THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Deep generative models for analysis and engineering of functional proteins

SANDRA VIKNANDER

Department of Life Sciences

CHALMERS UNIVERSITY OF TECHNOLOGY

Gothenburg, Sweden 2025

Deep generative models for analysis and engineering of functional proteins

SANDRA VIKNANDER
ISBN: 978-91-8103-187-4

Doktorsavhandlingar vid Chalmers tekniska högskola.
Ny serie nr 5645
ISSN0346-718X

Department of Life Sciences
Chalmers University of Technology
SE-412 96 Gothenburg
Sweden
Telephone + 46 (0)31-772 1000

Cover:
Conceptual representation of the neural network-driven protein engineering process explored in this thesis, where an input protein is modified to enhance thermal resistance.

Printed by Chalmers Reproservice
Gothenburg, Sweden 2025

Deep generative models for analysis and engineering of functional proteins

SANDRA VIKNANDER

Department of Life Sciences

Chalmers University of Technology

Abstract

Proteins are essential biological molecules that sustain life through diverse functions, from structural support to catalyzing biochemical reactions. Their catalytic efficiency makes them invaluable for industrial applications, where they often require optimization to function under specific conditions. While experimental and computational approaches have made progress in protein engineering, no universal method exists due to the complexity of protein structure and function. Recent advances in machine learning offer new possibilities by leveraging vast protein sequence data. However, key challenges remain, including the limited availability and uneven distribution of high-quality labels describing essential properties like enzymatic activity and thermal stability. Addressing these issues is critical for developing models capable of accurate trait selection. My work focuses on two key steps in protein engineering: diversification and selection. To improve selection, deep learning models were developed using transfer learning, data augmentation, and protein language models (pLMs) to predict physical and functional properties such as melting temperature, enzymatic temperature, protein abundance, and *in vitro* activity. These models not only enable precise trait selection but also provide insights into the relationships between sequence, thermal adaptation, and conformational stability. For diversification, a deep generative model was created to capture natural sequence diversity and extend it to generate novel variant libraries across protein families. This approach prioritizes functional sequences and allows for targeted engineering of proteins with enhanced properties. Moving beyond general sequence generation, a framework was developed to create variant pools optimized for specific traits, such as increased thermal stability. By integrating these advancements, we engineered functional protein variants from diverse wild-type sequences, achieving up to a 36°C increase in melting temperature. This work highlights the potential of generative machine learning to refine and accelerate the protein engineering cycle, paving the way for more efficient and scalable biotechnological applications.

Keywords: protein engineering, thermal stability, machine learning, deep learning, generative AI

List of Publications

This thesis is based on the work contained in the following papers and manuscripts:

Paper I: Expanding functional protein sequence spaces using generative adversarial networks
Donatas Repecka, Vykintas Jauniskis, Laurynas Karpus, Elzbieta Rembeza, Irmantas Rokaitis, Jan Zrimec, Simona Poviloniene, Audrius Laurynenas, **Sandra Viknander**, Wissam Abuajwa, Otto Savolainen, Rolandas Meskys, Martin K. M. Engqvist & Aleksej Zelezniak. Nature Machine Intelligence Volume 3 Issue 4, 324–333 (2021)

Paper II: Learning deep representations of enzyme thermal adaptation
Gang Li, Filip Buric, Jan Zrimec, **Sandra Viknander**, Jens Nielsen, Aleksej Zelezniak, Martin K. M. Engqvist. Protein Science Volume 31 Issue 12 e4480 (2022)

Paper III: Amino acid sequence encodes for protein abundance shaped by protein stability at reduced synthesis cost
Filip Buric*, **Sandra Viknander***, Xiaozhi Fu, Oliver Lemke, Oriol Gracia Carmona, Jan Zrimec, Lukasz Szyrwiel, Michael Muelleder, Markus Ralser, Aleksej Zelezniak. Protein Science Volume 34, Issue 1 e5239 (2025)

Paper IV: Computational scoring and experimental evaluation of enzymes generated by neural networks
Sean R. Johnson, Xiaozhi Fu, **Sandra Viknander**, Clara Goldin, Sarah Monaco, Aleksej Zelezniak, Kevin K. Yang. Nature Biotechnology (2024) https://doi.org/10.1038/s41587-024-02214-2

Paper V: Learning Thermal Adaptation through Adversarial and Evolutionary aware training
Sandra Viknander, Nikolaos Tatarakis, Xiaozhi Fu, Clara Goldin, Alexander Diaciuc, Aleksej Zelezniak (*Manuscript*)

Additional papers and manuscripts not included in this thesis:

Paper VI: Structure-based clustering and mutagenesis of bacterial tannases reveals the importance and diversity of active site-capping domains
Tom Coleman, **Sandra Viknander**, Alicia M. Kirk, David Sandberg, Elise Caron, Aleksej Zelezniak, Elizabeth Krenske, Johan Larsbrink. Protein Science Volume 33 Issue 12 e5202 (2024)

Paper VII: The Role of Metabolism in Shaping Enzyme Structures Over 400 Million Years of Evolution Oliver Lemke, Benjamin Murray Heineike, **Sandra Viknander**, Nir Cohen, Jacob Lucas Steenwyk, Leonard Spranger, Feiran Li, Federica Agostini, Cory Thomas Lee, Simran Kaur Aulakh, Jens Nielsen, Antonis Rokas, Judith Berman, Aleksej Zelezniak, Toni Ingolf Gossmann, Markus Ralser. bioRxiv (2024)
https://doi.org/10.1101/2024.05.27.596037

Contribution Summary

Paper I: I conducted the analysis of the discriminator and interpreted the attention maps.

Paper II: I contributed equally to the investigation, methodology, formal analysis, and software development.

Paper III: I co-designed the study, performed the PaxDB analysis, designed the models used for PaxDB and their corresponding analyses, conducted molecular dynamics (MD) simulations, and contributed equally to the formal analysis of these simulations. I contributed to the manuscript in its entirety, especially to the aforementioned parts.

Paper IV: I trained and generated sequences using ProteinGAN and performed structure prediction and evaluation using AlphaFold2 metrics.

Paper V: I designed the study, collected, analyzed, and preprocessed the data, and developed the generative model, including its training and evaluation. I performed MD simulations and analysis, did most of the experimental design, and wrote the manuscript.

Preface

This dissertation serves as partial fulfillment of the requirements to obtain the degree of Doctor of Teknologie at the Department of Life Science at Chalmers University of Technology. The PhD studies were carried out between May 2020 and March 2025 at the division of Systems and Synthetic Biology under the supervision of Aleksej Zelezniak.

The research was mainly funded by SciLifeLab fellows program Swedish Research council (Vetenskapsrådet) starting grant no. 2019-05356, Formas early-career research grant 2019-01403 , WALP Wallenberg Launchpad project 2021.0198 supported by the Knut and Alice Wallenberg Foundation The computations and data handling was enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS) and the Swedish National Infrastructure for Computing (SNIC) at the Chalmers Center for Computational Science and Engineering (C3SE), the National Supercomputer Centre in Sweden (NSC) and at the High-Performance Computing Center North, partially funded by the Swedish Research Council through grant agreements no. 2022-06725 and no. 2018-05973. GPU-intensive computations were enabled by the Berzelius resource provided by the Knut and Alice Wallenberg Foundation at the National Supercomputer Centre.

List of Abbreviations

ASR:	Ancestral Sequence Reconstruction
BERT:	Bidirectional Encoder Representations from Transformers
CASTing:	Combinatorial Active-Site Saturation Testing
CNN:	Convolutional Neural Network
COG:	Cluster of Orthologous Groups
CuSOD:	Copper Superoxide Dismutase
DNN:	Deep Neural Network
DSF:	Differential Scanning Fluorimetry
DSSP:	Define Secondary Structure of Proteins (algorithm)
FACS:	Fluorescence-Activated Cell Sorting
GAN:	Generative Adversarial Network
GO:	Gene Ontology
GPU:	Graphics Processing Unit
LIME:	Local Interpretable Model-agnostic Explanations
MD:	Molecular Dynamics
MDH:	Malate Dehydrogenase
MELT:	Meltome Atlas
ML:	Machine Learning
MSA:	Multiple Sequence Alignment
NAD:	Nicotinamide Adenine Dinucleotide
NADH:	Nicotinamide Adenine Dinucleotide (reduced form)
NGS:	Next-Generation Sequencing
OGT:	Optimal Growth Temperature
PCR:	Polymerase Chain Reaction
RMSE:	Root Mean Square Error
RMSF:	Root Mean Square Fluctuation
SASA:	Solvent-Accessible Surface Area
STD:	STandard Deviation
UMAP:	Uniform Manifold Approximation and Projection for Dimension Reduction
WT:	Wild Type

Table of Contents

Table of Figures

Background

The Role of Enzymes in Industrial Applications

Humans have harnessed the power of microbes, such as bacteria for making yogurt and cheese and yeast for baking bread and brewing beer, since ancient times. And thus, unknowingly, we were leveraging the remarkable capabilities of enzymes to produce these foods. Today, while the basic principles remain the same, we utilize enzyme technologies on a vastly greater industrial scale. With the advent of biotechnology and the ability to make artificial or recombinant proteins, we have been able to expand the applications of enzymes far beyond traditional food production. Thanks to their unparalleled catalytic efficiency, high specificity, and enzymatic processes have lower environmental impact over traditional chemistry, making them indispensable in a wide range of industries. From food and agriculture to chemicals, detergents, medicine, and biofuel production, enzymes play a critical role in advancing sustainable and efficient processes. While we have been successful in a wide array of areas in utilizing enzymes in novel ways, enzymes do have some inherent limitations that have stopped their adoption in an even greater capacity. Their catalytic efficiency depends on their intricate three-dimensional structures, which arise from the precise folding of their amino acid chains. While simultaneously being the source of their versatility, these structures also make them sensitive to environmental factors such as temperature, pH, and salt concentration. While natural evolution enables such adaptations over long timescales, it cannot keep pace with the rapidly changing demands of modern industry. As a result, enzymes frequently need to be adapted or engineered to function reliably in these challenging settings. This disparity highlights the need for innovative approaches to protein engineering that are capable of tailoring enzymes to meet industrial requirements efficiently and effectively.

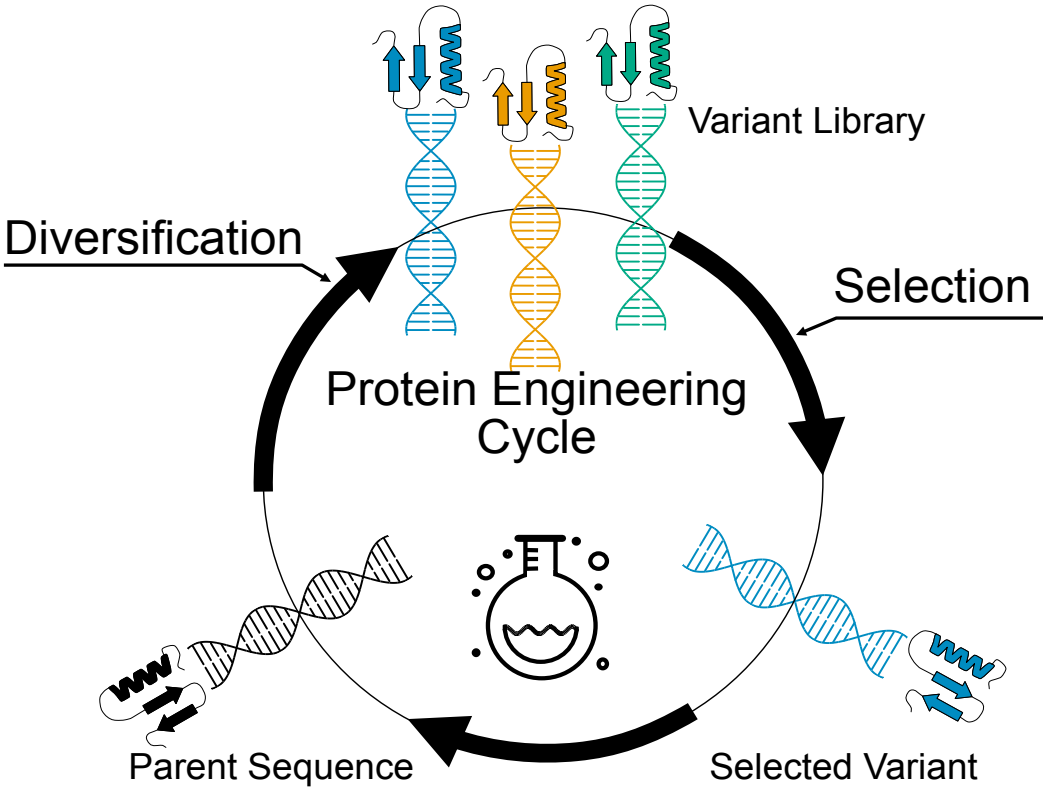

Figure 1: The protein engineering cycle.

Starting from a parent sequence, a variant library is generated through diversification. This library undergoes a selection phase where variants are evaluated, and the most fit candidates are identified. These selected variants are then used as a new template, and the cycle is repeated iteratively until a variant with the desired properties is achieved.

Proteins, though built from just a collection of 20 natural amino acids, exhibit an astounding diversity. The combinatorial possibilities of even a modest peptide chain of 100 residues exceed the number of particles in the known universe. Despite this staggering diversity, only an estimated 1 in 10^{77} such sequences is thought to fold into stable, functional structures, let alone act as effective enzymes[1–3]. Consequently, exhaustively searching the vast protein sequence space for functional enzymes through brute force is not only impractical but experimentally intractable. Fortunately, evolution has gifted us a vast collection of natural proteins that serve as valuable templates for engineering. By leveraging this evolutionary starting point, we can drastically narrow the search space and focus on modifying existing proteins to improve physicochemical characteristics or create novel functions. Nature's designs thus provide both inspiration and a foundational framework for optimizing and tailoring proteins to meet specific needs. The deliberate modification of protein structures to alter or enhance their function is referred to as protein engineering. While drawing inspiration from natural proteins significantly reduces the vast search space for potential modifications, the process remains highly complex. For instance, studies have shown that up to 70% of random

amino acid substitutions negatively impact protein function[4,5]. Moreover, achieving a desired phenotypic function or property often requires multiple coordinated substitutions, adding further complexity to the task.

As a result, protein engineering is often a labor-intensive and iterative process. Typically, this involves introducing several candidate substitutions into a pool of protein variants, followed by evaluating these variants for their fitness concerning the desired function or physicochemical property. The most promising variants are then selected for further refinement in subsequent rounds of engineering. This iterative cycle gradually optimizes the protein toward its intended function and characteristics. Figure 1 illustrates this iterative process. Starting with a template sequence, candidate substitutions are introduced into a variant pool. These variants are subjected to functional evaluation, and the best-performing variants are selected for further refinement, forming a continuous loop of design, evaluation, and optimization.

This iterative cycle of improvement, where a parent enzyme is diversified through substitutions and the fittest variants are selected, closely mirrors the natural phenomenon of evolution by natural selection. In nature, random mutations in parental genomes, coupled with the selective advantage conferred by beneficial mutations, have enabled microorganisms to evolve remarkable traits rapidly. Traits such as antibiotic resistance and the ability to metabolize nonnative herbicides and pesticides or even degrade some artificial polymers[6–9]. Directed evolution (DE) harnesses this evolutionary principle in a controlled laboratory setting. Mutations are introduced into the parent sequence using low-fidelity PCR cloning, generating diverse libraries of variants for the next generation[10]. The selection of improved variants depends on the phenotype of interest and the library requirements, with methods such as *in vivo* complementation of auxotrophic host strains, fluorescence-activated cell sorting (FACS), or microtiter plate screening commonly employed to identify the fittest candidates that will be used for the next iteration. While random mutagenesis has been successful in generating diverse libraries for protein engineering[11,12], the process inherently introduces a high proportion of neutral or, more importantly, deleterious mutations compared to beneficial ones. This accumulation of deleterious mutations limits the number of effective engineering cycles that can be conducted, as these harmful substitutions reduce the overall fitness of the variants with each iteration. To address this limitation, methods such as DNA shuffling have been developed. By fragmenting the genes of interest and reassembling the fragments, beneficial mutations can be combined across variants while neutral and deleterious substitutions are filtered out. This process effectively increases the number of engineering cycles that can be performed, facilitating the optimization of proteins and enhancing the likelihood of achieving the desired functional improvements[13]. However, even with the advancement of gene shuffling allowing for multiple iterations, the space of variants explored by DE remains small compared to the vast space enzymatic space. Given the sparsity of function, DE is not always capable of arriving at a desired outcome. In principle, the limitations of directed evolution could be addressed by generating ever-larger libraries to increase the likelihood of discovering

functional variants. However, practical constraints such as the efficiency of library screening and the vastness of the protein sequence space make this approach increasingly impractical.

In recent years, the field of protein engineering has shifted toward rational and semi-rational design strategies to refine the search space and focus on the most promising regions of sequence space. By leveraging insights from protein structure, sequence alignments, and evolutionary data, these approaches aim to predict which mutations are most likely to enhance function or stability. This targeted approach reduces the need for exhaustive random mutagenesis and instead directs efforts toward regions of the protein where modifications are more likely to yield beneficial outcomes.

One of the simplest and most effective ways to integrate evolutionary data into protein engineering is through multiple sequence alignments (MSAs). By analyzing conserved and variable residues across homologous proteins, engineers can identify positions with low conservation, which are more likely to tolerate or benefit from mutations. This principle forms the basis of Combinatorial Active-Site Saturation Testing (CASTing), a semi-rational approach that focuses mutagenesis on selected, functionally relevant residues. Despite its simplicity, CASTing has been shown to yield highly effective results in optimizing enzyme activity, stability, and specificity[14–16]. Beyond sequence-based methods, functional, structural, and physicochemical data can serve as additional modalities for guiding protein engineering. With these extra modalities, computational models such as Rosetta Design, YASARA, FoldX, and ABACUS[17–1920] can compute force fields and free energy between states to predict the effects of amino acid substitutions on protein stability and function. These *in silico* approaches can significantly reduce the number of variants that need to be tested, as many of the likely deleterious substitutions can be filtered out, accelerating the engineering cycle and improving overall efficiency[21,22].

Machine Learning and How It is Revolutionizing Biotechnology

The advent of modern high-throughput experimental methods in biology has led to an explosive growth of data. In the past, the scarcity of data necessitated a top-down approach, where hypotheses and rules were formulated from first principles and then validated through experiments. Today, the abundance of data enables a bottom-up approach, where we begin with the data itself and use statistical methods to uncover underlying relationships, leading to new insights. Machine learning (ML) techniques play a key role in extracting these relationships by fitting models to data. ML has found numerous applications in biotechnology, ranging from medical imaging, such as breast cancer detection[23,24], to drug discovery[25,26], molecular generation[27], and even solving the long-standing challenge of protein structure prediction[28]. These are just a few examples of a much broader trend in the application of ML in biotechnology.

Given the diversity of tasks and data types in biotechnology, it is valuable to gain an overview of different machine learning methodologies, their applications to biological data, and how

they are used throughout this thesis. Broadly, machine learning can be categorized into three major classes: discriminative models, unsupervised learning, and generative models.

Discriminative Models: learn a direct mapping between the input data X and a target variable Y, effectively modeling the conditional probability P(Y|X). When the target variable Y consists of discrete categories, the model is referred to as a classifier, learning a decision boundary that separates different classes. If Y is a continuous variable, the model performs regression, fitting a function that captures the relationship between X and Y (Figure 2, panel A). In *Papers I,V* we make use of discriminative models to classify real and generated sequences, which are then in turn used to train our generator models. In *Paper II,III,V* we are making use of regression models to fit T_m and protein abundance to sequence data.

Unsupervised Learning: can be used when labeled data is unavailable. These methods do not predict a specific output but instead identify patterns and structures within the data. A common approach is clustering, which groups data points based on their feature similarity. These groupings can then be used to infer potential classes or relationships within the dataset (Figure 2, panel B). In *Paper III* we make use of unsupervised learning when we project our data to a 1D manifold that groups sequence representations together.

Generative Modeling: Generative models aim to learn the joint probability distribution P(Y,X), allowing the model to generate new data points by sampling from the learned distribution. This enables the creation of new synthetic data instances conditioned on specific labels (Figure 2, panel C). In *Paper I,V* we trained such generative models to generate new proteins and thermostable variants.

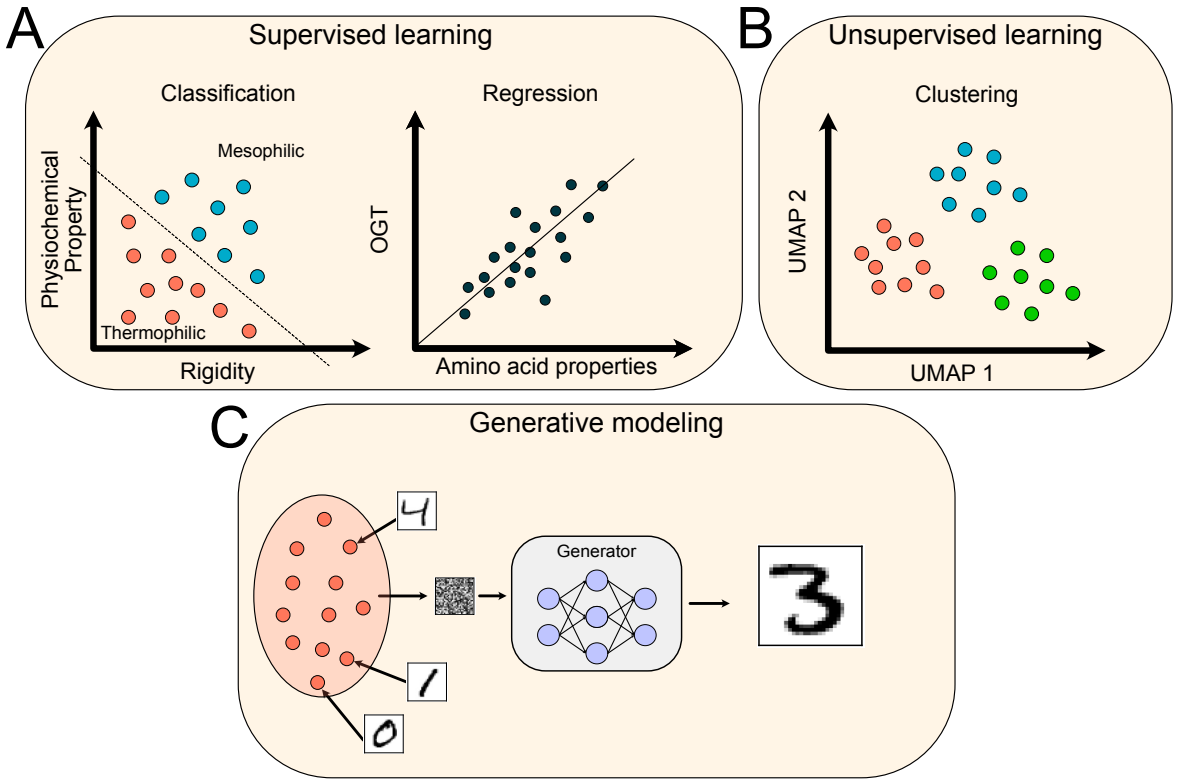

Figure 2: Machine learning approaches.

(A) Supervised learning is used when labeled data is available. Labels can be discrete classes, where a classification model learns a decision boundary to separate different categories (e.g., thermophilic vs. mesophilic proteins). Labels can also be continuous variables, where regression models learn relationships, such as predicting optimal growth temperature (OGT) based on amino acid properties. (B) Unsupervised learning is applied when labels are not available. Clustering methods, such as k-means or hierarchical clustering, group data points with similar features into distinct clusters, revealing underlying patterns in the data. (C) Generative modeling aims to learn the distribution of data, enabling the generation of new samples.

The models used in these respective categories could be classical machine learning approaches, such as linear regression, logistic regression, support vector machines, random forests, or shallow neural networks. However, with the rapid increase in computational power and the growing availability of biological data over the past decade, deep neural networks with multiple hidden layers have gained significant popularity. In this thesis, we frequently utilize two such architectures: Convolutional Neural Networks (CNNs) and Bidirectional Transformers. Given their importance, a brief explanation is warranted.

In *Papers I, II, V*, we primarily use convolutional models, although transformer modules are integrated into hybrid architectures in *Papers I,V*. For these convolutional models, we begin by encoding protein sequences using one-hot encoding, ensuring that each amino acid is represented as an orthogonal vector. This representation allows the sequence data to be processed by the network while preserving amino acid distinctiveness. CNNs operate under the assumption that meaningful patterns in the sequence exhibit local dependencies. Consequently, convolutions apply local transformation operations, meaning that each layer captures

information from a limited context window around each residue. This transformation is performed using a kernel (or filter) operation with a nonlinear activation σ, defined as:

$$y_i = \sigma\left(\sum_{j=0}^{k-1} w_j x_{i+j}\right) \quad [1]$$

for each position in the sequence [29] (Figure 3, panel A), where w_j are the learnable weights and k is the size of the kernel. This sliding-window approach enables CNNs to capture short-range sequence motifs effectively.

In contrast, *Papers III,IV* employ transformer-based models, which process sequences using a fundamentally different approach. Instead of a fixed one-hot encoding, these models use a tokenizer that converts residues into learned vector embeddings. Additionally, positional encodings are added to each residue embedding to retain order information. Unlike CNNs, which impose an implicit locality bias, transformers make no such assumption about local dependencies. Instead, they learn context dynamically through self-attention:

$$Attention\left(Q, K, V\right) = soft\max\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad [2]$$

where Q, K, and V are learned projections of the input embeddings and d_k is the dimension of the Q and K vectors. This mechanism allows transformers to model long-range dependencies across the sequence [30] (Figure 3, panel B). This is particularly relevant for proteins, as their 3D conformational structure can cause residues that are far apart in sequence to be functionally or structurally interdependent.

While deep learning architectures are often favored for their ability to learn complex patterns from data, it is important to recognize that increased flexibility also carries the risk of reduced generalizability to new data. This loss of generalizability arises when the model learns an overly complex decision boundary that captures noise in the data, or makes unreliable extrapolations to regions with limited data (Figure 4). This phenomenon, known as overfitting, and the use of transfer learning as a strategy to leverage data from a related domain to regularize the decision boundary, are further explored in *Paper II*.

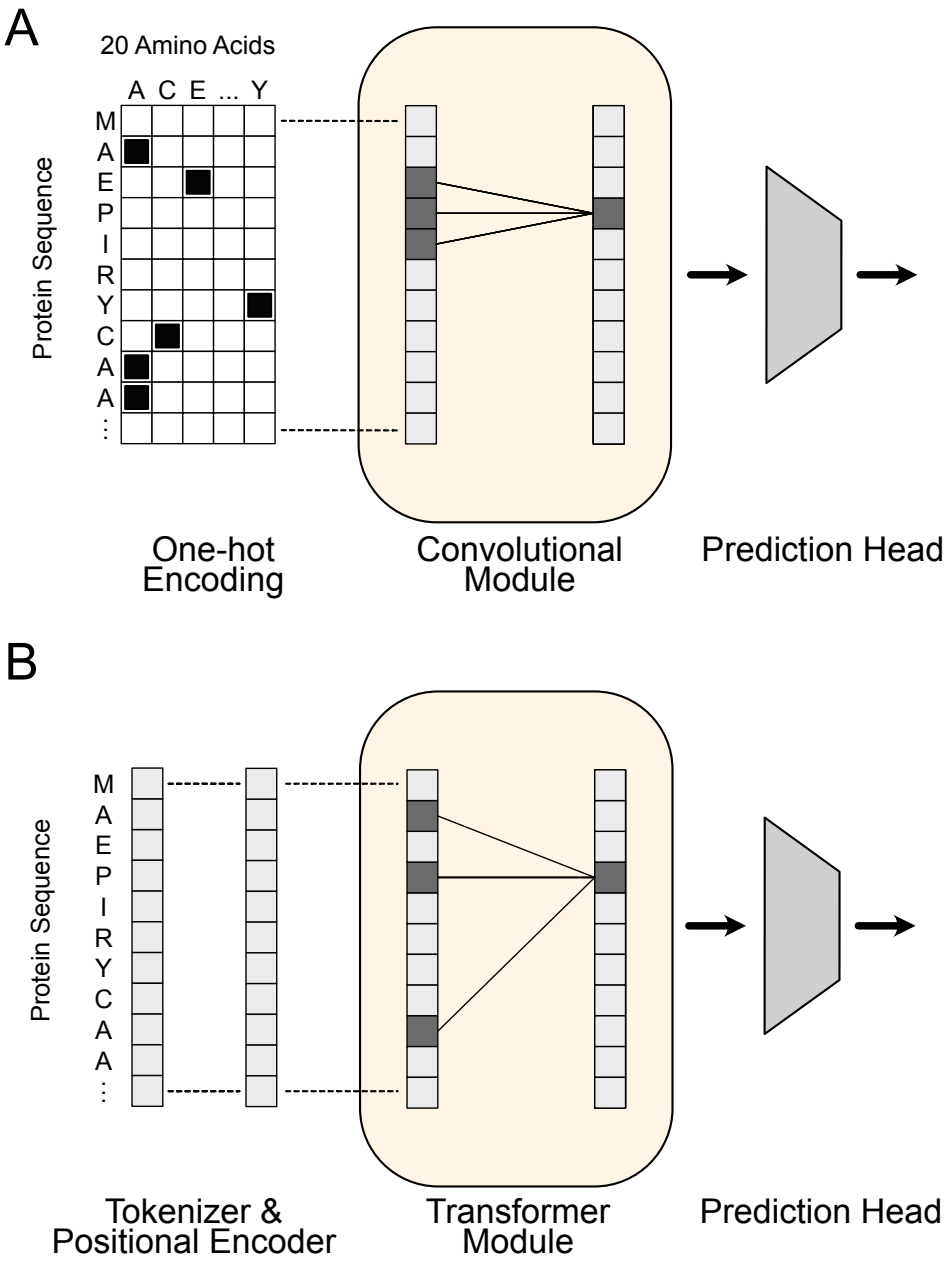

Figure 3: Convolutional and Transformer architectures used throughout the thesis.

(A) In the convolutional model, the sequence input is first converted into a one-hot encoded representation, which is then fed into convolutional layers. These layers extract features by applying local receptive fields, learning spatial patterns from fixed local contexts as information is passed forward through the network. (B) In the Transformer model, the input sequence is first processed by a tokenizer, which converts residues into vector representations (embeddings), followed by the addition of positional encoding to retain order information. Both of these transformations are learned. Unlike convolutional models, Transformers learn contextual relationships dynamically from the entire sequence using self-attention, allowing the model to determine relevant dependencies at different scales, rather than relying on fixed local patterns.

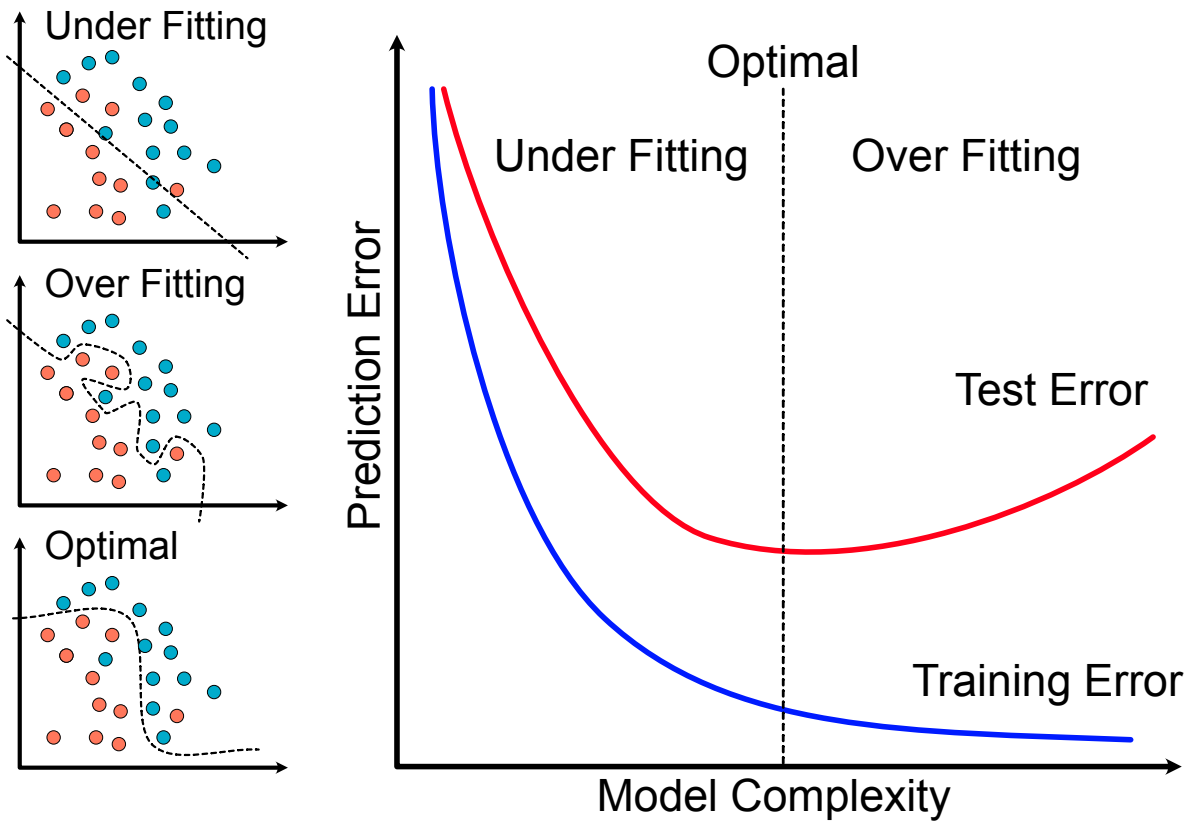

Figure 4: Bias-variance tradeoff.

The bias-variance tradeoff is the relationship between model complexity and generalization performance. As model complexity increases, both bias (training error) and generalization error initially decrease. However, beyond an optimal point, generalization error begins to rise due to increasing variance, leading to overfitting. Overfit models capture noise in the training data rather than generalizable patterns, resulting in poor performance on unseen data.

Protein Stability and Its Relevance to Protein Engineering

As indicated in the previous section, protein stability is often a critical property for their application. For example enzymes used in industry are often required to be thermally stable [31–33]. Likewise it is often essential that proteins used for therapeutics retain their native form and thus their activity for a long period[34]. Protein stability may also be critical for further engineering which may have destabilizing side effects. Increased protein stability also has additional benefits of being correlated with increased expression[35], and stabilization of a single chain can increase expression with as much as 100-fold[36]. Stability is also correlated with solubility[36] . Conversely, unstable proteins have a tendency to aggregate, possibly as an effect of partially exposing hydrophobic core residues [37,38]. Protein stability can mean several things, from thermodynamic stability, thermal stability, kinetic stability, and, in some cases, dynamic stability or conformational rigidity. These are used interchangeably in literature. Although they are related to one another, for the purposes of this thesis, we will define these terms and try to establish their relationships with one another.

Thermodynamic stability: One of the most fundamental concepts in protein stability is thermodynamic stability, which reflects the free energy of a system and the natural tendency for systems to minimize their available energy for performing work. The energy landscape of a system is defined by transitions between different states, with the Gibbs free energy (ΔG) describing the energy difference between them. For proteins, a simplified two-state model (Figure 5, panel A) considers the folded (F) and unfolded (U) states, connected by a reversible transition:

$$U \overset{K_{eq}}{\rightleftharpoons} F \quad [3]$$

At equilibrium, the populations of these states are governed by the equilibrium constant (K_{eq}), which is defined as the ratio of the rate constants for folding (k_F) and unfolding (k_U). Since equilibrium represents the point at which the rates of folding and unfolding are balanced, K_{eq} can also be expressed in terms of the concentrations of the folded ([F]) and unfolded ([U]) proteins:

$$K_{eq} = \frac{k_F}{k_U} = \frac{[F]}{[U]} \quad [4]$$

These rates are governed by thermodynamic principles and set by ΔG between the two states, which is the key determinant of protein stability. A system naturally tends toward the state with the lowest Gibbs free energy, meaning the difference in free energy between the unfolded and folded states, ΔG_U, determines whether folding is thermodynamically favorable:

$$\Delta G_U = G_U - G_F \quad [5]$$

ΔG_U is in turn, connected to the equilibrium constants as:

$$\Delta G_U = -RT \ln K_{eq} \quad [6]$$

Where R is the ideal gas constant and T is the temperature in kelvin. The gibbs free energy (ΔG) is also linked to, enthalpy (ΔH), entropy (ΔS), and temperature (T) and described by the Gibbs free energy equation[39]:

$$\Delta G = \Delta H - T\Delta S \quad [7]$$

Thermostability: In the equation for Gibbs free energy of folding ΔG_F , the ΔH_F will be the dominant term for stabilizing the folded state, due to favorable intramolecular interactions such as hydrogen bonding, van der Waals forces, and hydrophobic interactions. However, as temperature increases, the entropy ($T\Delta S_F$) term starts to dominate, reflecting the increased

number of microstates available to the system and the tendency towards disorder, which promotes unfolding. This interplay between enthalpy and entropy determines the thermal stability of a protein and the conditions under which it remains in its functional, folded state. A critical temperature in this analysis is the melting temperature (T_m), defined as the temperature at which $\Delta G(T) = 0$. At this point, the protein is equally likely to be found in the folded and unfolded states. T_m serves as a key metric in protein stability studies, providing insight into how proteins respond to temperature changes and guiding protein engineering efforts aimed at enhancing thermostability. While thermostability is related to the free energy, a high ΔG_U at a standard temperature of 25°C does not necessarily correspond to a high T_m[40,41].

Kinetic stability: While the transitions between the folded state and the unfolded state might be energetically favorable with a negative ΔG_U in some conditions, this does not necessarily translate to the transition being observed in a given time frame if the activation energy or free energy barrier $\Delta G_{U,\,Activation}$ is large[42]. However, this is not to say that the transition will not happen given enough time.

Dynamic or conformational stability: Proteins are inherently dynamic molecules, exhibiting significant fluctuations even in their native state. These fluctuations are crucial for protein function and enzymatic activity. Temperature influences these dynamics, with increased temperature leading to a higher rate of fluctuations. Although not a universal rule, more rigid proteins have been associated with greater thermostability[43]. Increasing rigidity in proteins has also been a successful strategy in engineering thermally stable proteins[44–46]. This increased rigidity at lower temperatures is also reflected in the optimal enzymatic temperature (T_{opt}), where many thermally stable proteins only become active at higher temperatures, allowing for sufficient conformational fluctuations[47].

While it is often useful to simplify these complex phenomena into a two-state reversible system, it is important to recognize that protein folding is far more complex. Transitions between protein states are not always reversible. As proteins unfold, they expose hydrophobic core residues, which can lead to aggregation with other unfolded or misfolded proteins. Aggregated states often reside in a lower free energy minimum and are typically separated by a substantial free energy barrier, making the transition back to the native state practically irreversible (Figure 5, panel B). Proteins do not exist in just two distinct states, proteins can exist in multiple conformations, each with higher free energy compared to the native state. This concept is often described using the folding funnel model, where a protein samples various partially folded local minima before reaching its native conformation[48] (Figure 5, panel C).

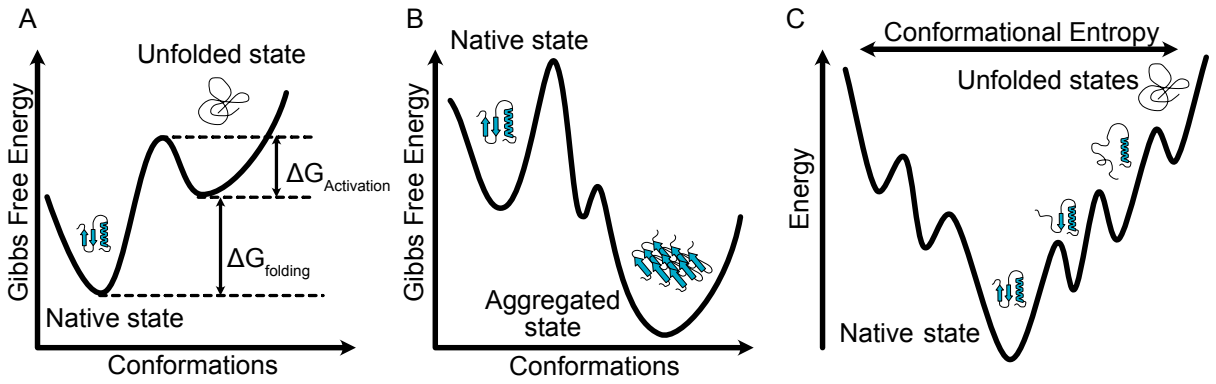

Figure 5: Protein folding occurs within a continuous energy landscape.

(A*) The transition between folded and unfolded states, with $\Delta G_{Folding}$ characterizing the free energy between the two states and the $\Delta G_{Activation}$ characterizing the energy needed to transition from the local minima of the unfolded state to the native state. (B) The transition between a natively folded state and a misfolded aggregate state is often characterized by a significant free energy barrier, making spontaneous reversal to the native state practically impossible. (C) The energy landscape of protein folding consists of multiple local minima corresponding to partially folded or unfolded conformations, where the free energy of the system goes down as the protein folds into its native state, and the number of available conformations increases with the free energy. The curve in the diagram is a representation of the energy landscape describing protein folding.*

Diversification with Deep Generative Modeling

ProteinGAN: Leveraging Generative Models for Enzyme Variant Library Design (Paper I)

The past few decades have witnessed an exponential increase in sequence data, largely driven by the advent of next-generation sequencing (NGS)[49] coupled with the automation of genome annotation[50]. This vast accumulation of protein sequences provides a resource that has yet to be fully utilized for guiding protein engineering. As previously mentioned, traditional approaches, such as directed evolution, rely on random mutagenesis and gene shuffling, largely ignoring this wealth of information. While techniques like CASTing make partial use of sequence data by constraining the mutational search space, they do not fully exploit the statistical patterns and functional relationships embedded within large protein sequence datasets. There are bioinformatics approaches that leverage sequence data, such as ancestral sequence reconstruction (ASR), which uses homologous sequences to construct phylogenetic trees. By interpolating sequence space between the extant sequences (leaf nodes), ASR allows for the inference of ancestral sequences. These methods have successfully engineered variants with desirable properties, including increased stability, enzymatic activity, and substrate promiscuity[51–53]. However, a key limitation of ASR is that it primarily generates the most statistically likely ancestors, thereby restricting the exploration of novel sequence space. As a result, while ASR refines existing functional diversity, it does not significantly expand it. In contrast, recent machine learning breakthroughs have demonstrated the power of large-scale data-driven models across diverse domains, including images, speech, and text, enabling the development of both discriminative (classification, regression)[54,55] and generative models capable of creating new data with both high fidelity and diversity[30,56–59]. Generative Adversarial Networks (GANs), in particular, revolutionized the ability to generate realistic images and music by learning complex distributions from datasets when they were introduced.

Inspired by these advances, we sought to leverage a Generative Adversarial Network (GAN) in *Paper I* to generate protein variant libraries that adhere to the underlying distribution of natural proteins while still introducing diversity. The GAN framework consists of two neural networks: a generator (G) and a discriminator (D), which are trained adversarially.

The generator takes random noise vectors sampled from an isotropic normal distribution as input and maps them to discrete protein sequences. The discriminator, on the other hand, is presented with both natural (wild-type) sequences and sequences generated by the generator. Its task is to distinguish between real and generated sequences. With adversarial training, the discriminator learns to improve its ability to classify natural (x) and generated (z) sequences through minimization of the likelihood loss function defined as:

$$\mathcal{L}_D = -\mathbb{E}_{x\sim p_{Data}}\left[\log D\left(x\right)\right] - \mathbb{E}_{z\sim p_z}\left[\log\left(1 - D\left(G\left(z\right)\right)\right)\right] \quad [8]$$

while the generator learns to produce sequences that increasingly resemble real proteins by optimizing its mapping from noise vectors to sequences. Here, expectation is taken over the empirical distribution x as well as the generated distribution z. This is achieved via gradient-based feedback from the discriminator, which guides the generator to refine its output through minimization of:

$$\mathcal{L}_G = -\mathbb{E}_{z\sim p_z}\left[\log D\left(G\left(z\right)\right)\right] \quad [9]$$

As training progresses, the generator becomes increasingly capable of capturing the complex sequence features characteristic of the protein family, ultimately learning to generate functional protein sequences. Likewise, the discriminator becomes more and more capable of discerning whether these complex features are present or not (Figure 6, panel A). Once the generator reaches this level of capability, it can be used to produce an arbitrary number of novel protein sequences by simply sampling from the isotropic normal distribution, providing a practically limitless pool of functional protein variants (Figure 6, panel B).

While the concept of training a GAN for protein generation sounds straightforward, proteins exhibit highly intricate and interdependent features that collectively define a functional protein family. Furthermore, the adversarial training setup of a generator and discriminator is inherently unstable, as either model can gain an advantage over the other, leading to poor learning dynamics. If the discriminator becomes too strong, it easily differentiates real sequences from generated ones, preventing the generator from improving. Conversely, if the generator overpowers the discriminator, it may exploit weaknesses in the model instead of learning meaningful sequence patterns. One common failure mode in GAN training is mode collapse, where the generator discovers a specific feature that consistently fools the discriminator and begins producing only a narrow set of sequences, losing diversity in the process. This defeats the purpose of having a generator capable of exploring the full sequence space. Numerous methods have been developed to mitigate mode collapse[60,61]. For ProteinGAN, we employed spectral normalization[62] as a regularization technique for the discriminator. This method constrains the discriminator's learning capacity, preventing it from focusing too heavily on any single feature. Doing so encourages the generator to explore a more diverse sequence space, leading to richer and more varied outputs. In addition to learning the correct sequence features, proteins must also maintain the correct sequential relationships between these features to remain functional. To address this, ProteinGAN incorporates self-attention[63] in the discriminator. Self-attention identifies pairwise relationships between residues in a sequence, allowing the model to evaluate whether key features appear in the correct positions relative to one another. This ensures that generated sequences not only contain the essential motifs of the protein family but also preserve their sequential relationship.

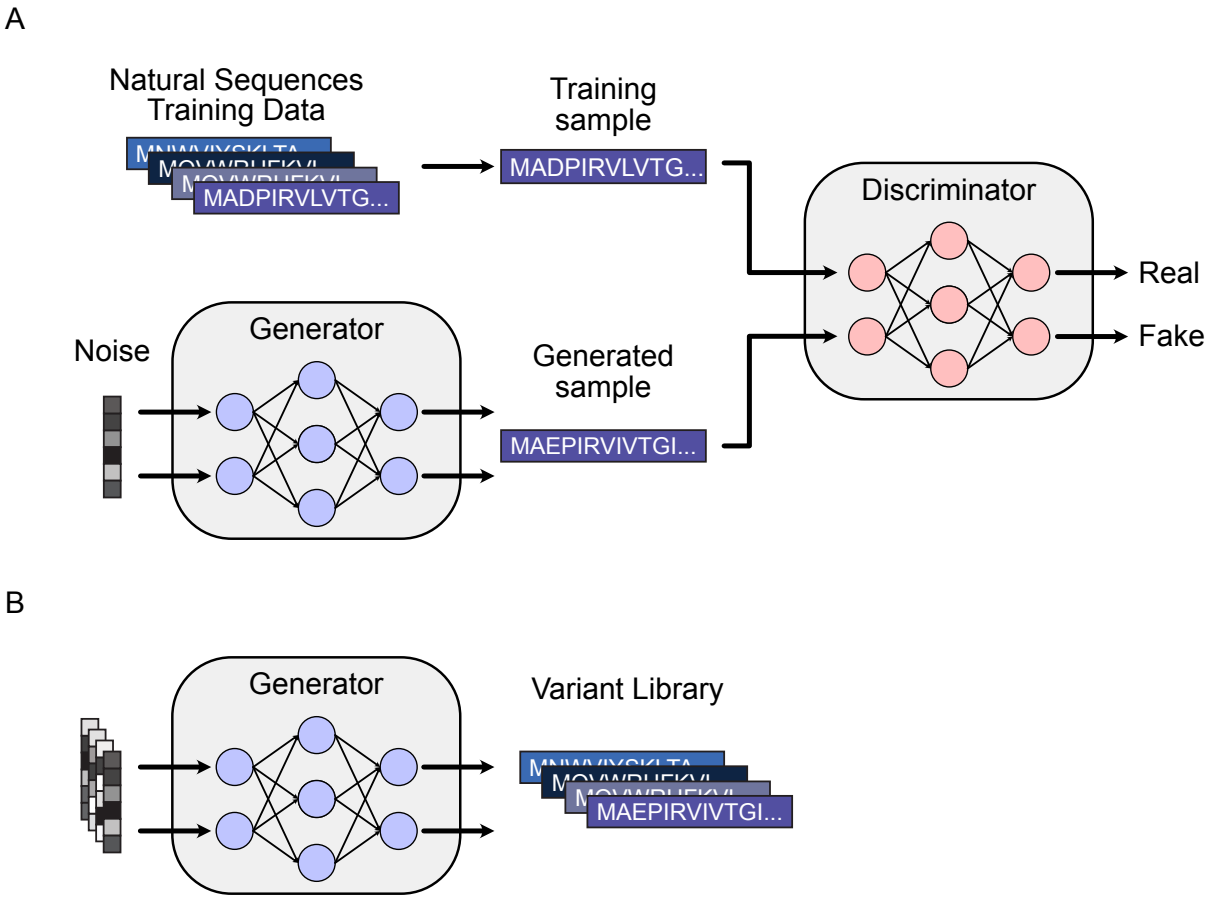

Figure 6: Training and application of a generative adversarial Network (GAN) for protein sequence design.

(A) The generative model is trained using adversarial training, where the generator network produces protein sequences, and the discriminator network evaluates whether the sequences are real (wild-type) or generated. Through iterative training, the generator learns to create sequences that resemble natural proteins, capturing the underlying distribution of wild-type sequences. (B) Once trained, the generator can be used to produce novel protein variants that expand upon natural diversity. These generated sequences can then be selected based on function, stability, or other desirable properties, similar to directed evolution.

To demonstrate ProteinGAN's ability to generate functional proteins, we selected the malate dehydrogenase (MDH) protein family as a test case. A dataset of 16,898 bacterial MDH sequences was obtained from the UniProt database. Of these, 16,706 sequences were used for training, while the remaining 192 were set aside as a validation set. Once ProteinGAN was trained, we evaluated whether the generated sequences captured key characteristics of the MDH family, including amino acid variability (Figure 7, panel A) and functionally important sites, such as substrate and cofactor binding sites and the H^+ acceptor residue (Figure 7, panel B). To gain deeper insights into how ProteinGAN learned these features, we analyzed the discriminator's attention maps for the training sequences. The extracted attention maps revealed that attention was primarily local, meaning that amino acids in close proximity within the sequence provided the most informative signals for discrimination (Figure 7, panel C). To further investigate the spatial relationships learned by the model, we correlated pairwise

attention scores with the Euclidean distances between residues in the corresponding 3D structures (Figure 7, panel D). ProteinGAN exhibited significant variance in these correlations across different sequences, suggesting that rather than relying on a fixed set of conserved features for the entire protein family, the discriminator learned to identify sequence-specific features that are important for an individual protein. This adaptability highlights the model's ability to capture functionally relevant sequence constraints while allowing for natural diversity.

Many conventional bioinformatics approaches, including generative models such as hidden Markov models and profile-based methods, have demonstrated an ability to capture evolutionary and structural information[64,65]. However, the functional validity of sequences generated by these models remains largely untested due to a lack of experimental validation. To assess the ability of ProteinGAN to generate functional proteins and demonstrate the potential of deep generative models for designing diverse variant libraries, we selected 60 generated sequences spanning 45–98% identity to their closest natural homologs in the training set for experimental validation (Figure 8, panel A). Of these, 55 sequences were successfully cloned and expressed, and 19 variants were purified. Enzyme activity assays revealed that 13 of the 19 purified variants were catalytically active, including one variant with as little as 66% sequence identity to its closest natural counterpart (Figure 8, panel B). Further biochemical characterization confirmed that the active variants specifically catalyzed the conversion of oxaloacetate to malate, with reaction yields comparable to those of natural MDH enzymes (Figure 8, panel C).

With ProteinGAN, for the first time, we demonstrated the potential of deep generative models to learn directly from natural protein sequence data and generate diverse functional variants. Notably, some generated sequences retained enzymatic activity despite having as little as 66% sequence identity to their closest natural counterpart in the training set. These findings highlight the potential of deep generative models as an *in silico* approach for designing novel protein variants and showcasing them as a potential tool in the diversification step of the protein engineering cycle.

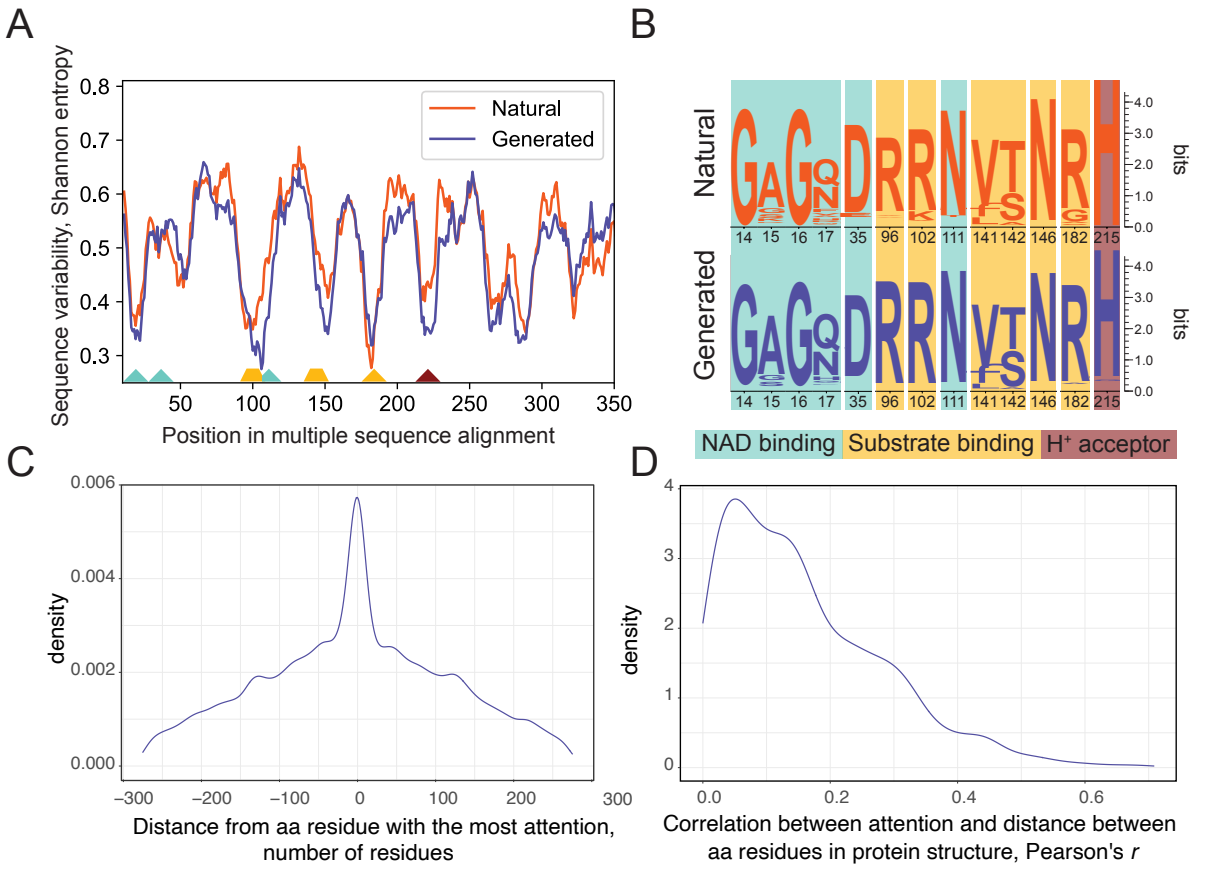

Figure 7: ProteinGAN generates protein sequences that closely follow the distribution of natural sequences while preserving functionally important sites.

(A) Amino acid (AA) variability of natural malate dehydrogenase (MDH) sequences and generated sequences. Sequence variability is represented as Shannon entropy values computed from multiple sequence alignments (MSA) of both generated and natural sequences. Lower entropy values indicate highly conserved, functionally relevant positions, while higher entropy reflects greater sequence diversity. (B) Sequence logos illustrate key conserved positions in the MSA, comparing natural and generated sequences. Functional sites associated with NAD binding, substrate binding, and proton acceptance are highlighted. (C) Distribution of positions where maximum attention is focused in real MDH sequences. Negative values correspond to the residues preceding the current position, while positive values correspond to the succeeding residues. (D) Correlation between attention scores and amino acid pairwise Euclidean distances in the corresponding protein structures. Figure reproduced from Paper I (Nature Machine Intelligence 2021)

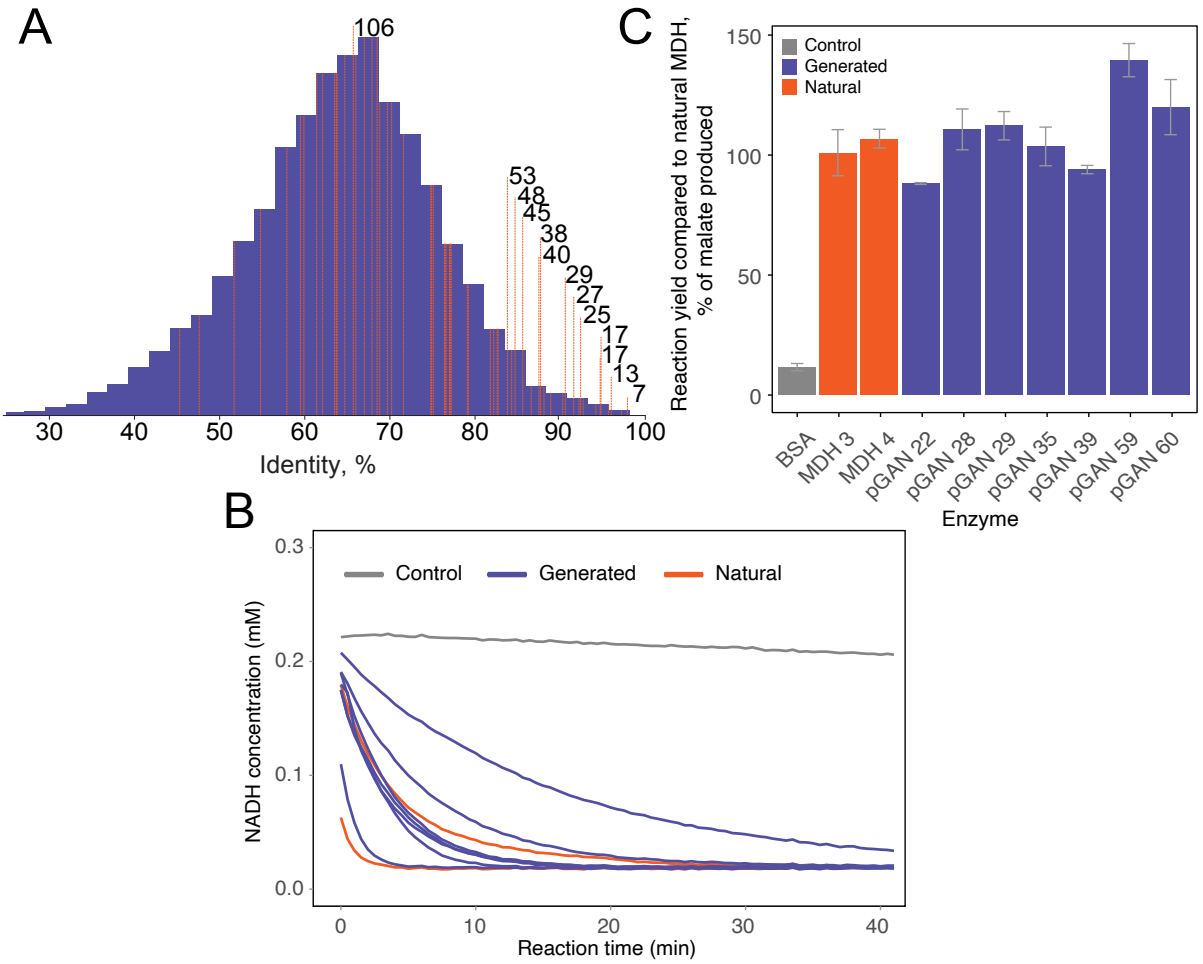

Figure 8: ProteinGAN generates diverse sequences with comparable activity to wild-type controls.

(A) Sequence diversity of ProteinGAN-generated enzymes, measured as the global sequence identity to the closest wild-type sequence in the training set. The histogram represents the full distribution, while dashed lines indicate experimentally validated sequences. Numbers above the dashed lines denote the total mutations (insertions, deletions, and substitutions) in active variants relative to their closest wild-type counterparts. (B) Enzymatic activity of generated and natural malate dehydrogenase (MDH) variants, measured by fluorescence-based monitoring of NADH consumption over time. Generated variants (blue) exhibit comparable kinetic profiles to natural MDH enzymes (orange). (C) Quantification of oxaloacetate-to-malate conversion yields, assessed via spectrophotometry. Generated enzymes (blue) demonstrate yield on par with natural MDH variants (orange), confirming that their catalytic function is equivalent to that of the wild-type controls. Figure reproduced from Paper I (Nature Machine Intelligence 2021)

Selection with Deep Discriminative Modeling

Machine Learning Prediction of Thermal Stability (Paper II)

Under the same environmental conditions, a given amino acid sequence reliably folds into the same three-dimensional structure, demonstrating that sequence alone encodes much, if not all, of the necessary structural information. Recent advances in computational modeling, such as AlphaFold2, RoseTTAFold, and ESMFold, have formally confirmed this by accurately predicting protein structures solely from sequence and alignment data [28,66,67]. Given the strong relationship between protein structure and function, including physicochemical properties and enzymatic activity, it is reasonable to expect that sequence data could also encode information related to these properties. In protein engineering, as mentioned earlier, one of the most critical physicochemical properties is stability, which relates to both the melting temperature (T_m) and the optimal enzymatic activity temperature (T_{opt}). In *Paper II*, we explore how deep learning models can predict these properties, enabling their use in guiding the selection of promising protein variants for engineering applications. Traditionally, predictive models for proteins have relied on handcrafted sequence features, such as *iFeatures* [68], which are then used as input for classical machine learning algorithms like support vector machines or random forests. However, deep neural networks offer an alternative approach, learning directly from raw sequence data to parameterize regression models. Despite their potential, early applications of these models in protein property prediction have been relatively limited in scope [69–71]. Another strategy to improve predictive performance is incorporating additional biological metadata, such as the host organism's optimal growth temperature (OGT) [72]. While this can enhance accuracy, it also increases reliance on experimentally determined values, which may not always be available for sequences of interest. A major challenge in developing machine learning models for predicting thermal stability, whether T_{opt} or T_m, is the limited availability of labeled training data. For T_{opt}, the BRENDA[73] database contained only 1,902 annotated enzymes as of 2019 (Figure 9, left). For T_m, the largest dataset available is the Meltome Atlas[74], which includes 41,725 enzyme entries (Figure 9, right). While these datasets represent extensive experimental efforts, they remain relatively small by machine learning standards, where models often require millions of data points to learn complex feature relationships effectively. Beyond dataset size, the composition of the data, both in terms of the independent features the model learns from, and the distribution of the dependent variable, is also critical [75]. For instance, the Meltome Atlas dataset is biased toward enzymes with lower thermal stability (<60°C), reflecting the preferential selection of mesophilic organisms in the study. However, as demonstrated by Leuenberger *et al.* [76], an alternative dataset incorporating a higher proportion of thermophilic organisms results in a more balanced distribution of T_m values (Figure 9, middle). This highlights how dataset design can influence the generalizability of predictive models.

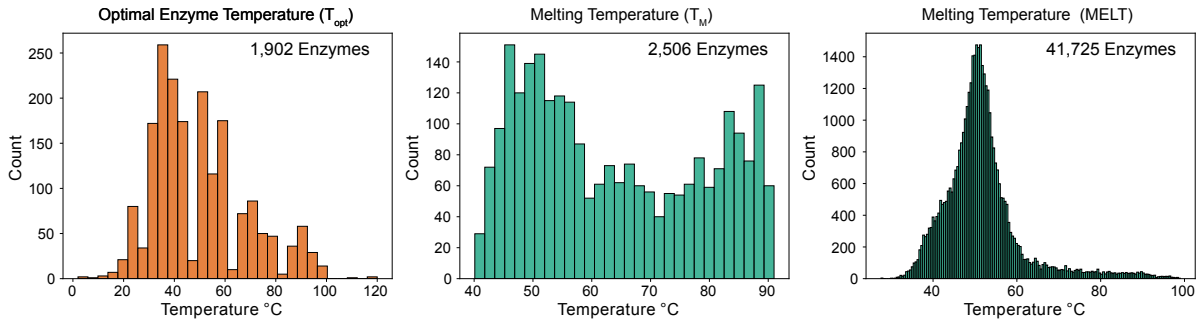

Figure 9: Temperature distributions of three enzyme datasets.

The histograms show the temperature distributions of three datasets used in this study. Left: A small dataset of 1,902 enzymes labeled with their optimal enzymatic activity temperature. Middle: A dataset of 2,506 enzymes with melting temperature annotations from Escherichia coli, Saccharomyces cerevisiae, and Thermus thermophilus. Right: A larger dataset of 41,725 enzymes, collected from the Meltome Atlas, provides melting temperature measurements across 13 species. Figure recreated as done in Paper II (Protein Science 2022).

To address the challenge of limited labeled data in one domain, it is common to leverage data from a related domain through a technique known as transfer learning. The underlying assumption is that both datasets share low-level features that are relevant across domains. By first training a model on a larger, related dataset, it can learn these common features before fine-tuning on the smaller target dataset. Selecting related datasets based on shared dependent variables is a standard practice. For example, a model trained to classify cat breeds would likely have learned features that are also useful for distinguishing dog breeds. In some cases, transfer learning has proven effective even when the source and target domains differ significantly, as large datasets can still help models capture fundamental patterns necessary for effective learning [77,78]. For transfer learning to be most effective, the source dataset is typically much larger than the target dataset of interest. This technique has been extraordinarily successful across various fields, including medical imaging and natural language processing [79,80]. For thermal stability prediction, we utilize a related dataset in the form of optimal growth temperature (OGT) data, compiled by Enquist, M. K. M.[81], containing OGT values for 8,184 organisms. Each protein in these organisms' proteomes is labeled with its respective OGT (Figure 10, panel A)[81]. This dataset consists of over 3 million nonredundant sequences with associated OGT values, providing a significantly larger pool of labeled data compared to datasets for T_m or T_{opt}. Notably, the OGT dataset exhibits a bias toward mesophilic sequences (OGT <45°C), similar to how the Meltome Atlas dataset is skewed toward lower thermal stability enzymes (Figure 10, panel B). To establish a strong feature extractor, we trained a deep neural network (DNN) consisting of a convolutional feature extraction module followed by a fully connected regression head (Figure 10, panel D). When trained on the OGT dataset, this model explained 59% of the variance in OGT, achieving a root mean squared error (RMSE) of 5.5°C on a held-out test set of 150,776 sequences (Figure 10, panel C). To evaluate the effectiveness of transfer learning using the OGT dataset, we tested its performance on the three

smaller datasets: T_{opt}, T_m, and the Meltome Atlas (MELT). We compared two transfer learning strategies:

- **FrozenCNN**: The convolutional feature extractor was kept fixed, and only the regression head was fine-tuned on the new dataset.
- **TuneALL**: The entire model, including the feature extractor, was fine-tuned on the new dataset.

As controls, we evaluated four additional approaches:

- **iFeatures**: A classical machine learning approach using handcrafted sequence features with the best-performing shallow ML models.
- **UniRep**: A general protein transfer learning model[82] using the same classical ML regression methods as iFeatures.
- **FromScratch**: Training the same deep learning architecture from scratch without pretraining on the OGT dataset.
- **FrozenAll**: Using the OGT-trained model without any fine-tuning on the target datasets.

By comparing these strategies, we assess the extent to which transfer learning from OGT data improves thermal stability prediction and whether fine-tuning the entire model **TuneALL** outperforms freezing the feature extractor **FrozenCNN**.

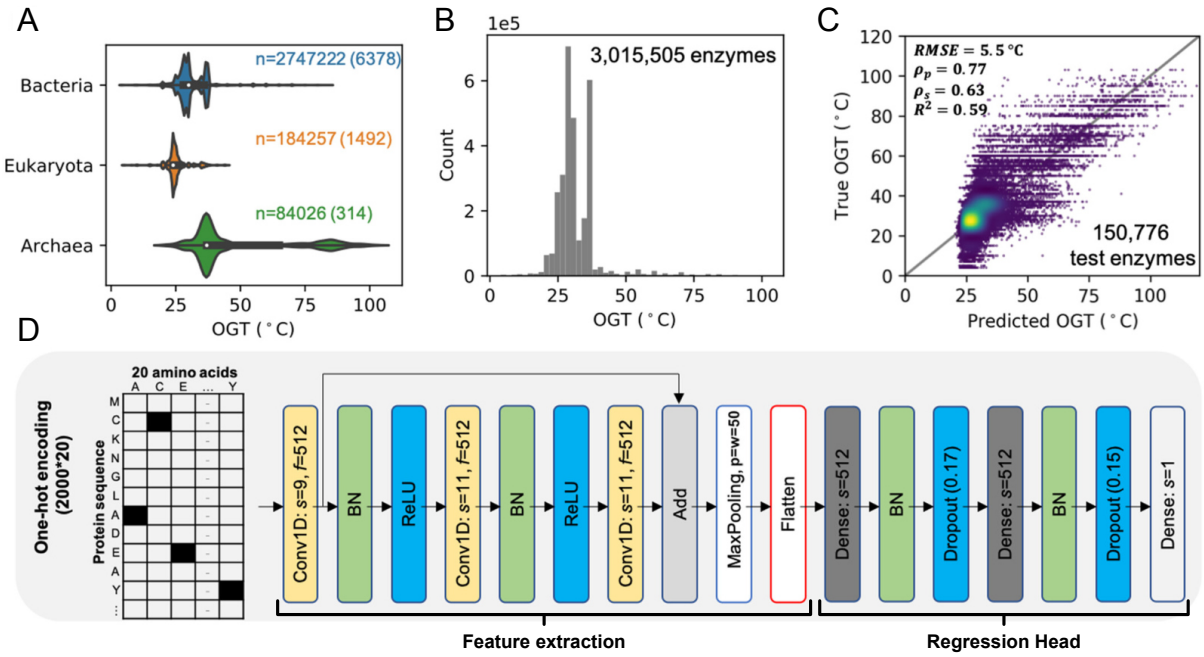

Figure 10: Learning sequence features from optimal growth temperature (OGT) data.

(A) OGT data from 8,184 species spanning the three domains of life: Bacteria, Eukaryota, and Archaea. The number of species (in parentheses) and the number of associated protein sequences (preceding number) are indicated for each domain. (B) Temperature distribution of 3,015,505 proteome sequences from these organisms, illustrating the overall bias in available OGT data. (C) The performance of the deep learning regression model was trained on OGT-annotated sequences. Model predictions are evaluated on 150,776 test sequences, which were randomly split from the dataset. The model achieves an RMSE of 5.5°C and an R^2 value of 59%. (D) Schematic of the regression model architecture. Input protein sequences are one-hot encoded, padded to a maximum length of 2,000, and processed through a two-part model architecture: A feature extraction module consisting of convolutional layers with residual connections, batch normalization (BN), and ReLU activations, and a regression head, composed of fully connected layers with dropout for regularization, followed by a final dense layer with a single neuron using a linear activation function. The feature extraction module is designed to be used separately for transfer learning in related tasks. Figure reproduced and annotated from Paper II (Protein Science 2022).

When evaluated on holdout test data, it is evident that the model pretrained on the OGT data set outperforms the classical iFeature and UniRep-based model, as well as the deep learning models trained from scratch. The effect of transfer learning was most noticeable for the two smallest T_{opt} and T_m data sets (Figure 11, panels A, B, D, E), where the explained variance of the best models were 57% and 73%, respectively. The increase in performance was still noticeable for the significantly larger MELT data set (Figure 11, panels C, F) with an explained variance of 58% for the best model. For all of the data sets, the OGT pretrained models had a significantly increased performance (Welch's t-test p-value < 0.05).

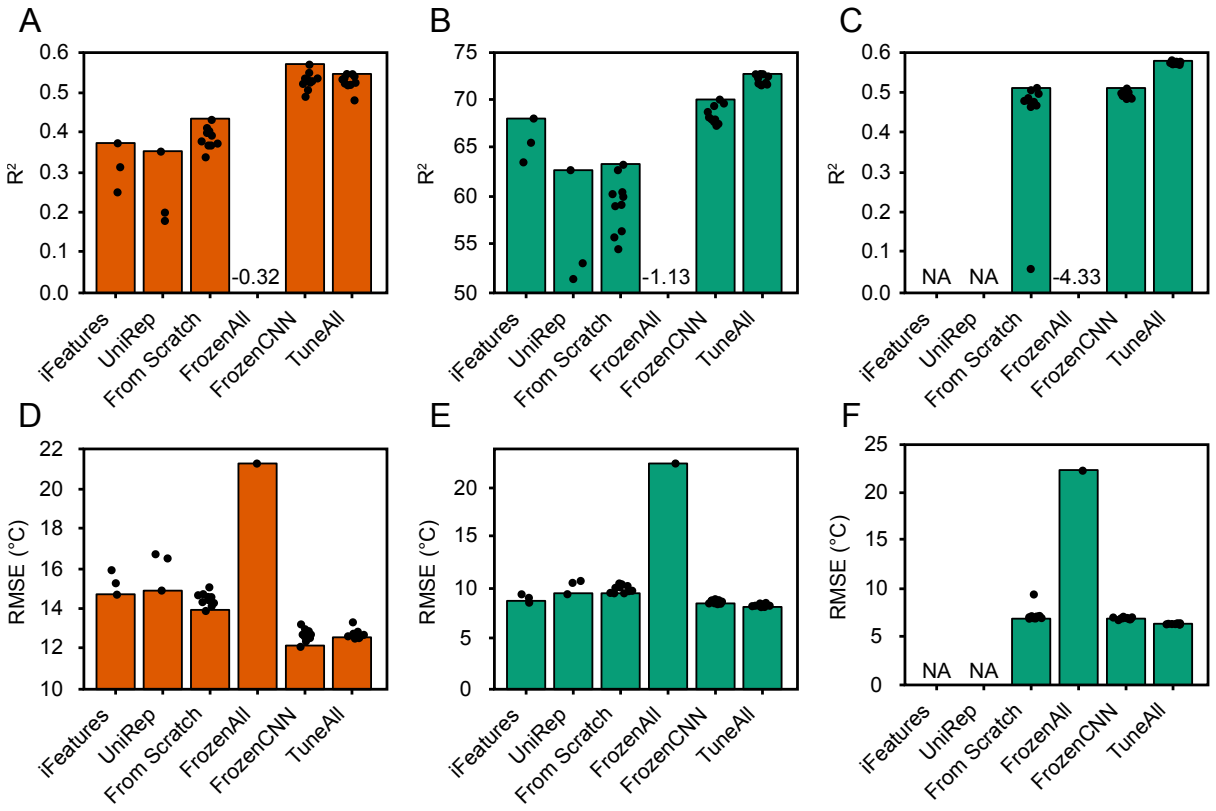

Figure 11: Comparison of traditional feature-based machine learning and deep learning with transfer learning.

Panels (A–C) show the coefficient of determination (R^2) for various models, while panels (D–F) present the root mean square error (RMSE) in degrees. (A, D) display results on 190 test sequences from the T_{opt} dataset. (B, E) show performance on 251 test sequences from the melting temperature (T_m) dataset[76]. (C, F) illustrate performance on 4173 test sequences from the MELT dataset. The iFeature and UniRep models are evaluated using three different shallow ML methods. The training From Scratch, FrozenCNN, and TuneALL have all been repeated 10 times. The FrozenAll was only conducted once. Due to computational constraints, the iFeature and UniRep models were omitted from the MELT data. The bars represent the best iteration for each model. Figure panels A-C reproduced from Paper II (Protein Science 2022).

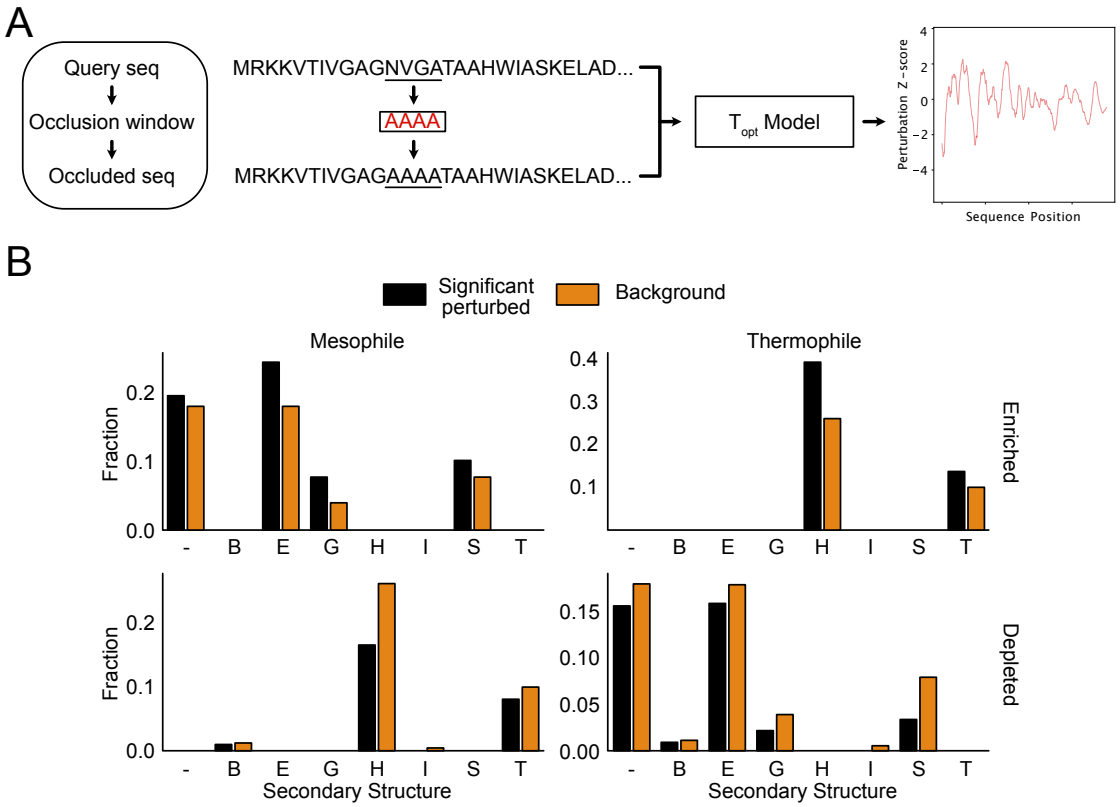

Figure 12: Structural elements influencing Topt prediction differ between mesophiles and thermophiles.
(A) Perturbation maps were generated by systematically occluding sequences using contiguous alanine substitution windows. Each window was sequentially applied along the entire sequence, and the resulting changes in predicted T_{opt} values were converted to z-scores. These z-scores were then mapped to sequence-related features, such as amino acid identity and secondary structure elements. (B) Significantly perturbed DSSP[83] motifs, highlighting structural elements that are critical for T_{opt} prediction. Regions with strong perturbation indicate positions where sequence information is highly influential in the model's predictions. Figure panel B reproduced from Paper II (Protein Science 2022).

Since the feature extractor module of our regression model demonstrated strong performance in learning relevant features for thermal stability, particularly for T_{opt}, the next logical step was to probe the model to understand which features it had learned. Neural networks are often regarded as black-box models, meaning their internal decision-making processes are not directly interpretable. However, several techniques have been developed to shed light on the logic learned by these models. One such method is saliency mapping, where gradients that maximize class or value prediction are propagated back through the model, highlighting which input features most influence the direction of prediction [84]. Another approach is Local Interpretable Model-agnostic Explanations (LIME), which involves making small perturbations to the original data, recording the changes in predictions, and training an interpretable model to approximate the local behavior of the original model [85]. Additionally, attention mechanisms, used in *Paper I*, are commonly used for interpretation [86]. However, our current model does not incorporate attention layers, and interpreting saliency maps for discrete data, such as one-hot encoded sequences, can be challenging. Thus, we choose to employ a

modified version of LIME for our interpretation study. Instead of general perturbations, we applied occlusion-based perturbations by systematically substituting segments of five amino acids of the sequence with alanine (A) residues. This occlusion was performed sequentially across the entire sequence, and the corresponding predictions were recorded. Instead of training a linear regression model to predict the effects of perturbations, we smoothed the perturbed values using a moving average filter (Figure 12, panel A). For further analysis, we focused only on significantly perturbed sites (|STD| > 2). To distinguish between different stability profiles, we divided the sequences into mesophilic (OGT 20–45°C) and thermophilic (OGT > 45°C) groups. The first aspect analyzed was amino acid composition, where we examined the impact of specific residues on the model's predictions. The results aligned well with previously established amino acid preferences for thermostability [87]. Specifically, residues such as K, E, and R (charged), I, L (hydrophobic), and F, Y (aromatic) were more relevant for predicting thermophilic sequences, reflecting their known role in enhancing thermal stability. Conversely, residues such as A, S, and W, which are more commonly associated with cold-adapted proteins [87], had a greater influence on the prediction of mesophilic sequences. We also analyzed the importance of secondary structural elements in prediction. The results revealed that mesophilic proteins exhibited a broader range of structurally important motifs, whereas thermophilic proteins showed a more specific dependence on turns and α-helices (Figure 12, panel B). The increased relevance of helical structures in thermophiles is consistent with their stabilizing nature and the enrichment of arginine (R) residues at helix termini, which are known to enhance thermostability[88]. While showcasing that the model manages to capture known biological features, much of the model's internal decision-making remains opaque. This occlusion study only examines how the removal of information influences predictions. The choice of window size and amino acid substitution introduces a degree of bias, potentially affecting the results. Additionally, the perturbations capture only local, sequential features, meaning that global feature relationships learned by the model remain unexplored. Addressing this limitation through global occlusion variations would be computationally prohibitive, making it a challenging avenue for further investigation. Thus, for further analysis, it might be worth looking into applying attention to explicitly gain the ability to probe the global features learned by these models.

In this work, we demonstrated the potential of transfer learning to enhance the prediction of protein thermostability, even when using a biased and low-precision dataset such as the OGT dataset. While the model's raw predictive performance may not yet be precise enough to fully replace experimental methods like thermal assays for candidate sequence selection, it shows promise as a filtering tool to reduce the number of sequences requiring experimental screening.

Deep Learning Finds a Relationship between Amino Acid Sequence and Abundance (Paper III)

In the previous section, we explored how deep neural networks can predict physicochemical properties such as melting temperature and maximal enzymatic temperature. These properties are intuitively linked to the amino acid sequence, given the well-established relationships between sequence, structure, and function. However, systemic properties like protein abundance are more complex, as they are primarily governed by regulatory elements at the DNA level and follow the principles of the central dogma. Despite this complexity, certain aspects of protein abundance may still be encoded within the amino acid sequence itself. During steady state, protein abundance can be modeled as a function of both production and degradation rates. While these processes emerge from intricate cellular mechanisms, evolutionary constraints may have imprinted signals within the protein sequence that encode variability in abundance. For instance, protein production is largely regulated at the transcriptional level, particularly during initiation [89–91], which is encoded in the transcript sequence [92,93]. However, the N-terminal region of the amino acid sequence has also been shown to influence this process [94]. The metabolic cost incurred by the use of certain amino acids may also influence the transcription rate of proteins[95–97]. Similarly, protein degradation is governed by the complex interplay of proteolysis [98], yet the C-terminal region plays a crucial role in determining degradation rates, contributing to variability in abundance [99,100]. Since most research has focused on genomic or transcriptomic data and, given evidence that protein abundance is partially encoded in the amino acid sequence, *Paper III* aimed to develop a machine learning model to predict protein abundance directly from sequence. For this task, we utilized two datasets: (i) a collection of 136 proteomes with abundance measurements from PaxDB [101] and (ii) an experimental dataset containing 21 independent abundance measurements for the *Saccharomyces cerevisiae* proteome from Ho et al. [102]. Due to the challenges of interpreting global features in the convolution-based model from *Paper II* and the increased complexity of the current task, we opted for an architecture incorporating attention mechanisms. We believed this choice would not only enhance the model's performance but, more importantly, improve its interpretability. To begin, we sought to assess the relative information encoded in amino acid sequences across the Tree of Life by leveraging both datasets. Each proteome in PaxDB, along with the corresponding median abundance value, was treated as an independent dataset, as was the dataset from Ho et al. Given the relatively small size of our datasets and the demonstrated benefits of transfer learning in *Paper II*, we first utilized a pretrained Bidirectional Encoder Representation from Transformer (BERT) model, specifically ESM-1b [103], to extract meaningful sequence embeddings. This approach has been highly effective in various downstream biological tasks [103,104]. Using these embeddings as feature representations, we trained small neural networks as regression heads to predict the median abundance values for each dataset. Despite the simplicity of this approach, relying only on a pretrained model for feature extraction and a lightweight neural network for regression, more than 50% of the datasets achieved an explained variance

exceeding 40% on an independent test set. Notably, this trend was consistent across different domains of life (Figure 13, panel A).

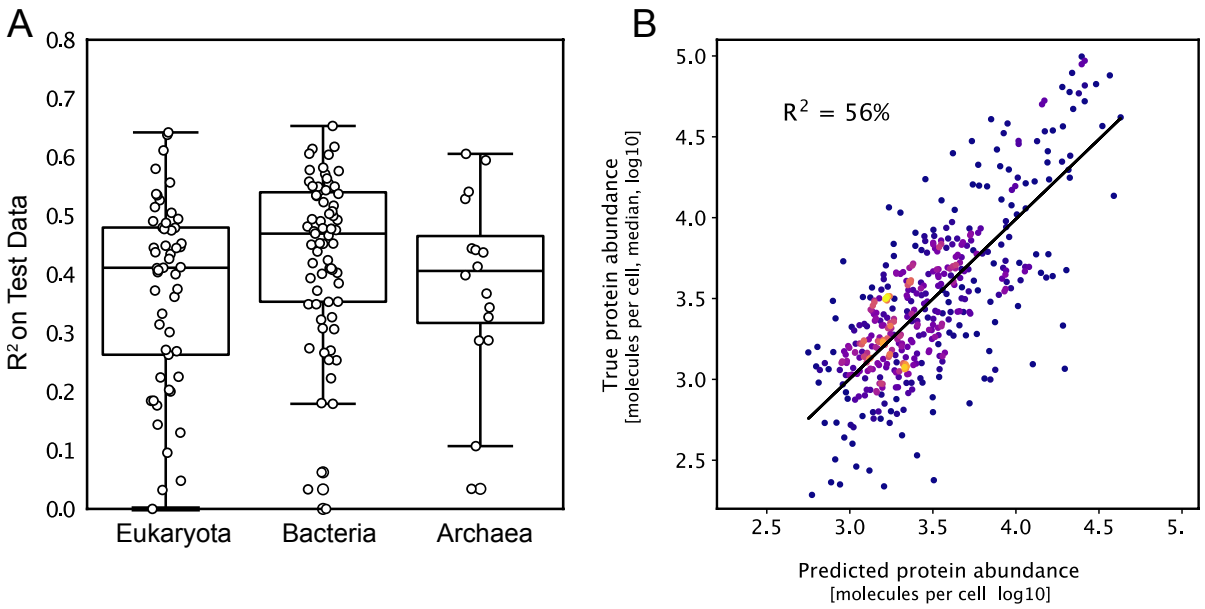

Figure 13: Amino acid sequence encodes information about protein abundance across the domains of life.

(A) Performance of neural networks trained on ESM-1b sequence embeddings to predict protein abundance across 137 diverse proteomes from PaxDB and Ho et al.. Each point represents a protein, and the model's predictive accuracy is evaluated across different taxonomic groups. (B) Performance of a BERT-based deep learning model with a regression head trained from scratch on Saccharomyces cerevisiae sequences and abundance values from Ho et al. The scatter plot shows predicted versus observed abundance values, with color intensity reflecting data density. Figure reproduced from Paper III (Protein Science 2025).

While ESM-1b proved highly effective for feature extraction in the regression task, interpreting its internal attention maps would not provide meaningful insights into protein abundance predictions. This is because the model itself was not fine-tuned during training. Only the regression heads learned to map embeddings to abundance values. Even if we had fine-tuned ESM-1b alongside the regression head, interpreting its attention patterns would remain challenging, as we would not be able to discern whether relationships were learned during pretraining or during the fine-tuning process. This is in contrast to the approach in *Paper II*, where the model was pretrained on a related and correlated task, making interpretation more feasible. In the case of ESM-1b, however, pretraining was conducted on a task entirely unrelated to protein abundance prediction, limiting the usefulness of its internal representations for interpretation. Instead, we trained a smaller version of the bidirectional transformer architecture model from scratch on the Ho et al. data set, which showed the best performance in the previous experiment. In order to train this model effectively on the small amount of data, we made use of data augmentation where each protein were sampled for each experimental value and for every sampled protein a shuffled version of the sequence were introduced with

its abundance value set zero (1e-5), effectively increasing the the number of data points to 199,206. The shuffling was made in order to discourage the model from only learning the amino acid composition and instead learn more semantic features of the sequences. Although not reaching quite the same performance as when the ESM-1b model was used to extract embeddings, the BERT trained from scratch still explained 56% of the variance of protein abundance (Figure 13, panel B). With this BERT model, we could then extract attention profiles for all the sequences in our dataset, including the shuffled sequences, as control. These attention profiles that relate to the relative importance of each residue could then be correlated to various physicochemical amino acid indexes and metabolic cost associations (Figure 14, panel A). The attention profiles were found to correlate with amino acid cost associations when mapping sequence to protein abundance, with attention being drawn toward either energy-intensive amino acids or those with lower synthesis costs. These correlations do not directly link amino acid cost to predicted abundance but rather highlight the latent features the model has learned. As a control, shuffling protein sequences resulted in negligible correlations, reinforcing that attention weights capture meaningful positional information (Figure 14, panel B). Similarly, attention profiles showed strong correlations with physicochemical amino acid indices and structural elements, such as secondary structure and protein domains (Figure 14, panel C). Notably, helices appeared particularly important for model predictions, suggesting that helical content may be a key feature the model has learned. Across the various covered domains, the associated GO terms were diverse, encompassing translation, protein folding, post-translational modification, carbohydrate and ion transport, stress response, organelle fission, cell cycle, cell division, and sporulation. This suggests that no single domain is uniquely informative; rather, the structured nature of these domains contributes to the model's ability to predict protein abundance.

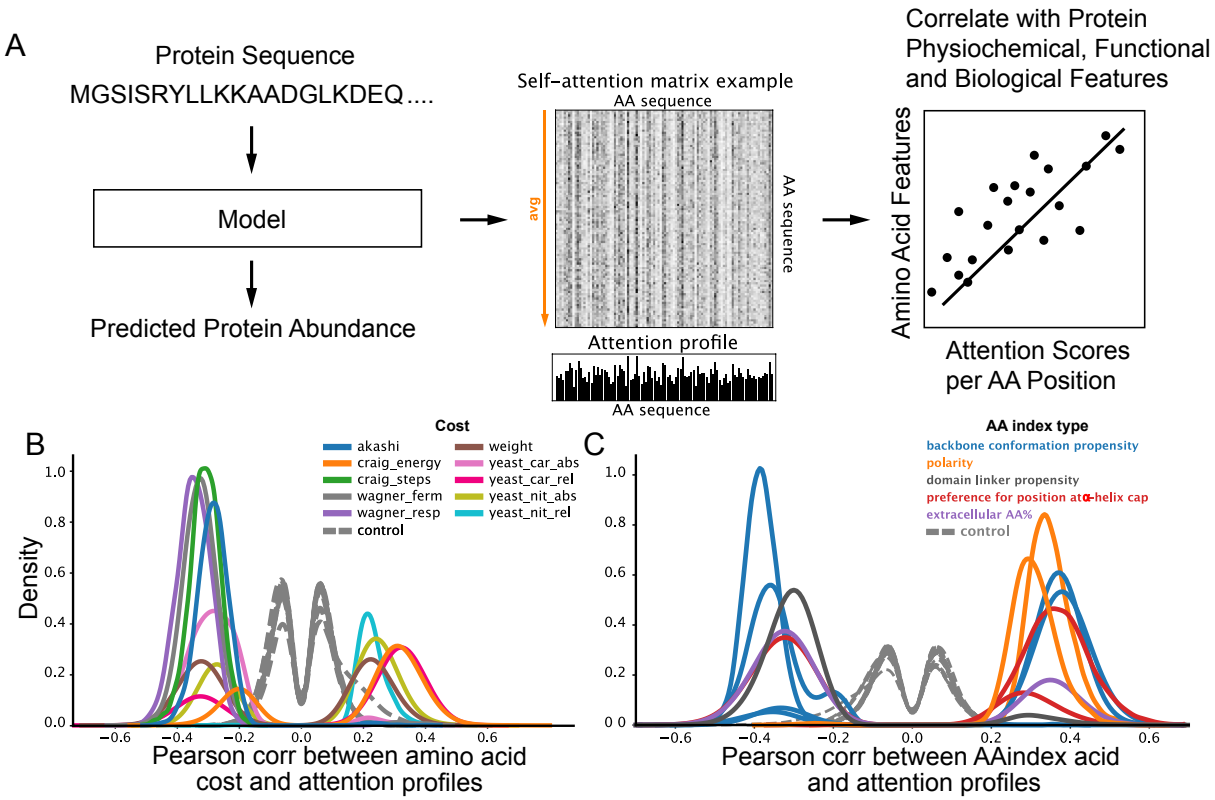

Figure 14: Model attention profiles correlate with amino acid properties and metabolic costs.

(A) Attention profiles extracted from a trained BERT model reflect how the model distributes focus across residues during sequence-based predictions. These profiles can be used to investigate correlations between amino acid indices and amino acid costs. (B) The correlation between amino acid costs and model attention scores suggests that residues with higher biosynthetic costs may be differentially attended to by the model. (C) Correlation between amino acid indices and attention profiles, highlighting relationships between model focus and biochemical properties such as backbone conformation propensity, polarity, and domain linker propensity. Figure reproduced and altered from Paper III (Protein Science 2025).

Given that the model exhibited moderate correlations with biological and physicochemical properties linked to protein abundance, a natural next step is to investigate whether these relationships extend beyond correlation and capture causal determinants of predicted abundance. Specifically, we aim to understand which sequence features contribute to high or low predicted abundance and whether these features can be systematically modified to alter abundance predictions. More intriguingly, we test whether high-abundance characteristics can be transplanted into low-abundance sequences to enhance their predicted abundance. By probing the learned sequence representations in this way, we try to peer directly into this black box model. To investigate how individual amino acids contribute to protein abundance predictions, we analyzed the embedded space learned by the Transformer encoder. Specifically, we trained a parametric UMAP[105] projection to reduce the high-dimensional representation to a one-dimensional scale, creating an "embedded ordering". This ordering provides a ranking of amino acids within each sequence based on their predicted contribution

to abundance. The fundamental assumption is that training induces a structured manifold in the embedded space, where sequences with similar abundance values align closer along a geodesic path. This enables a meaningful ordering of residues, where lower-ranked residues correspond to lower predicted abundance and vice versa. The projection was trained using the start token embeddings, as these integrate sequence-wide information, and correctness was assessed via Spearman correlation with abundance targets. Using this embedded ordering, we designed a mutation strategy: residues with the lowest order values were selected for substitution, aiming to increase predicted abundance. Guide sequences of the ten highest-abundance proteins provided substitution candidates, where each selected residue was replaced by the closest-matching residue from a guide sequence after applying a fixed shift in ordering. Control experiments included random substitutions to assess baseline effects (Figure 15, panel A). Variants with 80% identity created with this substitution strategy displayed an increased predicted abundance of more than 600%, while introducing random substitutions at the same rate incurred an overall negative impact on predicted abundance (Figure 15, panel B).

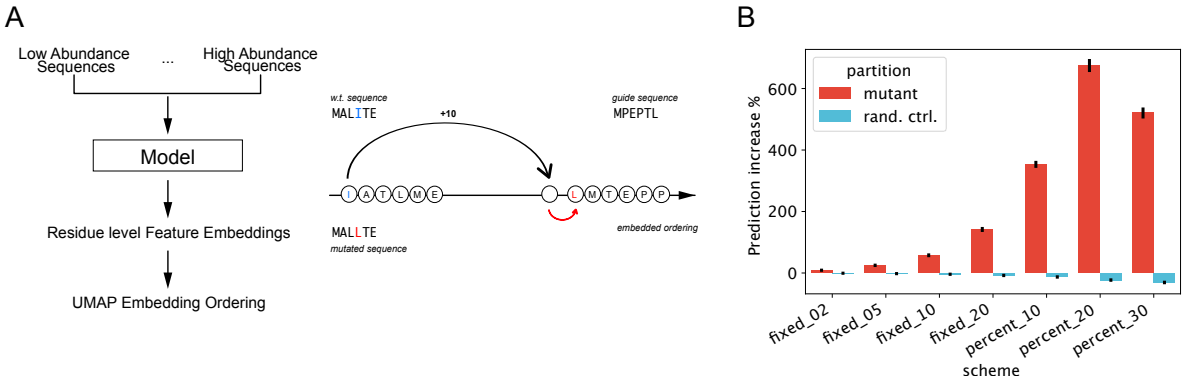

Figure 15: MGEM substitutions enhance predicted abundance of low-abundance proteins.

(A) The MGEM protocol extracts residue-level feature embeddings for all the sequences, reflecting their ordering from low to high abundance. These embeddings are projected into a one-dimensional UMAP space, where residues are ranked based on their projection values. Substitutions are introduced by replacing the lowest-ranked residues in the query sequence with the closest corresponding residues from the guide sequence after applying a small shift in the projection values. (B) Percentage increase of the predicted abundance of MGEM variants, compared to variants with random substitutions. Figure reproduced and altered from Paper III (Protein Science 2025).

While this result was intriguing on its own, it did not yet provide the full picture. One key concern is whether the observed increase in predicted abundance reflects meaningful biological features or is merely an artifact of the model's training scheme. To address this, we sought to investigate the structural and dynamic consequences of the mutations through molecular dynamics simulation. Specifically, we selected 100 non-membrane enzymes with their variants and performed molecular dynamics simulations to assess how the introduced substitutions influenced protein dynamics. This analysis allowed us to determine whether the predicted abundance changes were accompanied by meaningful alterations in protein behavior. Given

the high computational cost of molecular dynamics simulations, a fixed simulation time of 100 ns was chosen for all protein variants and their corresponding wild-type sequences. However, our model does not capture the full variation in protein abundance and lacks explicit knowledge of protein folding and stability. As a result, some introduced mutations may have led to structural destabilization. To ensure a meaningful comparison, we only considered simulations where both the wild-type and mutant variants converged to a stable final conformation within the 100 ns trajectory. Out of the 100 simulated wild-type and variant pairs, 46 met the convergence criterion and were retained for further analysis. Among these, 33% of variants exhibited a significant decrease in Root Mean Square Fluctuation (RMSF), and 59% of atomic fluctuations were reduced by at least two standard deviations compared to their wild-type counterparts (Figure 16, panel A). Notably, this reduction in fluctuation was not confined to the mutation sites; approximately 80% of the decrease occurred in non-mutated residues, suggesting that the substitutions impact global protein dynamics rather than just local regions (Figure 16, panel B). Additionally, 84% of the variants displayed a higher degree of intramolecular interactions than their corresponding wild-types (Figure 16, panel C), and solvent-accessible surface area (SASA) was also reduced in variants compared to wild-types. These shifts in dynamic and structural properties are associated with increased thermostability [88,106]. To further investigate this link, we used our predictive model from *Paper II* to estimate changes in optimal growth temperature (OGT). As a group, the variants generated using MGEM exhibited a significant increase in predicted OGT values. Taken together, these findings suggest that the model trained to predict protein abundance implicitly captured features related to conformational stability, which is itself closely tied to thermal stability.

In this project, we set out to develop a deep learning model for predicting protein abundance. More broadly, our goal was to demonstrate the potential of machine learning in modeling cellular and systemic properties using only the protein sequence, rather than being limited to physicochemical characteristics. By analyzing the learned features, we found that the model captured information related to protein stability, suggesting a connection between protein abundance and intrinsic physical properties.

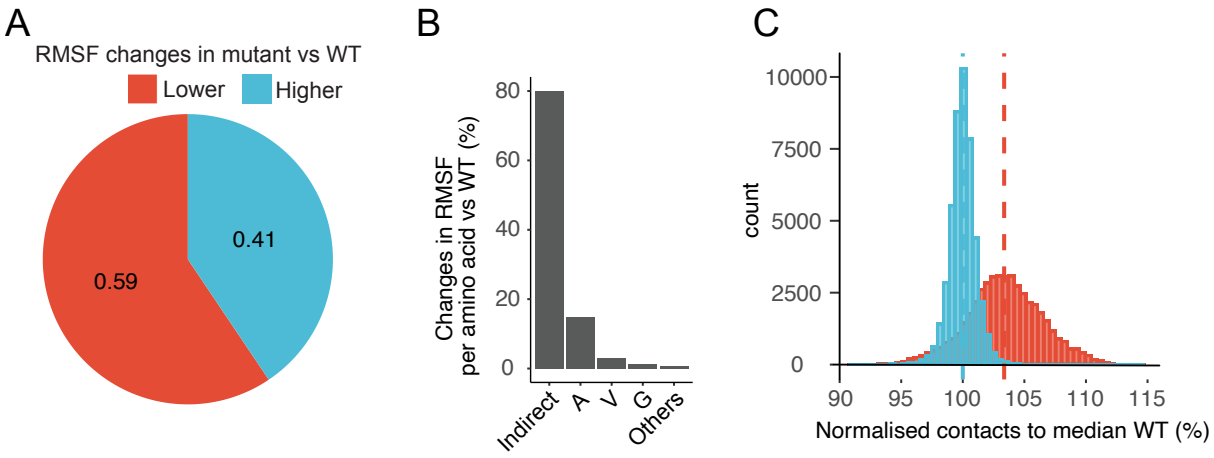

Figure 16: MGEM variants exhibit increased rigidity and enhanced residue contacts.

(A) Distribution of residue fluctuations in MGEM variants compared to wild-type (WT). The fraction of residues with significantly lower root mean square fluctuation (RMSF) values (≥2 standard deviations below WT) is shown in red, while those with increased fluctuations are in blue. (B) The proportion of significant RMSF changes (absolute z-score >2) per introduced mutation. "Indirect" refers to regions of the protein sequence without substitutions. (C) Normalized contact distribution between residues in WT and MGEM variants that increase protein abundance. Contacts are measured within an 8 Å proximity of the carbon backbone, using frames from the second half of a 100 ns trajectory. Figure reproduced from Paper III (Protein Science 2025).

COMPSS: a Scoring Metric to Maximize Yield of Functional Generated Proteins (Paper IV)

In the previous projects *Papers II & III*, we demonstrated how deep learning models could aid the selection phase of protein engineering by predicting specific properties, such as physicochemical characteristics like thermal stability or systemic properties like protein abundance. However, our findings from *Paper I* suggested that selecting for these properties alone may not be sufficient. Despite 19 out of 60 tested variants being expressed and active, more than two-thirds either failed to express, could not be purified, or were inactive. This highlights a critical challenge in protein engineering: ensuring that selected variants are not only optimized for specific traits but are also functional and expressible. To address this, we aimed to develop a selection method that would increase the yield of experimentally active variants in each iteration of the protein engineering cycle. Recent advances in machine learning for biotechnology have led to the emergence of increasingly large and powerful models, some with billions of parameters trained on vast datasets comprising billions of sequences[107,108]. These models have been applied to complex tasks such as inverse protein folding[109], where a sequence is inferred from a structure, and protein language modeling[103,110,111], which captures how evolution shapes protein function. Some have even tackled problems once thought nearly impossible, such as protein folding[28]. The success of these models in their respective fields, driven by both scale and data diversity, is remarkable. Given their extensive knowledge of protein structure and function, we hypothesize that these models encode biologically relevant features that influence both expression and activity and therefore could serve as a powerful framework for improving the selection of functional proteins generated by generative models.

To assess the ability of large-scale protein models to aid in the selection of functional proteins, we tested 10 different metrics, including sequence alignment scores (Identity and BLOSUM62) and a structure-based relaxation metric (Rosetta-relax)[18] (Figure 17). We chose to include these three different modalities as we hypothesized that each encodes different types of information towards the likelihood of a novel sequence being active. The sequence based metrics encode rudimentary features like the inclusion of methionine at the N-terminal to more complex language features in the language models. The alignment metrics will encode evolutionary aspects of the proteins. Finally the structure metrics will encode the structurally relevant features. For sequence generation, we used two deep learning generative models, ProteinGAN, which we developed in *Paper I*, and ESM-MSA[112], as well as a bioinformatics-based ancestral sequence reconstruction (ASR) approach[113].

While ESM-MSA is not a generative model in the traditional sense, it can approximate generative capabilities when sequences are iteratively masked and predicted[114,115]. Similarly, ASR is not typically classified as a generative method, as it is constrained by the statistical properties of phylogenetic trees, but it has been successfully applied to resurrect ancient proteins and engineer new functional variants[51–53,116].

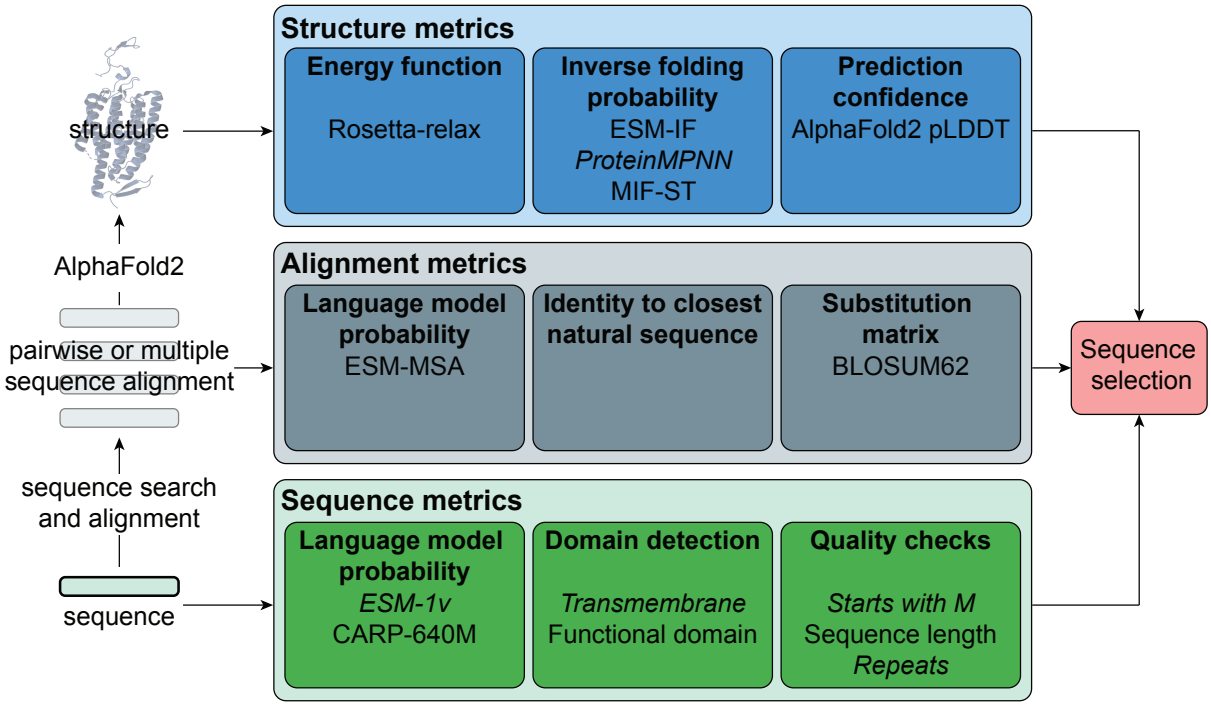

Figure 17: Three modalities, sequence, alignment and structure were tested as scoring metrics.

The metrics were evaluated first for their individual ability to select sequences to maximize their likelihood of being experimentally active. Figure reproduced from Paper IV (Nature Biotechnology 2024).

To experimentally validate our findings, we focused on two protein families: Malate Dehydrogenase (MDH) and Copper Superoxide Dismutase (CuSOD). The screening capability of different models was evaluated by testing the activity of 144 sequences, with 18 sequences per generative model per protein family, along with 18 wild-type sequences for each family as controls.

No single metric consistently outperformed all others. However, certain categories of metrics performed notably well. Inverse folding models (ESM-IF[117], ProteinMPNN[109], and MIF-ST[118]) and structure-based relaxation (Rosetta-relax) demonstrated the strongest predictive capabilities. While Rosetta-relax performed well, it required significantly more computational resources than ProteinMPNN, which exhibited comparable performance. In contrast, traditional alignment-based metrics, such as sequence identity and BLOSUM62, performed the worst overall. Interestingly, AlphaFold2 pLDDT[28], which measures structure prediction confidence, showed strong predictive accuracy for CuSOD but performed poorly for MDH sequences. Among protein language models, ESM-1v[110] and CARP-640M[119] had similar performance, though ESM-1v was more consistent across sequences generated by different models. Given the relatively low correlation between the scores for the language models and the inverse folding models, and the superior performance of both of them compared to sequence identity and AlphaFold2 pLDDT scores (Figure 18), we selected ProteinMPNN and ESM-1v

as the most promising candidates for further calibration of a joint scoring metric. It is worth noting, however, that the ProteinMPNN metric relies on structural inputs generated using AlphaFold2.

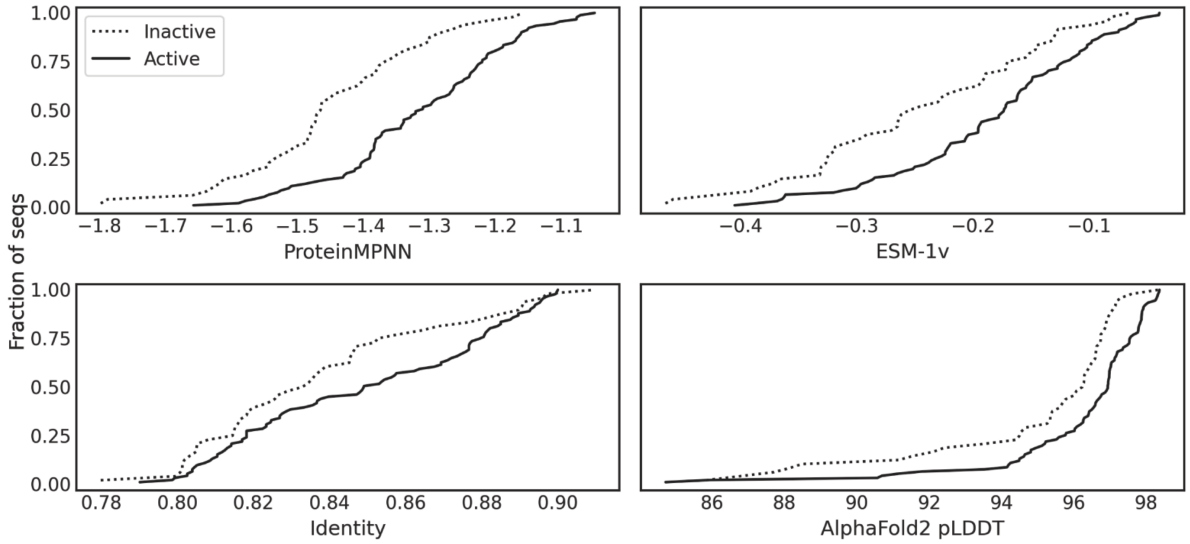

Figure 18: Inverse folding and protein language models provide greater insight into protein activity than sequence identity or AlphaFold2 confidence.

Cumulative distribution plots comparing the scores of sequences that were experimentally active (solid lines) and inactive (dotted lines) protein sequences across four different metrics: (top left) ProteinMPNN, (top right) ESM-1v, (bottom left) sequence identity to the wild-type, and (bottom right) AlphaFold2 pLDDT. The inverse folding model (ProteinMPNN) and protein language model (ESM-1v) show a clearer separation between active and inactive sequences, suggesting they capture functional determinants more effectively than sequence identity or AlphaFold2 confidence scores. Figure reproduced from Paper IV (Nature Biotechnology 2024).

To maximize performance and leverage the strengths of both ESM-1v and ProteinMPNN as filtering metrics, we combined them into a COmposite Metric for Protein Sequence Selection (COMPSS), as a sequential filtering pipeline, preceded by a rudimentary identity filter to ensure all sequences remained within the same identity range. The order of the filters was chosen based on computational efficiency: identity scoring, which has the lowest computational cost, was applied first, followed by ESM-1v, and finally ProteinMPNN (Figure 19, panel A). The filtering thresholds for each step were optimized for each enzyme family to maximize the enrichment of active sequences. To validate this workflow, we generated a new set of sequences using ProteinGAN (*Paper I*) and ESM-MSA. From each model and enzyme family, 18 sequences that passed the filtering criteria were randomly selected. Additionally, for each selected sequence, a closely related variant (within 1% sequence identity) that failed the ESM-1v filter was included as a control to assess the filtering's ability to distinguish between closely related but functionally distinct sequences (Figure 19, panel B). The selected enzyme sequences exhibited high *in vitro* activity, with 94% of ESM-MSA-derived CuSODs and all MDH variants being active. Overall, 74% of the generated sequences were active, representing

a 77% higher success rate than control sequences that failed the selection filter (Figure 19, panel C). Furthermore, 83% of active sequences selected by COMPSS maintained activity levels within an order of magnitude of their wild-type counterparts.

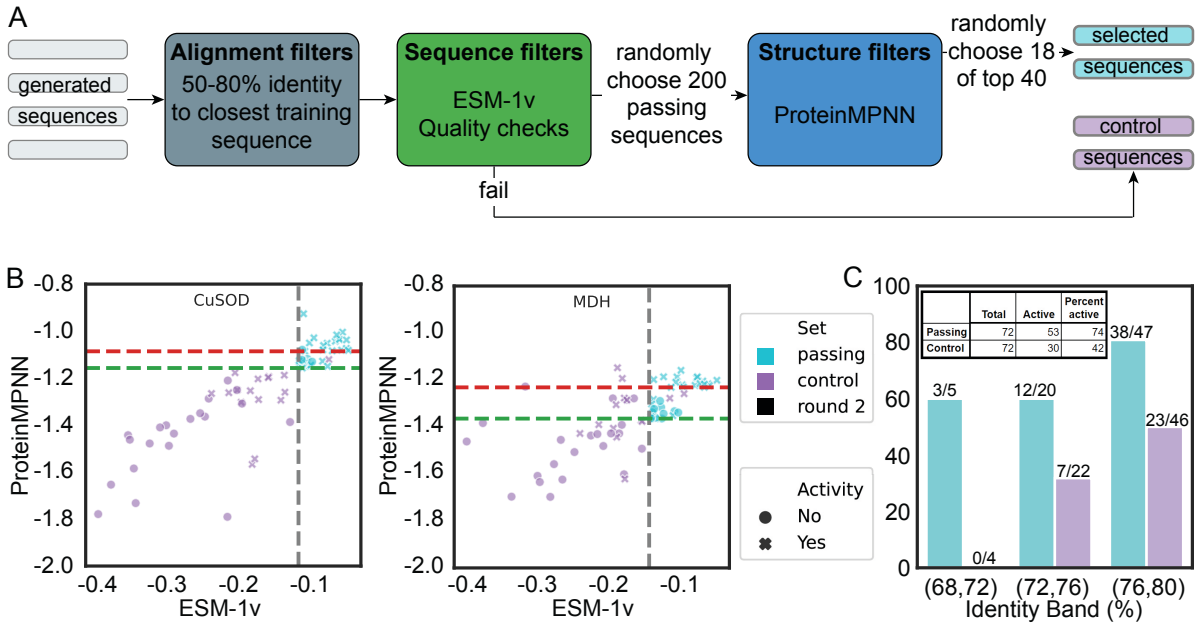

Figure 19: The COMPSS pipeline enhances sequence selection and performance.

(A) Overview of the COMPSS filtering pipeline used for selection. (B) Comparison of ESM-1v and ProteinMPNN scores for selected sequences in two enzyme families: CuSOD and MDH. Selected sequences (teal) and control sequences (violet) are plotted. The vertical dashed gray line represents the top 10th percentile cutoff for ESM-1v scores, calculated based on test sequences. The horizontal dashed lines correspond to ProteinMPNN scores for the 40th-ranked sequence in a batch of 200 candidates. The lower green line marks the ProteinGAN model's score threshold, while the upper red line marks the ESM-MSA threshold. Control sequences that appear to the right of the gray line are those that passed ESM-1v scoring but failed at least one quality check. (C) Proportion of active enzymes across identity bands. Sequences are grouped by sequence identity bands (68–72%, 72–76%, and 76–80%). Bars indicate the number of active sequences in selected (teal) and control (violet) sets. Figure reproduced from Paper IV (Nature Biotechnology 2024).

In this project, we aimed to develop a selection step capable of identifying functional sequences generated by deep generative models. To this end, we developed COMPSS, a filtering workflow that integrates protein language modeling and inverse folding approaches. Our results demonstrate that COMPSS effectively enriches active sequences, increasing the likelihood of selecting functional candidates for further optimization and engineering. This highlights its potential as a valuable tool in the protein engineering cycle, improving the efficiency of selecting viable protein variants.

Putting It All Together: Streamlining the Protein Engineering Cycle

Enhancing Enzyme Phenotypic Properties with Machine Learning (Paper V)

In the previous chapters, we have explored individual components of the protein engineering cycle. In *Paper I*, we focused on generative models as a means to replace the diversification step, while *Papers II–IV* demonstrated how different models can enhance the selection phase. However, a key objective of my PhD studies has been to investigate the potential of machine learning in a more integrated, holistic approach to protein engineering. Specifically, our goal was to develop a machine learning-driven workflow capable of handling the entire protein engineering cycle internally, creating from an experimental point of view a linear workflow.

For this final project, we aimed to combine the insights gained from previous studies into a unified protein engineering framework. Given our success in predicting thermal stability properties such as T_m and T_{opt} in *Paper II*, we chose thermal stability as the primary target for optimization. As discussed in *Paper II*, one of the main challenges in this area is the limited availability of labeled data. Despite datasets like the OGT dataset being relatively large, they still represent less than 1% of the total available sequence data (Figure 20, panel A). Additionally, annotated datasets are heavily skewed toward low-temperature proteins. When attempting to expand available data using our regression model to predict thermophilic sequences, we further observed that many protein families (EggNOG[120] annotated Clusters of Orthologous Groups (COGs)) lack any predicted thermophilic representatives (Figure 20, panel B). Thus for our optimization framework we wanted to be able to extend thermal adaptation to these families that were predicted to not have it.

To fully leverage our accumulated knowledge from the previous projects we developed new models for this task rather than integrating the previous models. Although these new models are based on similar architectures, their training strategies and datasets have been specifically optimized for engineering thermophilic proteins.

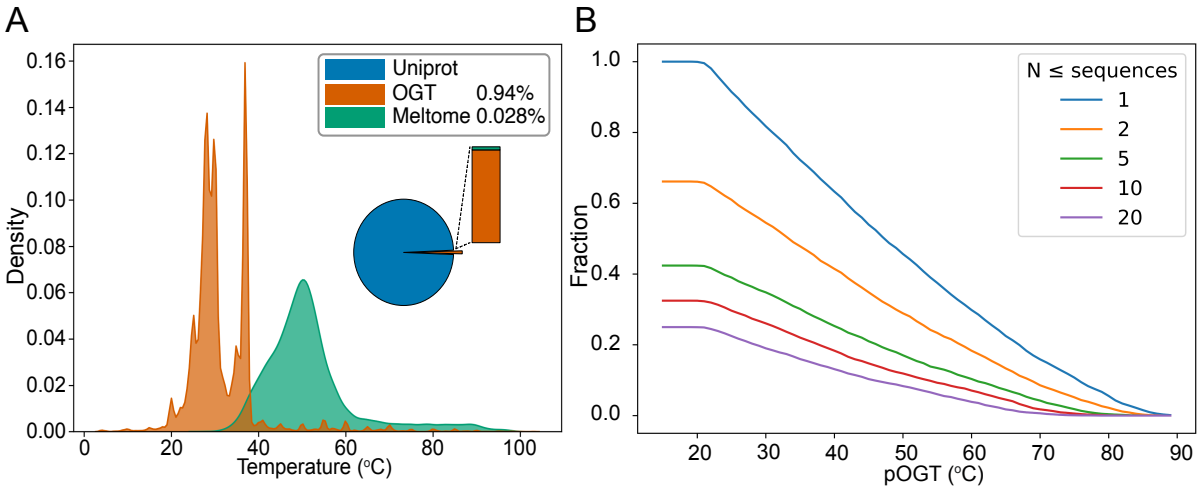

Figure 20: Limited sequence data have thermal stability annotations, and few COGs are predicted to contain thermophiles.

(A) Temperature distribution of available datasets measuring protein thermal properties. The Optimal Growth Temperature (OGT) dataset represents organism-level thermal adaptation, while the Meltome Atlas (T_m dataset) measures protein melting temperatures. The fraction of sequences in these datasets relative to the entire UniProt database is also shown. (B) COG-level distribution of predicted thermophilic sequences. The fraction of EggNOG COGs that contain at least N sequences with a predicted temperature above the threshold on the X-axis. Predictions were made using a deep learning model (adapted from Paper II) trained on the OGT dataset. Swiss-Prot[121] sequences longer than 512 amino acids were excluded from the analysis.

In any machine learning project, the quality and relevance of the training data are crucial for success. Here, we again use the OGT dataset from *Paper II* to train an improved version of our OGT regression model that will serve as our selection model. Additionally, for our generative model that will generate the variant libraries, we constructed two datasets containing both mesophilic (OGT <45°C) and thermophilic (OGT >60°) sequences:

- **Pretraining dataset:** ~65,000 mesophilic sequences from 15 enzymatic COGs and ~65,000 thermophilic sequences from 1,096 enzymatic COGs, ensuring that mesophilic and thermophilic sequences originate from entirely different COGs.
- **Finetuning dataset:** 1,098 mesophilic MDH sequences and 87 thermophilic MDH sequences, which were not part of the pretraing set.

Both datasets are considered **unpaired**, meaning that sequences in the mesophilic and thermophilic groups do not have direct one-to-one counterparts in the opposing group (Figure 21). For the **pretraining dataset**, this unpaired nature is clear since mesophiles and thermophiles come from entirely different COGs, preventing any direct orthology between sequences. For the **finetuning dataset**, one might assume pairing is possible since all sequences belong to the same protein family (MDH). However, the large discrepancy in sample sizes, where mesophilic sequences vastly outnumber thermophilic ones, complicates direct pairing. If we attempted to create paired data, multiple mesophilic sequences would need to be mapped to the same thermophile, reducing sequence diversity and potentially biasing the

model. To preserve diversity and avoid overfitting, we also treat the finetuning dataset as unpaired.

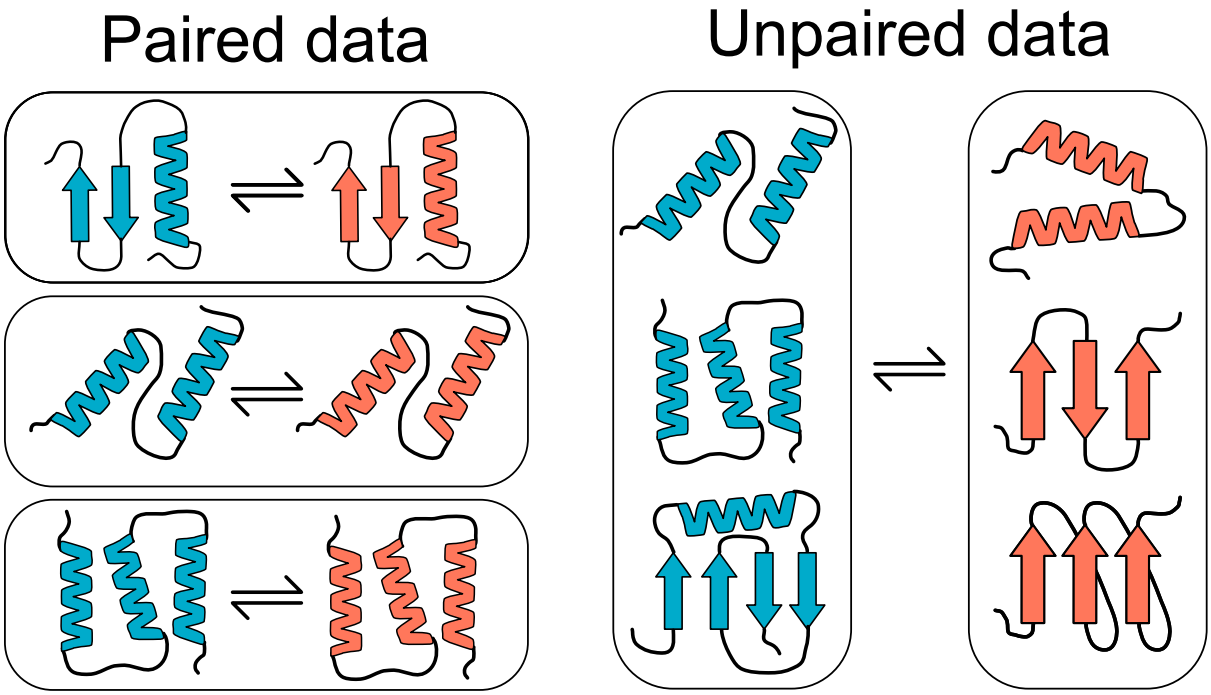

Figure 21: Concept of paired and unpaired data in protein analysis.

Paired data (left side) consists of corresponding protein sequences or structures that are explicitly linked, such as homologous proteins with known evolutionary relationships or experimentally characterized variants of the same protein. These pairs allow for direct comparisons, such as sequence-function relationships, structural stability, or evolutionary constraints. Unpaired data (on the right side) lack explicit one-to-one correspondences; instead, the groups are related many-to-many. These datasets consist of individual protein sequences, without evolutionary relatedness (non-homologous), or structures without predefined pairings, requiring inference methods to establish relationships or patterns.

Due to the unpaired nature of our datasets, it is not possible to train a direct mapping function between mesophilic and thermophilic sequences, as seen in machine translation. Instead, we once again draw inspiration from adversarial training, which we previously used in *Paper I*. However, unlike traditional GANs that generate data from noise, our approach starts with distinct mesophilic sequences and aims to implicitly map them to the thermophilic distribution while preserving the enzymatic function. Rather than using the same GAN architecture as before, we took inspiration from CycleGAN[122], a variant of GAN specifically designed for training with unpaired data to jointly learn to map between X (mesophiles) and Y (thermophiles) using two mapping functions G:X→Y and F:Y→X. The adversarial training is parametrized by discriminator functions D_x and D_y for the respective mappings from X→Y and Y→X. The adversarial learning objective is then given by:

$$\mathcal{L}_{D_Y} = -\mathbb{E}_{y \sim p_{data}(y)} \left[\log \left(D_Y \left(y \right) \right) \right] - \mathbb{E}_{x \sim p_{data}(x)} \left[\log \left(1 - D_Y \left(G \left(x \right) \right) \right) \right] \quad [10]$$

$$\mathcal{L}_{D_X} = -\mathbb{E}_{x \sim p_{data}(x)} \left[\log \left(D_X \left(x \right) \right) \right] - \mathbb{E}_{y \sim p_{data}(y)} \left[\log \left(1 - D_X \left(F \left(y \right) \right) \right) \right] \quad [11]$$

$$\mathcal{L}_G = -\mathbb{E}_{x \sim p_{data}(x)} \left[\log \left(D_Y \left(G \left(x_i \right) \right) \right) \right] \quad [12]$$

$$\mathcal{L}_F = -\mathbb{E}_{y \sim p_{data}(y)} \left[\log \left(D_X \left(F \left(y \right) \right) \right) \right] \quad [13]$$

The expectation is taken over the empirical distributions. Additionally, to help maintain consistency between the mapped sequences and the original distributions, a cycle consistency loss is added:

$$\mathcal{L}_{cycle} \left(G, F \right) = \mathbb{E}_{x \sim p_{data}(x)} \left[\left(F \left(G \left(x \right) \right) - x \right)^2 \right] + \mathbb{E}_{y \sim p_{data}(y)} \left[\left(G \left(F \left(y \right) \right) - y \right)^2 \right] \quad [14]$$

The cycle consistency loss further helps in avoiding mode collapse by keeping essential features in each sequence. Given the limited size of our datasets, we also incorporated transfer learning, building on our successes in *Papers II & III*. For this, we selected ESM1v, as we previously demonstrated its effectiveness in identifying functional proteins in *Paper IV*. However, directly integrating transfer learning into GAN training poses challenges. Namely, the risk of imbalance, where either the generator or discriminator becomes stronger than the other, leading to unstable training[56]. To address this, we implemented a teacher-student setup[123], which adaptively distills knowledge from ESM1v during adversarial training through the additional loss:

$$\mathcal{L}_{EVO} = \mathbb{E}_{x \sim pdata(x)} \left[\sum_i CE \left(G(x)_i, x_i \right) \cdot l_i \right] + \mathbb{E}_{y \sim pdata(y)} \left[\sum_i CE \left(F(y)_i, y_i \right) \cdot l_i \right] \quad [15]$$

where *CE* denotes cross-entropy and l_i is the likelihood of the wild-type amino acid at each position *i* as predicted by the ESM1v model.

This approach ensures that transfer learning is applied dynamically, maintaining stability throughout training. Similar techniques of dynamically introducing transfer learning have seen success in previous works[124].

By combining CycleGAN training loss with the adaptive distillation loss of ESM1v, the full loss is then the Adaptive Adversarial Cycle and Evolutionary consistency loss (AACE):

$$\mathcal{L}_{AACE} = \mathcal{L}_{GAN}(G, D_Y, X, Y) + \mathcal{L}_{GAN}(F, D_X, Y, X) + \mathcal{L}_{Cycle}(G, F) + \lambda \mathcal{L}_{EVO}(G, F) \quad [16]$$

The EVO loss is dynamically scaled by λ which is controlled using a PID controller that I introduced and tested for neural network applications in *Paper V*. The scaling of the EVO loss further helps in controlling the rate of substitutions that each generator introduces (Figure 22, panel A). After training the generative model on both the pretraining and finetuning datasets, we integrated the trained generator with the regression model to construct the THermal Optimizing Representations (THOR) framework, designed to optimize proteins for thermal stability (Figure 22, panel B). The framework was assembled into two versions representing the two data sets:

- THOR-OG80: Generator trained only on the pretraining data set
- THOR-MDH80: Generator trained on the pretraining set and finetuned on the finetuning set.

To validate THOR as a framework for engineering thermally stable proteins, MDH was chosen as the candidate family. This selection allowed us to assess the framework's ability to generalize thermal adaptation to the unseen MDH orthologous group with THOR-OG80, which has not seen any MDH sequences during training, as well as its ability to capture thermal adaptation within the MDH orthologous group with THOR-MDH80 (that has been finetuned on MDH sequences).

Evaluation was done using three wild-type sequences, selected from the test sequences used as controls in *Paper IV*. For each wild-type sequence, THOR-OG80 produced three optimized variants. Meanwhile, the THOR-MDH80 generator generated 39, 4, and 2 variants for the respective wild-type sequences.

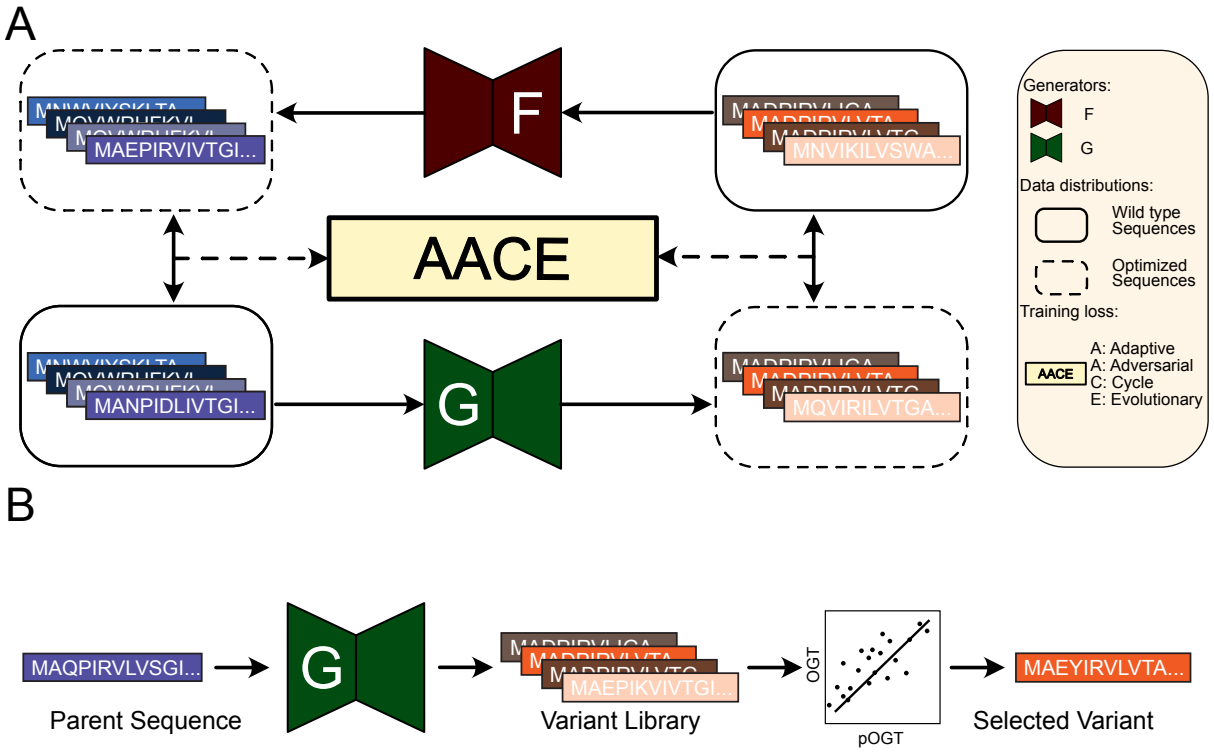

Figure 22: THOR: a unified framework for optimization of protein thermal tolerance.

(A) To enable the generator to learn from unpaired mesophilic and thermophilic sequence distributions, we use a CycleGAN-inspired framework with two mapping functions, G and F. These functions establish bidirectional transformations between the mesophilic and thermophilic domains through adversarial training. To maintain sequence integrity, the adversarial process is regularized by cycle consistency, ensuring the mappings between domains remain reversible and evolutionary consistent, preserving essential identity to wild-type sequences. Together, these elements form an adaptive adversarial training framework incorporating cycle and evolutionary consistency (AACE). (B) The generator and regression model are paired together to form the THOR framework that can optimize unseen mesophilic sequences by generating variant libraries of thermophiles, after which the regression model will select the best candidates.

The variants were then evaluated using Differential Scanning Flourimetry (DSF) thermal assay[125,126]. The THOR-OG80 generator successfully increased T_m for two of the three wild-type sequences, with a maximum increase of 15.4°C and 17.2°C, respectively. In contrast, the THOR-MDH80 generator improved the T_m for all three wild-types, achieving increases of 36.4°C, 30.9°C, and 18.5°C, respectively (Figure 23, panel A). The THOR-MDH80 variants were further evaluated to find the temperature where they would be irreversibly denatured, which was tested by heat treating the variants and their respective wild-types for 10 min at ten-degree intervals. Each enzyme then had its activity evaluated after cooling down to room temperature. For two out of the three wild-type sequences, engineered variants retained activity at or above the wild-type's heat tolerance (Figure 23, panel B). For one of the wild-type sequences, three of its variants had a 20°C increase to the temperature at which they would be irreversibly unfolded.

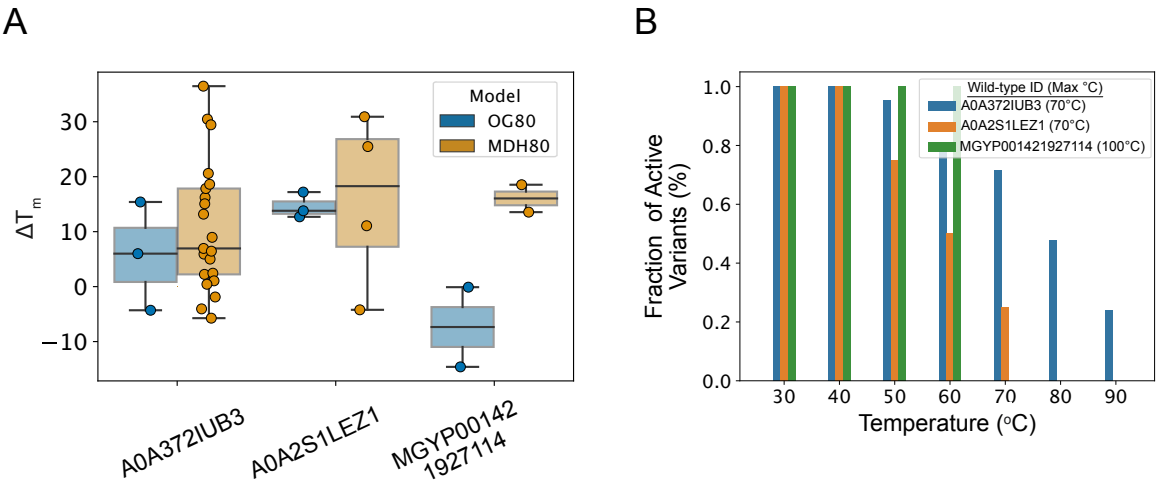

Figure 23: Experimental validation confirms increased thermal tolerance of THOR-optimized variants.

(A) Melting temperature shifts for THOR-optimized MDH variants compared to their corresponding wild-type sequences. (B) Fraction of active variants after heat treatment. Variants generated by the fine-tuned MDH80 model retained enzymatic activity at room temperature even after 10 minutes of heat treatment, demonstrating improved thermal tolerance.

In this project, we set out to create a unified framework for engineering thermally stable proteins, one that combines a generative model introducing sequence diversity with a regression model that identifies candidate sequences with the highest predicted thermal adaptation. Additionally, to ensure the generated sequences remain functional, we integrated a version of COMPSS directly into the generator's training process. By taking this holistic approach to integrating machine learning into the protein engineering cycle, we moved closer to a framework that reduces the need for multiple iterative steps, making the optimization process more efficient and streamlined.

Conclusions and Outlook

The overarching goal of this thesis was to explore how generative machine learning can aid and streamline protein engineering by leveraging data-driven approaches for sequence design, property prediction, and functional selection. Through five interconnected studies, I demonstrated how deep learning models can contribute to different stages of the protein engineering cycle.

In *Paper I*, we focused on the first step of the cycle, sequence diversification, by developing ProteinGAN, a deep generative model capable of generating diverse yet functional proteins. The ability to rapidly sample a vast number of protein sequences *in silico* has immense potential, but experimental validation remains a bottleneck. Synthesizing and testing all generated sequences is infeasible, emphasizing the need for efficient selection strategies.

Thus, in *Papers II-IV*, the focus shifted to functional selection, the challenge of identifying viable candidates from the vast sequence space. One of the most commonly optimized properties in protein engineering is thermal stability, as increased stability is often required either for the application or to facilitate further engineering that can often be destabilizing. In *Paper II*, we developed a regression model capable of predicting three key measures of thermal adaptation OGT, T_m, and T_{opt}. While the study did not explicitly apply this model to select sequences, its performance (R^2: 59%, 58%, 57%) suggests it could serve as an effective filtering tool, reducing the number of experimental validations needed. While ProteinGAN was able to generate functional proteins at a high rate, there was still room for improvement, as approximately two-thirds of generated sequences lacked activity. To address this, we developed COMPSS in *Paper IV*, a scoring metric that integrates sequence-based protein language modeling with structural inverse folding scores. This metric successfully enriched active sequences, increasing the proportion of functional proteins by up to 77%. Together, the studies *I,II,IV* demonstrated the ability of deep learning to contribute to both key aspects of protein engineering: diversification and selection.

However, the true potential of these methods is realized when they are combined into a unified framework, as demonstrated in *Paper V*. Here, we introduced THOR, a framework for designing thermally stable proteins by integrating generative modeling, functional scoring, and predictive selection. A new generative model was trained specifically to optimize thermal adaptation, with part of COMPSS directly incorporated into its training process to ensure functional viability. Finally, our regression model was used to select the most promising candidates for experimental validation. Using this approach, we successfully increased the T_m of three wild-type enzymes by 36°C, 30°C, and 18°C, demonstrating the power of deep generative models in protein engineering.

Beyond their engineering applications, deep learning models also offer a powerful bottom-up approach to discovering new biological relationships. While deep models are often criticized as "black boxes" due to their lack of interpretability, various techniques can be used to probe

their learned representations. In *Paper II*, we applied sequence occlusion analysis to assess the contributions of different residues to T_{opt} predictions. This analysis revealed biologically meaningful features, aligning with known motifs such as amino acid propensity and the link between thermal stability and helical content. In *Paper III*, we took this further by analyzing attention profiles in a protein abundance prediction model, correlating learned features with amino acid properties. We also developed a novel approach to leverage the model's internal sequence representations to guide mutations that increase predicted abundance. Molecular dynamics simulations of these designed mutations showed increased rigidity, reinforcing the connection between conformational stability and protein abundance.

Collectively, these studies illustrate how generative machine learning can reshape protein engineering by enabling efficient sequence design, property prediction, and functional selection. Furthermore, deep learning is not only a powerful engineering tool, it also has the potential to uncover new biological insights.

While this thesis's findings represent significant advances, the field of machine learning-driven biotechnology is evolving rapidly, both through refinements of existing methods and the emergence of new frontiers. Until now, this work has focused on protein engineering, but the field is now shifting towards the generative design of DNA sequences. As these approaches continue to mature, they open the door to reprogramming life itself, unlocking unprecedented possibilities in synthetic biology, drug development, and beyond.

Acknowledgments

First and foremost, I express my deepest gratitude to my supervisor, Aleksej, for his unwavering trust, guidance, encouragement, and patience. His mentorship has been invaluable not only in shaping my scientific thinking but also in allowing me the freedom to explore my ideas while providing the critical grounding necessary for meaningful research. I am profoundly thankful for his support throughout this journey.

I would also like to extend my sincere gratitude to my co-supervisor, Pernilla, whose expertise and unwavering support have been instrumental in shaping my research. From the very beginning, she helped me plan, structure, and navigate the many challenges of this project, ensuring that everything was set up and executed efficiently. Her insights and meticulous approach not only kept me on track but also helped ground my work in sound chemistry.

I would also like to express my deepest gratitude to Filip, who played an instrumental role in guiding me through the early phase of my PhD, a time that was particularly challenging due to the remote work policies in the wake of the pandemic. His support, patience, and willingness to share his knowledge made an immense difference as I navigated those initial hurdles. Beyond that, he has continued to be a source of invaluable guidance throughout my journey, always ready to offer insights, advice, and encouragement. And in the final stretch, his keen eye and thoughtful feedback helped refine and polish this thesis. I truly appreciate everything he has done for me. Thank you, Filip.

I am equally grateful to my incredible group members and collaborators: Nikos, Xiaozhi, Sasha, Danish, Jan, Martin, Gang, Mariia, Clara, Kevin, and Sean, among others. Beyond the rewarding projects we've worked on together, each of you has contributed to my growth in ways that go far beyond research. I have learned much from our discussions, challenges, and shared successes. Thank you for your generosity in sharing your knowledge, insights, and, most importantly, your friendship.

To all SysBio members, I cannot thank you enough for creating such a collegial and welcoming atmosphere. Your feedback, support, and words of wisdom have helped shape my academic and personal growth.

Finally, my heartfelt thanks go to my wife, family, and friends, whose constant support has meant the world to me. Your encouragement, patience, and belief in me have carried me through the highs and lows of this journey. Thank you to everyone that I have shared a conversation about science, life, literature, philosophy, entertainment. Every exchange, no matter how small, has left an imprint on me, and for that, I am profoundly grateful.

This work would not have been possible without you. Thank you from the bottom of my heart.

References

1. Axe, D. D. Estimating the prevalence of protein sequences adopting functional enzyme folds. *J. Mol. Biol.* **341**, 1295–1315 (2004).
2. Taverna, D. M. & Goldstein, R. A. Why are proteins marginally stable? *Proteins* **46**, 105–109 (2002).
3. Keefe, A. D. & Szostak, J. W. Functional proteins from a random-sequence library. *Nature* **410**, 715–718 (2001).
4. Rockah-Shmuel, L., Tóth-Petróczy, Á. & Tawfik, D. S. Systematic mapping of protein mutational space by prolonged drift reveals the deleterious effects of seemingly neutral mutations. *PLoS Comput. Biol.* **11**, e1004421 (2015).
5. Guo, H. H., Choe, J. & Loeb, L. A. Protein tolerance to random amino acid change. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 9205–9210 (2004).
6. Gorontzy, T. *et al.* Microbial degradation of explosives and related compounds. *Crit. Rev. Microbiol.* **20**, 265–284 (1994).
7. Singh, B. K. & Walker, A. Microbial degradation of organophosphorus compounds. *FEMS Microbiol. Rev.* **30**, 428–471 (2006).
8. Davies, J. & Davies, D. Origins and evolution of antibiotic resistance. *Microbiol. Mol. Biol. Rev.* **74**, 417–433 (2010).
9. Mohanan, N., Montazer, Z., Sharma, P. K. & Levin, D. B. Microbial and Enzymatic Degradation of Synthetic Plastics. *Front. Microbiol.* **11**, 580709 (2020).
10. Cadwell, R. C. & Joyce, G. F. Randomization of genes by PCR mutagenesis. *Genome Res.* **2**, 28–33 (1992).
11. Dube, D. K. *et al.* Artificial mutants generated by the insertion of random oligonucleotides into the putative nucleoside binding site of the HSV-1 thymidine kinase

gene. *Biochemistry* **30**, 11760–11767 (1991).

12. Moore, J. C. & Arnold, F. H. Directed evolution of a para-nitrobenzyl esterase for aqueous-organic solvents. *Nat. Biotechnol.* **14**, 458–467 (1996).
13. Stemmer, W. P. Rapid evolution of a protein in vitro by DNA shuffling. *Nature* **370**, 389–391 (1994).
14. Reetz, M. T. *et al.* Expanding the substrate scope of enzymes: combining mutations obtained by CASTing. *Chemistry* **12**, 6031–6038 (2006).
15. Clouthier, C. M., Kayser, M. M. & Reetz, M. T. Designing new Baeyer—Villiger monooxygenases using restricted CASTing. *ChemInform* **38**, (2007).
16. Parra, L. P., Agudo, R. & Reetz, M. T. Directed evolution by using iterative saturation mutagenesis based on multiresidue sites. *Chembiochem* **14**, 2301–2309 (2013).
17. Krieger, E., Koraimann, G. & Vriend, G. Increasing the precision of comparative models with YASARA NOVA--a self-parameterizing force field. *Proteins* **47**, 393–402 (2002).
18. Das, R. & Baker, D. Macromolecular modeling with rosetta. *Annu. Rev. Biochem.* **77**, 363–382 (2008).
19. Buß, O., Rudat, J. & Ochsenreither, K. FoldX as protein engineering tool: Better than random based approaches? *Comput. Struct. Biotechnol. J.* **16**, 25–33 (2018).
20. Xiong, P., Chen, Q. & Liu, H. Computational protein design under a given backbone structure with the ABACUS statistical energy function. *Methods Mol. Biol.* **1529**, 217–226 (2017).
21. Floor, R. J. *et al.* Computational library design for increasing haloalkane dehalogenase stability. *Chembiochem* **15**, 1660–1672 (2014).
22. Wijma, H. J. *et al.* Computationally designed libraries for rapid enzyme stabilization. *Protein Eng. Des. Sel.* **27**, 49–58 (2014).
23. Khalid, A. *et al.* Breast Cancer Detection and Prevention Using Machine Learning.

Diagnostics (Basel) **13**, (2023).

24. Islam, T. *et al.* Predictive modeling for breast cancer classification in the context of Bangladeshi patients by use of machine learning approach with explainable AI. *Scientific Reports* **14**, 1–17 (2024).
25. Stokes, J. M. *et al.* A Deep Learning Approach to Antibiotic Discovery. *Cell* **180**, 688–702.e13 (2020).
26. Liu, G. *et al.* Deep learning-guided discovery of an antibiotic targeting Acinetobacter baumannii. *Nature Chemical Biology* **19**, 1342–1350 (2023).
27. Loeffler, H. H. *et al.* Reinvent 4: Modern AI–driven generative molecule design. *Journal of Cheminformatics* **16**, 1–16 (2024).
28. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
29. O'Shea, K. & Nash, R. An Introduction to Convolutional Neural Networks. (2015).
30. Vaswani, A. *et al.* Attention is all you need. *arXiv [cs.CL]* (2017) doi:10.48550/ARXIV.1706.03762.
31. Ahmad, S., Kamal, M. Z., Sankaranarayanan, R. & Rao, N. M. Thermostable Bacillus subtilis lipases: in vitro evolution and structural insight. *J Mol Biol* **381**, 324–340 (2008).
32. Goedegebuur, F. *et al.* Improving the thermal stability of cellobiohydrolase Cel7A from Hypocrea jecorina by directed evolution. *J. Biol. Chem.* **292**, 17418–17430 (2017).
33. Kumar, V., Dangi, A. K. & Shukla, P. Engineering Thermostable Microbial Xylanases Toward its Industrial Applications. *Mol Biotechnol* **60**, 226–235 (2018).
34. Arndt, M. A. E. *et al.* Generation of a highly stable, internalizing anti-CD22 single-chain Fv fragment for targeting non-Hodgkin's lymphoma. *Int J Cancer* **107**, 822–829 (2003).
35. Kowalski, J. M., Parekh, R. N., Mao, J. & Wittrup, K. D. Protein folding stability can determine the efficiency of escape from endoplasmic reticulum quality control. *J Biol*

Chem **273**, 19453–19458 (1998).

36. Shusta, E. V., Holler, P. D., Kieke, M. C., Kranz, D. M. & Wittrup, K. D. Directed evolution of a stable scaffold for T-cell receptor engineering. *Nat Biotechnol* **18**, 754–759 (2000).
37. Traxlmayr, M. W. *et al.* Directed evolution of Her2/neu-binding IgG1-Fc for improved stability and resistance to aggregation by using yeast surface display. *Protein Eng Des Sel* **26**, 255–265 (2013).
38. Porebski, B. T. *et al.* Circumventing the stability-function trade-off in an engineered FN3 domain. *Protein Eng Des Sel* **29**, 541–550 (2016).
39. Richard Elliott, J. & Lira, C. T. *Introductory Chemical Engineering Thermodynamics*. (Prentice Hall, 2012).
40. Privalov, P. L. Stability of proteins: small globular proteins. *Adv. Protein Chem.* **33**, 167–241 (1979).
41. Yadav, S. & Ahmad, F. A new method for the determination of stability parameters of proteins from their heat-induced denaturation curves. *Anal. Biochem.* **283**, 207–213 (2000).
42. Sanchez-Ruiz, J. M. Protein kinetic stability. *Biophys Chem* **148**, 1–15 (2010).
43. Karshikoff, A., Nilsson, L. & Ladenstein, R. Rigidity versus flexibility: the dilemma of understanding protein thermal stability. *FEBS J* **282**, 3899–3917 (2015).
44. Suplatov, D., Voevodin, V. & Švedas, V. Robust enzyme design: bioinformatic tools for improved protein stability. *Biotechnol J* **10**, 344–355 (2015).
45. Rader, A. J., Yennamalli, R. M., Harter, A. K. & Sen, T. Z. A rigid network of long-range contacts increases thermostability in a mutant endoglucanase. *J Biomol Struct Dyn* **30**, 628–637 (2012).
46. Radestock, S. & Gohlke, H. Exploiting the Link between Protein Rigidity and

Thermostability for Data-Driven Protein Engineering. *Engineering in Life Sciences* **8**, 507–522 (2008).

47. Radestock, S. & Gohlke, H. Protein rigidity and thermophilic adaptation. *Proteins* **79**, 1089–1108 (2011).
48. Wolynes, P. G., Onuchic, J. N. & Thirumalai, D. Navigating the Folding Routes. *Science* (1995) doi:10.1126/science.7886447.
49. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics* **17**, 333–351 (2016).
50. Petty, N. K. Genome annotation: man versus machine. *Nature Reviews Microbiology* **8**, 762–762 (2010).
51. Furukawa, R., Toma, W., Yamazaki, K. & Akanuma, S. Ancestral sequence reconstruction produces thermally stable enzymes with mesophilic enzyme-like catalytic properties. *Sci Rep* **10**, 15493 (2020).
52. Koblan, L. W. *et al.* Improving cytidine and adenine base editors by expression optimization and ancestral reconstruction. *Nat Biotechnol* **36**, 843–846 (2018).
53. Chen, X. *et al.* Directed reconstruction of a novel ancestral alcohol dehydrogenase featuring shifted pH-profile, enhanced thermostability and expanded substrate spectrum. *Bioresour Technol* **363**, 127886 (2022).
54. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. *arXiv [cs.CV]* (2015) doi:10.48550/ARXIV.1512.03385.
55. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of deep bidirectional Transformers for language understanding. *arXiv [cs.CL]* (2018) doi:10.48550/ARXIV.1810.04805.
56. Goodfellow, I. J. *et al.* Generative Adversarial Networks. *arXiv [stat.ML]* (2014) doi:10.48550/ARXIV.1406.2661.

57. van den Oord, A. *et al.* WaveNet: A generative model for raw audio. *arXiv [cs.SD]* (2016) doi:10.48550/ARXIV.1609.03499.
58. Karras, T., Laine, S. & Aila, T. A style-based generator architecture for generative adversarial networks. *arXiv [cs.NE]* (2018) doi:10.48550/ARXIV.1812.04948.
59. Brown, T. B. *et al.* Language Models are Few-Shot Learners. *arXiv [cs.CL]* (2020) doi:10.48550/ARXIV.2005.14165.
60. Arjovsky, M., Chintala, S. & Bottou, L. Wasserstein GAN. (2017).
61. Durall, R., Chatzimichailidis, A., Labus, P. & Keuper, J. Combating Mode Collapse in GAN training: An Empirical Analysis using Hessian Eigenvalues. (2020).
62. Miyato, T., Kataoka, T., Koyama, M. & Yoshida, Y. Spectral normalization for generative adversarial networks. *arXiv [cs.LG]* (2018).
63. Zhang, H., Goodfellow, I., Metaxas, D. & Odena, A. Self-attention generative adversarial networks. *arXiv [stat.ML]* (2018) doi:10.48550/ARXIV.1805.08318.
64. Romero, P. A., Krause, A. & Arnold, F. H. Navigating the protein fitness landscape with Gaussian processes. *Proc Natl Acad Sci U S A* **110**, E193–201 (2013).
65. Söding, J. Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21**, 951–960 (2005).
66. Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
67. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
68. Chen, Z. *et al.* iFeature: a Python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics* **34**, 2499–2502 (2018).
69. Gado, J. E., Beckham, G. T. & Payne, C. M. Improving Enzyme Optimum Temperature Prediction with Resampling Strategies and Ensemble Learning. *J Chem Inf Model* **60**,

4098–4107 (2020).

70. Min, S., Kim, H., Lee, B. & Yoon, S. Protein transfer learning improves identification of heat shock protein families. *PLoS One* **16**, e0251865 (2021).
71. Yu, C.-H. *et al.* ColGen: An end-to-end deep learning model to predict thermal stability of de novo collagen sequences. *J Mech Behav Biomed Mater* **125**, 104921 (2022).
72. Li, G., Rabe, K. S., Nielsen, J. & Engqvist, M. K. M. Machine Learning Applied to Predicting Microorganism Growth Temperatures and Enzyme Catalytic Optima. *ACS Synth Biol* **8**, 1411–1420 (2019).
73. Jeske, L., Placzek, S., Schomburg, I., Chang, A. & Schomburg, D. BRENDA in 2019: a European ELIXIR core data resource. *Nucleic Acids Res* **47**, D542–D549 (2019).
74. Jarzab, A. *et al.* Meltome atlas-thermal proteome stability across the tree of life. *Nat Methods* **17**, 495–503 (2020).
75. Pelletier, E. D., Jeffries, S. D., Song, K. & Hemmerling, T. M. Comparative Analysis of Machine-Learning Model Performance in Image Analysis: The Impact of Dataset Diversity and Size. *Anesth Analg* **139**, 1332–1339 (2024).
76. Leuenberger, P. *et al.* Cell-wide analysis of protein thermal unfolding reveals determinants of thermostability. *Science* **355**, (2017).
77. Remya, R. K. & Wilscy, M. Pretrained convolutional neural networks as feature extractor for image splicing detection. in *2018 International Conference on Circuits and Systems in Digital Enterprise Technology (ICCSDET)* (IEEE, 2018). doi:10.1109/iccsdet.2018.8821242.
78. Weiss, K., Khoshgoftaar, T. M. & Wang, D. A survey of transfer learning. *Journal of Big Data* **3**, 1–40 (2016).
79. Kim, H. E. *et al.* Transfer learning for medical image classification: a literature review. *BMC Medical Imaging* **22**, 1–13 (2022).

80. Ruder, S., Peters, M. E., Swayamdipta, S. & Wolf, T. Transfer learning in natural language processing. in *Proceedings of the 2019 Conference of the North* (Association for Computational Linguistics, Stroudsburg, PA, USA, 2019). doi:10.18653/v1/n19-5004.

81. Engqvist, M. K. M. Correlating enzyme annotations with a large set of microbial growth temperatures reveals metabolic adaptations to growth at diverse temperatures. *BMC Microbiol* **18**, 177 (2018).

82. Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M. & Church, G. M. Unified rational protein engineering with sequence-based deep representation learning. *Nature Methods* **16**, 1315–1322 (2019).

83. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).

84. Simonyan, K., Vedaldi, A. & Zisserman, A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. (2013).

85. Ribeiro, M. T., Singh, S. & Guestrin, C. 'Why Should I Trust You?': Explaining the Predictions of Any Classifier. (2016).

86. Bahdanau, D., Cho, K. & Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. (2014).

87. Pinney, M. M. *et al.* Parallel molecular mechanisms for enzyme temperature adaptation. *Science* **371**, (2021).

88. Kumar, S., Tsai, C. J. & Nussinov, R. Factors enhancing protein thermostability. *Protein Eng* **13**, 179–191 (2000).

89. Laursen, B. S., Sørensen, H. P., Mortensen, K. K. & Sperling-Petersen, H. U. Initiation of protein synthesis in bacteria. *Microbiol Mol Biol Rev* **69**, 101–123 (2005).

90. Merrick, W. C. & Pavitt, G. D. Protein Synthesis Initiation in Eukaryotic Cells. *Cold*

Spring Harb Perspect Biol **10**, (2018).

91. Verma, M. *et al.* A short translational ramp determines the efficiency of protein synthesis. *Nature Communications* **10**, 1–15 (2019).
92. Vogel, C. *et al.* Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. *Mol Syst Biol* **6**, 400 (2010).
93. Zur, H. & Tuller, T. Transcript features alone enable accurate prediction and understanding of gene expression in S. cerevisiae. *BMC Bioinformatics* **14 Suppl 15**, S1 (2013).
94. Zhao, W., Liu, S., Du, G. & Zhou, J. An efficient expression tag library based on self-assembling amphipathic peptides. *Microbial Cell Factories* **18**, 1–11 (2019).
95. Akashi, H. & Gojobori, T. Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis. *Proc Natl Acad Sci U S A* **99**, 3695–3700 (2002).
96. Raiford, D. W. *et al.* Do amino acid biosynthetic costs constrain protein evolution in Saccharomyces cerevisiae? *J Mol Evol* **67**, 621–630 (2008).
97. Swire, J. Selection on synthesis cost affects interprotein amino acid usage in all three domains of life. *J Mol Evol* **64**, 558–571 (2007).
98. Shantha Raju, T. Proteolysis of Proteins. in *Co? and Post-Translational Modifications of Therapeutic Antibodies and Proteins* 183–202 (John Wiley & Sons, Ltd, 2019).
99. Correa Marrero, M. & Barrio-Hernandez, I. Toward Understanding the Biochemical Determinants of Protein Degradation Rates. *ACS Omega* **6**, 5091–5100 (2021).
100. Weber, M. *et al.* Impact of C-terminal amino acid composition on protein expression in bacteria. *Mol Syst Biol* **16**, e9208 (2020).
101. Huang, Q., Szklarczyk, D., Wang, M., Simonovic, M. & von Mering, C. PaxDb 5.0: Curated Protein Quantification Data Suggests Adaptive Proteome Changes in Yeasts.

Mol Cell Proteomics **22**, 100640 (2023).

102. Ho, B., Baryshnikova, A. & Brown, G. W. Unification of protein abundance datasets yields a quantitative Saccharomyces cerevisiae proteome. *Cell Syst.* **6**, 192–205.e3 (2018).

103. Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. U. S. A.* **118**, e2016239118 (2021).

104. Ibtehaz, N. & Kihara, D. Application of Sequence Embedding in Protein Sequence-Based Predictions. (2021).

105. Sainburg, T., McInnes, L. & Gentner, T. Q. Parametric UMAP embeddings for representation and semi-supervised learning. (2020).

106. Razvi, A. & Scholtz, J. M. Lessons in stability from thermophilic proteins. *Protein Sci* **15**, 1569–1578 (2006).

107. Madani, A. *et al.* Large language models generate functional protein sequences across diverse families. *Nat Biotechnol* **41**, 1099–1106 (2023).

108. Hayes, T. *et al.* Simulating 500 million years of evolution with a language model. *Science* **387**, 850–858 (2025).

109. Dauparas, J. *et al.* Robust deep learning-based protein sequence design using ProteinMPNN. *Science* **378**, 49–56 (2022).

110. Meier, J. *et al.* Language models enable zero-shot prediction of the effects of mutations on protein function. *bioRxiv* 2021.07.09.450648 (2021) doi:10.1101/2021.07.09.450648.

111. Brandes, N., Ofer, D., Peleg, Y., Rappoport, N. & Linial, M. ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics* **38**, 2102–2110 (2022).

112. Rao, R. *et al.* MSA Transformer. *bioRxiv* 2021.02.12.430858 (2021)

doi:10.1101/2021.02.12.430858.

113. Foley, G. *et al.* Engineering indel and substitution variants of diverse and ancient enzymes using Graphical Representation of Ancestral Sequence Predictions (GRASP). *PLoS Comput Biol* **18**, e1010633 (2022).

114. Johnson, S. R., Monaco, S., Massie, K. & Syed, Z. Generating novel protein sequences using Gibbs sampling of masked language models. *bioRxiv* 2021.01.26.428322 (2021) doi:10.1101/2021.01.26.428322.

115. Wang, A. & Cho, K. BERT has a Mouth, and It Must Speak: BERT as a Markov Random Field Language Model. (2019).

116. Merkl, R. & Sterner, R. Ancestral protein reconstruction: techniques and applications. *Biol Chem* **397**, 1–21 (2016).

117. Hsu, C. *et al.* Learning inverse folding from millions of predicted structures. *bioRxiv* 2022.04.10.487779 (2022) doi:10.1101/2022.04.10.487779.

118. Yang, K. K., Zanichelli, N. & Yeh, H. Masked inverse folding with sequence transfer for protein representation learning. *Protein Eng Des Sel* **36**, (2023).

119. Yang, K. K., Fusi, N. & Lu, A. X. Convolutions are competitive with transformers for protein sequence pretraining. *Cell Syst* **15**, 286–294.e2 (2024).

120. Huerta-Cepas, J. *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res* **47**, D309–D314 (2019).

121. UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Res* **53**, D609–D617 (2025).

122. Zhu, J.-Y., Park, T., Isola, P. & Efros, A. A. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. (2017).

123. Sanh, V., Debut, L., Chaumond, J. & Wolf, T. DistilBERT, a distilled version of BERT:

smaller, faster, cheaper and lighter. (2019).

124. Ham, H., Jun, T. J. & Kim, D. Unbalanced GANs: Pre-training the Generator of Generative Adversarial Network using Variational Autoencoder. (2020).

125. Niesen, F. H., Berglund, H. & Vedadi, M. The use of differential scanning fluorimetry to detect ligand interactions that promote protein stability. *Nature Protocols* **2**, 2212–2221 (2007).

126. Protocol for performing and optimizing differential scanning fluorimetry experiments. *STAR Protocols* **4**, 102688 (2023).