLU activation function[63] with uniform[64] weight initialization were used.

4.6 | Hyper-parameter optimization

Two small OGT data sets with 10,000 samples each were used for tuning hyper-parameters: (a) the first one was randomly sampled from the OGT training data set (Figure S2a); (b) the second one was sampled in a way that the resulting samples showed a uniform OGT distribution (Figure S2b). Each data set was randomly split into training (90%) and validation (10%) data sets. The hyper-parameters were tuned using values randomly sampled from the defined parameter spaces (Table S1). Around 100–200 parameter sets were randomly sampled and tested. The best hyper-parameter set was chosen based on the one with the lowest validation loss on each small OGT data set. Then the model with these two hyper-parameter sets was tested with the big OGT training data set (2,864,729 enzymes). This data set was further split into training (95%) and validation (5%) data sets. After manually tuning a few hyper-parameters, the best model with the lowest validation loss was chosen as the final hyper-parameter set (Figure S3).

4.7 | Feature extraction for enzymes in T_{opt} and two protein T_{m} data sets

A set of 5,494 rationally designed features was extracted with iFeature.[38] These features included k-mer compositions (AAC, 20 features; DPC, 400), composition of k-spaced amino acid pairs (CKSAAP, 2400), dipeptide deviation from expected mean (DDE, 400), grouped amino acid composition (GAAC, 5), composition of k-spaced amino acid group pairs (CKSAAGP, 150), grouped dipeptide composition (GDPC, 25), grouped tripeptide composition (GTPC, 125), Moran autocorrelation (Moran, 240), Geary autocorrelation (Geary, 240), normalized Moreau-Broto (NMBroto, 240), composition-transition-distribution (CTDC, 39; CTDT, 39; CTDD, 195), conjoint triad (CTriad, 343), conjoint k-spaced triad (KSCTriad, 343), pseudo-amino acid composition (PAAC, 50), amphiphilic PAAC (APAAC, 80), sequence-order-coupling number (SOCNumber, 60), and quasi-sequence-order descriptors (QSOrder, 100).

4.8 | UniRep

A representation with 1900 × 3 features was extracted for each protein sequence with the previously published deep learning model UniRep,[39] which is a Multiplicative Long-Short-Term-Memory (mLSTM) Recurrent Neural Network that was trained on the UniRef50 data set.[65]

4.9 | Supervised classical ML methods

Two linear regression algorithms BayesianRidge and Elastic Net as well as three non-nonlinear algorithms Decision Tree, Random Forest, and Support Vector Machine were evaluated on each feature set (iFeatures and UniRep). Input features were firstly scaled to a standard normal distribution by $x_{N,i} = \frac{x_i - u_i}{\sigma_i}$, where x_i is the values of feature i of all samples, u_i and σ_i are the mean and standard deviation of x_i, respectively. This was done by taking all samples, including train and test data sets together. The training data set was further randomly split into training and validation data sets. The validation data set was used to tune the hyper-parameters via a greedy search approach. The optimized model was tested on the held out test data set and the R^2 score was calculated. In Figure 3a,b, the R^2 score of three best regression models were shown for iFeature and UniRep. All ML analyses in this section were performed with scikit-learn (v0.20.3)[66] using default settings.

4.10 | UMAP projection of deep sequence representations

The TOPT data set was filtered on sequences of maximum length 2000, that had both OGT and values, yielding 1,040 sequences. The Flatten layer outputs of the pre-trained DeepET model were collected for these (a set of 20,480-dimensional vectors), and a three-component UMAP[46] nonlinear projection was fitted using the umap-lean 0.5.3 Python package, with parameters *n_neighbors* = 15 and *min_dist* = 0.1 (using the default Euclidean distance). All other parameters were left as default. The perspective of the 3D scatter plot (elevation and azimuth) was chosen to give the best view of the overall point cloud. The related nonlinear projection method t-SNE[67] was also employed (using the implementation in the scikit-learn 1.0.1 Python package, with PCA initialization and 1,000 iterations) but did not yield a clear sorting or clustering of points for various *perplexity* values.

4.11 | Relevance profile analysis

The sequence perturbation study was performed on a set of 1,554 enzymes, a subset of the TOPT data set. Relevance profiles were obtained for each sequence by sliding a 5-amino-acid-long occlusion window on the sequence, 1 amino acid at a time (thus resulting in overlapping windows). For each window position, a T_{opt} prediction was obtained and the perturbation or relevance score was calculated as $(prediction_{occluded} - prediction_{wt})/prediction_{wt}$. As the sliding window position was bounded by the sequence, to obtain a perturbation vector of the same length as the sequence (and thus, a relevance score for each amino acid), the sequence was flanked by two repeats of the terminal amino acids. A moving average was then performed on the resulting relevance score vector for each protein. The width of the occlusion window was chosen to be small and match short secondary structure feature lengths. The impact of the occlusion width was tested by performing the perturbation procedure with windows of length 2, 5, 10, and 20, which respectively represent 0.42%, 1%, 2%, and 4.2% of the average sequence length in the set, 477. The resulting profiles showed large overlaps (Figure S7) and the choice of width had no impact on the resulting set of significantly covered protein domains (Figure S8).

The amino acid enrichment of relevance profiles was assessed by performing one-sided hypergeometric tests between the background amino acid counts in all sequences and the counts of amino acids occurring at significant (absolute z-score > 2) relevance profile positions, to test for both overrepresentation (enrichment) and underrepresentation (depletion) of amino acids. A p value threshold of .05 was set for significance. Cryophiles (OGT < 20°C, 5 sequences) were excluded due to very low counts. The remaining set included 1,220 mesophiles and 323 thermophiles. An analogous procedure was performed to assess the relevance of secondary structure, starting from per-residue sequence annotations obtained with DSSP 3 from PDB files. Due to either lack of PDB entries or structural errors within the files, the structural annotation set only included 874 mesophiles and 279 thermophiles. As a sanity check, positions where no annotation was available appeared as underrepresented (depleted) and were removed from results.

The InterPro database (retrieved June 24, 2021) was filtered to only domains at most half of the length of the protein, to ensure balance when performing control. The per-domain control consisted in taking the sequence S_{out} outside of a given domain and counting the number of significantly perturbed (absolute z-score > 2) positions. This number was divided by the number of windows in S_{out} of the same length as the domain, to give an expected count corresponding to repeatedly randomly sampling subsequences the same length as the domain. The final control coverage fraction was taken as the above average divided by the domain length.

GO slims were produced starting from the GO terms provided for each domain in the InterPro database and the Generic GO Subset provided by the GO Consortium (version August 21, 2021).[68,69] We selected domains that had at least 30% of their length covered by significantly perturbed (absolute z-score > 2) positions. The processing was performed using the Python packages GOATOOLS 1.1.6[70] and obonet 0.3.0. The full list of terms of GO slims is given in Tables S3 and S4.

4.12 | Software

Python v3.6 (www.python.org) scripts were used for the computations and data analysis, using the packages NumPy 1.18.1,[71] SciPy 1.6.2,[72] tensorflow 1.14,[73] keras 2.2.4, Biopython 1.76,[74] and PySpark 3.1.2. The code and data are available at Zenodo (https://doi.org/10.5281/zenodo.6351465).

AUTHOR CONTRIBUTIONS

Gang Li: Conceptualization (equal); data curation (equal); formal analysis (equal); investigation (equal); methodology (equal); resources (equal); software (equal); validation (equal); visualization (equal); writing – original draft (equal). **Filip Buric:** Formal analysis (equal); investigation (equal); methodology (equal); software (equal); validation (equal); visualization (equal); writing – original draft (equal); writing – review and editing (equal). **Jan Zrimec:** Investigation (equal); methodology (equal); writing – original draft (equal); writing – review and editing (equal). **Sandra Viknander:** Investigation (equal); methodology (equal); software (equal). **Jens Nielsen:** Funding acquisition (equal); supervision (equal). **Aleksej Zelezniak:** Funding acquisition (equal); methodology (equal); supervision (equal). **martin engqvist:** Conceptualization (equal); funding acquisition (equal); methodology (equal); project administration (equal); resources (equal); supervision (equal); writing – original draft (equal); writing – review and editing (equal).

ACKNOWLEDGMENTS

GL and JN have received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie program, project PAcMEN (Grant agreement no. 722287). JN also acknowledges funding from the Novo Nordisk Foundation (Grant no. NNF10CC1016517), the Knut and Alice Wallenberg Foundation. The study was supported by SciLifeLab funding and the Swedish Research Council (Vetenskapsrådet) starting Grant no. 2019-05356. AZ was supported by the Marius Jakulis Jason Foundation and JZ by the Slovenian Research Agency (ARRS) Grant no. J2-3060 and Public Scholarship, Development, Disability and Maintenance Fund of the Republic of Slovenia Grant no. 11013-9/2021-2. The computations were performed on resources at Chalmers Center for Computational Science and Engineering (C3SE) provided by the Swedish National Infrastructure for Computing (SNIC). Mikael Öhman and Thomas Svedberg at C3SE are acknowledged for technical assistance.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The data underlying this article were deposited to the Zenodo repository and are available at https://doi.org/10.5281/zenodo.6351465.

ORCID

Gang Li https://orcid.org/0000-0001-6778-2842
Filip Buric https://orcid.org/0000-0003-0991-9040
Jan Zrimec https://orcid.org/0000-0002-7099-961X
Sandra Viknander https://orcid.org/0000-0003-1809-8627
Jens Nielsen https://orcid.org/0000-0002-9955-6003
Aleksej Zelezniak https://orcid.org/0000-0002-3098-9441
Martin K. M. Engqvist https://orcid.org/0000-0003-2174-2225

REFERENCES

1. Rothschild LJ, Mancinelli RL. Life in extreme environments. Nature. 2001;409:1092–1101.
2. Hickey DA, Singer GAC. Genomic and proteomic adaptations to growth at high temperature. Genome Biol. 2004;5:117.
3. Engqvist MKM. Correlating enzyme annotations with a large set of microbial growth temperatures reveals metabolic adaptations to growth at diverse temperatures. BMC Microbiol. 2018;18:177.
4. Fersht AR, Daggett V. Protein folding and unfolding at atomic resolution. Cell. 2002;108:573–582.
5. Leuenberger P, Ganscha S, Kahraman A, et al. Cell-wide analysis of protein thermal unfolding reveals determinants of thermostability. Science. 2017;355:eaai7825.
6. Zakhartsev M, Yang X, Reuss M, Pörtner HO. Metabolic efficiency in yeast Saccharomyces cerevisiae in relation to temperature dependent growth and biomass yield. J Therm Biol. 2015;52:117–129.
7. Jarzab A, Kurzawa N, Hopf T, et al. Meltome atlas-thermal proteome stability across the tree of life. Nat Methods. 2020;17:495–503. https://doi.org/10.1038/s41592-020-0801-4.
8. Nguyen V, Wilson C, Hoemberger M, et al. Evolutionary drivers of thermoadaptation in enzyme catalysis. Science. 2017;355:289–294.
9. Savitski MM, Reinhard FBM, Franken H, et al. Tracking cancer drugs in living cells by thermal profiling of the proteome. Science. 2014;346:1255784.
10. Arnold FH, Wintrode PL, Miyazaki K, Gershenson A. How enzymes adapt: Lessons from directed evolution. Trends Biochem Sci. 2001;26:100–106.
11. Mateus A, Bobonis J, Kurzawa N, et al. Thermal proteome profiling in bacteria: Probing protein state. Mol Syst Biol. 2018;14:e8242.
12. Sawle L, Ghosh K. How do thermophilic proteins and proteomes withstand high temperature? Biophys J. 2011;101:217–227.
13. Daniel RM, Danson MJ. A new understanding of how temperature affects the catalytic activity of enzymes. Trends Biochem Sci. 2010;35:584–591.
14. Hobbs JK, Jiao W, Easter AD, Parker EJ, Schipper LA, Arcus VL. Change in heat capacity for enzyme catalysis determines temperature dependence of enzyme catalyzed rates. ACS Chem Biol. 2013;8:2388–2393.
15. Venev SV, Zeldovich KB. Thermophilic adaptation in prokaryotes is constrained by metabolic costs of proteostasis. Mol Biol Evol. 2018;35:211–224.
16. Chen K, Gao Y, Mih N, O'Brien EJ, Yang L, Palsson BO. Thermosensitivity of growth is determined by chaperone-mediated proteome reallocation. Proc Natl Acad Sci U S A. 2017;114:11548–11553.
17. Li G, Hu Y, Wang H, et al. Bayesian genome scale modelling identifies thermal determinants of yeast metabolism. bioRxiv. 2020. https://doi.org/10.1038/s41467-020-20338-2
18. Chang RL, Andrews K, Kim D, Li Z, Godzik A, Palsson BO. Structural systems biology evaluation of metabolic thermotolerance in *Escherichia coli*. Science. 2013;340:1220–1223.
19. Singer GAC, Hickey DA. Thermophilic prokaryotes have characteristic patterns of codon usage, amino acid composition and nucleotide content. Gene. 2003;317:39–47.
20. Gromiha MM, Oobatake M, Sarai A. Important amino acid properties for enhanced thermostability from mesophilic to thermophilic proteins. Biophys Chem. 1999;82:51–67.
21. Zeldovich KB, Berezovsky IN, Shakhnovich EI. Protein and DNA sequence determinants of thermophilic adaptation. PLoS Comput Biol. 2007;3:e5.
22. Zhao N, Pang B, Shyu C-R, Korkin D. Charged residues at protein interaction interfaces: Unexpected conservation and orchestrated divergence. Protein Sci. 2011;20:1275–1284.
23. Szilágyi A, Závodszky P. Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: Results of a comprehensive survey. Structure. 2000;8:493–504.
24. England JL, Shakhnovich BE, Shakhnovich EI. Natural selection of more designable folds: A mechanism for thermophilic adaptation. Proc Natl Acad Sci U S A. 2003;100:8727–8731.
25. Khan S, Vihinen M. Performance of protein stability predictors. Hum Mutat. 2010;31:675–684.

26. Li G, Zrimec J, Ji B, et al. Performance of regression models as a function of experiment noise. arXiv [q-bioBM]. 2019. https://doi.org/10.1177/11779322211020315
27. Gado JE, Beckham GT, Payne CM. Improving enzyme optimum temperature prediction with resampling strategies and ensemble learning. J Chem Inf Model. 2020;60:4098–4107.
28. Min S, Kim H, Lee B, Yoon S. Protein transfer learning improves identification of heat shock protein families. PLoS One. 2021;16:e0251865.
29. Yu C-H, Khare E, Narayan OP, Parker R, Kaplan DL, Buehler MJ. ColGen: An end-to-end deep learning model to predict thermal stability of de novo collagen sequences. J Mech Behav Biomed Mater. 2022;125:104921.
30. Yang KK, Lu AX, Fusi N. Convolutions are competitive with transformers for protein sequence pretraining. bioRxiv. 2022; 2022.05.19.492714. https://doi.org/10.1101/2022.05.19.492714.
31. Zhang Y, Guan F, Xu G, et al. A novel thermophilic chitinase directly mined from the marine metagenome using the deep learning tool Preoptem. Bioresour Bioprocess. 2022;9:1–14.
32. Li G, Rabe KS, Nielsen J, Engqvist MKM. Machine learning applied to predicting microorganism growth temperatures and enzyme catalytic optima. ACS Synth Biol. 2019;8:1411–1420.
33. He K, Zhang X, Ren S, Sun J. Identity mappings in deep residual networks. In: Leibe B, Matas J, Sebe N, Welling M, editors. Computer vision – ECCV. Volume 9908. New York: Springer International Publishing, 2016; p. 630–645.
34. Bileschi ML, Belanger D, Bryant D, et al. Using deep learning to annotate the protein universe. bioRxiv. 2019;626507. https://doi.org/10.1101/626507.
35. Tan C. Sun F, Kong T, Zhang W, Yang C, Liu C. A survey on deep transfer learning. in Artificial neural networks and machine learning – ICANN 2018 270–279. New York: Springer International Publishing.
36. Pesciullesi G, Schwaller P, Laino T, Reymond J-L. Transfer learning enables the molecular transformer to predict regio- and stereoselective reactions on carbohydrates. Nat Commun. 2020;11:4874.
37. Chaudhari M, Thapa N, Ismail H, et al. DTL-DephosSite: Deep transfer learning based approach to predict dephosphorylation sites. Front Cell Dev Biol. 2021;9:662983.
38. Chen Z, Zhao P, Li F, et al. iFeature: A Python package and web server for features extraction and selection from protein and peptide sequences. Bioinformatics. 2018;34:2499–2502.
39. Alley EC, Khimulya G, Biswas S, AlQuraishi M, Church GM. Unified rational protein engineering with sequence-based deep representation learning. Nat Methods. 2019;16:1315–1322. https://doi.org/10.1038/s41592-019-0598-1.
40. Heinzinger M, Elnaggar A, Wang Y, et al. Modeling aspects of the language of life through transfer-learning protein sequences. BMC Bioinform. 2019;20:723.
41. Rao R, Bhattacharya N, Thomas N, et al. Evaluating protein transfer learning with TAPE. Adv Neural Inf Process Syst. 2019;32:9689–9701.
42. Rives A, Meier J, Sercu T, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc Natl Acad Sci U S A. 2021;118: e2016239118.
43. Ng H-W, Nguyen VD, Vonikakis V, Winkler S. Deep learning for emotion recognition on small datasets using transfer learning. Proceedings of the 2015 ACM on international conference on multimodal interaction. New York: Association for Computing Machinery, 2015; p. 443–449.
44. Yang KK, Wu Z, Arnold FH. Machine-learning-guided directed evolution for protein engineering. Nat Methods. 2019;16:687–694.
45. Steed R, Panda S, Kobren A, & Wick, M. Upstream mitigation is not all you need: Testing the bias transfer hypothesis in pre-trained language models. In Proceedings of the 60th annual meeting of the association for computational linguistics (Volume 1: Long Papers), p. 3524–3542.
46. McInnes L, Healy J, Melville J. UMAP: Uniform manifold approximation and projection for dimension reduction. arXiv: 1802.03426 [statML]. 2018.
47. Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. Computer vision – ECCV. New York: Springer International Publishing, 2014; p. 818–833.
48. Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol. 2015;33:831–838.
49. Pezeshgi Modarres H, Mofrad MR, Sanati-Nezhad A. Protein thermostability engineering. RSC Adv. 2016;6:115252–115270.
50. Panja AS, Bandopadhyay B, Maiti S. Protein thermostability is owing to their preferences to non-polar smaller volume amino acids, variations in residual physico-chemical properties and more salt-bridges. PLoS One. 2015;10:e0131495.
51. Kumar S, Tsai CJ, Nussinov R. Factors enhancing protein thermostability. Protein Eng. 2000;13:179–191.
52. Blum M, Chang HY, Chuguransky S, et al. The InterPro protein families and domains database: 20 years on. Nucleic Acids Res. 2021;49:D344–D354.
53. Syariz MA, Lin CH, Heriza D, Lasminto U, Sukojo BM, Jaelani LM. A transfer learning technique for inland chlorophyll-a concentration estimation using Sentinel-3 imagery. NATO Adv Sci Inst Ser E Appl Sci. 2021;12:203.
54. Jin X, Barbieri F, Kennedy B, Davani AM, Neves L, Ren X. On transferability of bias mitigation effects in language model fine-tuning. arXiv:2010.12864 [csCL]. 2020.
55. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. Proceedings of the 34th international conference on machine learning (eds. Precup, D & Teh, Y. W.), Vol. 70, Cambridge, MA: Proceedings of Machine Learning Research (PMLR), 2017; p. 3319–3328.
56. Kamiński K, Ludwiczak J, Jasiński M, et al. Rossmann-toolbox: A deep learning-based protocol for the prediction and design of cofactor specificity in Rossmann fold proteins. Brief Bioinform. 2022;23:bbab371.
57. Li W, Godzik A. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22:1658–1659.
58. Jeske L, Placzek S, Schomburg I, Chang A, Schomburg D. BRENDA in 2019: A European ELIXIR core data resource. Nucleic Acids Res. 2019;47:D542–D549.
59. Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv: 1502.03167 [csLG]. 2015.
60. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A simple way to prevent neural networks from overfitting. United States: Journal of Machine Learning Research (JMLR); 2014.

61. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ, editors. Advances in neural information processing systems 25. Red Hook, NY: Curran Associates, Inc. 2012; p. 1097–1105.
62. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv:1412.6980 [cs.LG].
63. Nair V, Hinton GE. Rectified linear units improve restricted boltzmann machines. Proceedings of the 27th international conference on machine learning (ICML-10). Madison WI: Omnipress. 2010; p. 807–814.
64. He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. Proceedings of the IEEE international conference on computer vision. Cambridge, MA: IEEE, 2015; p. 1026–1034.
65. Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH, the UniProt Consortium. UniRef clusters: A comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics. 2015;31:926–932.
66. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in python. J Mach Learn Res. 2011;12:2825–2830.
67. Van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res. 2008;9(11):2579–2605.
68. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: Tool for the unification of biology. Nat Genet. 2000;25:25–29.
69. Gene Ontology Consortium. The gene ontology resource: Enriching a GOld mine. Nucleic Acids Res. 2021;49:D325–D334.
70. Klopfenstein DV, Zhang L, Pedersen BS, et al. GOATOOLS: A Python library for gene ontology analyses. Sci Rep. 2018;8:10872.
71. Harris CR, Millman KJ, van der Walt SJ, et al. Array programming with NumPy. Nature. 2020;585:357–362.
72. Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: Fundamental algorithms for scientific computing in Python. Nat Methods. 2020;17:261–272.
73. Abadi M, Agarwal A, Barham P, et al. TensorFlow: Large-scale machine learning on heterogeneous distributed systems. arXiv:1603.04467 [csDC]. 2016.
74. Cock PJA, Antao T, Chang JT, et al. Biopython: Freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009;25:1422–1423.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Li G, Buric F, Zrimec J, Viknander S, Nielsen J, Zelezniak A, et al. Learning deep representations of enzyme thermal adaptation. Protein Science. 2022;31(12):e4480. https://doi.org/10.1002/pro.4480