Protein Representations: Encoding Biological Information for Machine Learning in Biocatalysis

David Harding-Larsen [a], Jonathan Funk [a], Niklas Gesmar Madsen [a], Hani Gharabli [a], Carlos G. Acevedo-Rocha [a], Stanislav Mazurenko [b, c], Ditte Hededam Welner [a, *]

[a] The Novo Nordisk Center for Biosustainability, Technical University of Denmark, Søltofts Plads, Bygning 220, 2800 Kgs. Lyngby, Denmark
[b] Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Faculty of Science, Masaryk University, Kamenice 5, 625 00 Brno, Czech Republic
[c] International Clinical Research Center, St. Anne's University Hospital Brno, Pekarska 53, 656 91 Brno, Czech Republic
* Corresponding author. E-mail address: diwel@biosustain.dtu.dk

Abstract

Enzymes offer a more environmentally friendly and low-impact solution to conventional chemistry, but they often require additional engineering for industrial settings, an endeavor that is challenging and laborious. To address this issue, the power of machine learning can be harnessed to produce predictive models that facilitate *in* silico study and engineering of novel enzymatic properties. However, the conversion from the biological domain to the computational realm requires special attention to ensure the training of accurate and precise models. In this review, we examine the critical step of encoding protein information to numeric representations for use in machine learning. We selected the most important approaches for encoding the three distinct biological protein representations — primary sequence, 3D structure, and dynamics — to explore their requirements for employment and inherent biases. Combined representations of proteins and substrates are also introduced as emergent tools in biocatalysis. We propose the division of fixed representations, a collection of rule-based encoding strategies, and learned representations extracted from the latent spaces of large neural networks. To select the most suitable protein representation, we propose two main factors governing this choice. The first one is the model setup, being influenced by the size of the training dataset and the choice of architecture. The second factor is the model objectives, concerning the assayed property, the difference between wild-type models and mutant predictors, and requirements for explainability. This review is aimed at serving as a source of information and guidance for properly representing enzymes in future machine learning models for biocatalysis.

Keywords

Machine Learning; Biocatalysis; Protein Representations; Enzyme Engineering; Representation Learning; Protein Dynamics, Predictive Models

1. Introduction

In the current time of climate change and increasing resource depletion, enzyme technology has emerged as a more environmentally friendly and potentially low-impact approach to industrial processes traditionally mediated by conventional chemistry (Buller et al., 2023; Hauer, 2020; Radley et al., 2023; Reetz et al., 2024; Sheldon and Woodley, 2018; Wu et al., 2021). Instead of complicated pathways with a plethora of reagents, extreme conditions, and protection groups, enzymes offer a renewable alternative with high selectivity and tunability (Sheldon and Woodley, 2018; Woodley, 2022; Wu et al., 2021). Early examples consist of enzyme-based detergents (Kirk et al., 2002) and the employment of nitrile hydratases to synthesize acrylamide (Yamada and Kobayashi, 1996). Recent advances in bioinformatics strategies have enabled the discovery of enzymes with specialized activity (Buller et al., 2023; Hon et al., 2020; Oberg et al., 2023), as well as the engineering of enzymes towards enhanced activity, substrate specificity, enantioselectivity, and thermostability (Galanie et al., 2020; Qu et al., 2020; Renata et al., 2015). Especially the directed evolution (DE) approach of mimicking Darwinian evolution, which was co-awarded with a Nobel Prize to Frances Arnold (Arnold, 2018, 1998, 1996) has seen significant use for enzyme engineering (Bornscheuer and Pohl, 2001; Cherry et al., 1999; Cherry and Fidantsef, 2003; Giver et al., 1998; Stimple et al., 2020; Turner, 2009; Zhao and Arnold, 1999). Enzymatic biocatalysis has had a profound impact in areas such as pharmaceutical drug discovery (Devine et al., 2018; Savile et al., 2010), the cosmetic industry (Heath et al., 2022; Khan and Rathod, 2015), and waste degradation (Bilal et al., 2019; Mohanan et al., 2020), and multiple enzymatic processes have even been developed sequentially to create biocatalytic cascades (France et al., 2017; Gandomkar et al., 2019; Huffman et al., 2019; Nazor et al., 2021; Santacoloma et al., 2011; Sperl and Sieber, 2018).

The growing use of enzymes has, nonetheless, revealed several challenges when utilizing them for industrial catalysis purposes because they did not evolve to perform optimally in industrial bioreactors where high stability, selectivity, and activity are important to maximize product yields. Despite improvements in protein engineering, enhancing multiple enzyme properties such as activity and stability simultaneously is still a difficult endeavor (Acevedo-Rocha et al., 2018; Calzadiaz-Ramirez et al., 2020; Stimple et al., 2020; Tokuriki et al., 2012), as well as the prediction and control of substrate specificity and regioselectivity — crucial properties for industrial purposes — are often challenging (Harding-Larsen et al., 2023; M. Yang et al., 2018). In this context, machine learning (ML) algorithms have emerged as powerful tools, capable of modeling complex relationships within protein and enzyme datasets. In biocatalysis, ML has facilitated the study and engineering of proteins and led to novel insights for improving enzymatic processes (Kouba et al., 2023; Markus et al., 2023; Mazurenko et al., 2020; Yang et al., 2019). Notable examples include activity and substrate specificity predictors (Robinson et al., 2020), deep learning (DL) models for the estimation of metabolic enzyme activities (Li et al., 2022) and for functional predictions of enzymes (Gligorijević et al., 2021), models for protein solubility predictions (Yang et al., 2016; Y. Yang et al., 2021), and numerous approaches for predicting protein stability changes upon mutagenesis (Blaabjerg et al., 2023; Folkman et al., 2016; Iqbal et al., 2022; Li et al., 2020; Teng et al., 2010). ML has also enabled a more efficient multiparametric optimization

Abbreviations: BLOSUM, BLOck SUbstitution Matrix; CNN, convolutional neural network; DL, deep learning; EC,enzyme commission; ELBO, evidence lower bound; GFP, green fluorescent protein; GNN, graph neural network; KNN, k-nearest neighbors; MD, molecular dynamics; MLDE, machine learning-assisted directed evolution; MSM, Markov state models; OHE, one-hot encoding; PLM, protein language model; QM/MM, quantum mechanics/molecular mechanics; VAE, variational autoencoder, XAI, explainable AI strategy (Kunka et al., 2023; Ma et al., 2021), facilitated *de novo* enzyme design (Yeh et al., 2023), and prediction of non-additive epistatic effects (Cadet et al., 2018, 2022; Li et al., 2021). Finally, ML has been combined with DE in the aptly termed "machine learning-assisted" directed evolution (MLDE), where it has significantly improved the exploration of the sequence-function landscape in the search for enhanced variants (Bruce J. Wittmann et al., 2021; Wu et al., 2019; Xu et al., 2020; Yang et al., 2024, 2019).

Traditionally, the focus within ML research has often been to refine the algorithms, whereas data representation is treated as a secondary concern. This viewpoint posits that given sufficient data and computational resources, ML models should inherently discern and leverage the most salient features relevant to the task at hand. However, this view overlooks the challenge of producing such large protein datasets of high quality (*i.e.*, reproducibility) and neglects the critical role of data representation in enhancing or limiting a model's ability to learn (Bengio et al., 2013; Iuchi et al., 2021). Our work addresses the topic of protein representations as a critical step for uniting biology and data science. In biology, a protein is commonly represented by its primary or tertiary structure through categorical or symbolic information, while ML traditionally requires numeric inputs in the forms of vectors, matrices, and tensors. This poses an exciting task of representing proteins in a manner that is both informative for ML models and reflective of the underlying biological properties.

Interestingly, the concept of inductive biases introduces a nuanced understanding of how ML models approach learning tasks. Inductive biases refer to the assumptions made by a model about the patterns it expects to find in the data before any data is indeed observed. They guide the learning algorithm towards certain solutions over others, effectively shaping the hypothesis space that the model explores (Baxter, 2000). Selecting the right inductive biases — through the strategic representation of data — can significantly facilitate the learning process, enabling models to learn more efficiently and effectively from fewer examples (Baxter, 2000).

In the context of biocatalysis, these inductive biases arise either manually or by representation learning, and the choices made during the encoding process strongly affect the information captured in the representations. In this review, we investigate the methodologies for protein representation utilizing the protein sequence, structure, or dynamics. We also analyze the assumptions of the inductive biases that are captured in the different representation techniques. We conclude with a discussion about different factors influencing the choice of protein representation.

2. Sequence Representations

A simple description of a protein is the one-dimensional sequence representation of the molecular structure using an alphabet of 20 amino acids. This leads to an alphanumeric expression of the biomolecular components to easily differentiate between proteins. While simple, the string of single-letter residue codes contains a vast amount of information, from the physicochemical properties of every amino acid to the evolutionary trace of the protein. Sequences are even intrinsically linked to 3D structures and functional properties, making them a rich source of information critical for protein design. However, the development of ML models for predicting protein functions requires precise feature extraction from those sequences. A spectrum of methodologies to identify optimal features are available, ranging from simple to complex ones. This section outlines the evolution of feature extraction techniques, emphasizing the transition from elementary assumptions to sophisticated models.

Finally, we will treat a mixed representation where structural insights are used to influence the sequence representation.

2.1 Fixed Sequence Representations

The methods for capturing biological information stored in the sequence representation are varied, often focusing on different elements of this information. One category of methods is the so-called "fixed" representations, a collection of rule-based approaches to convert between the protein sequence and numerical vectors by incorporating specific parts of the amino acid characteristics (Figure 1) (Markus et al., 2023). The simplest of all is the one-hot encoding (OHE) technique, a prevalent method in ML for transforming categorical data into a binary format. Here, each residue is represented as a vector $v_i = (0,0,...,1,...,0)$ with '1' placed at the i^{th} index corresponding to its lettering, creating a binary 20 × N matrix with a single non-zero entry in each column, where N is the length of the protein sequence. Although OHE offers no protein information aside from the amino acid identities, it is used extensively as a fast and effective method for converting biological information into numerical vectors (Elabd et al., 2020; Goldman et al., 2022; Greenhalgh et al., 2021; Hsu et al., 2022; Michael et al., 2023; Raimondi et al., 2019; Bruce J. Wittmann et al., 2021; M. Yang et al., 2018). However, the sparse and high-dimensional nature of OHE can lead to computational inefficiencies, particularly in models dealing with long protein sequences. Moreover, many ML algorithms require the input of a fixed size throughout their training and inference, necessitating an additional data pre-treatment step in OHE, *e.g.*, trimming long sequences or extending short ones with zeros.

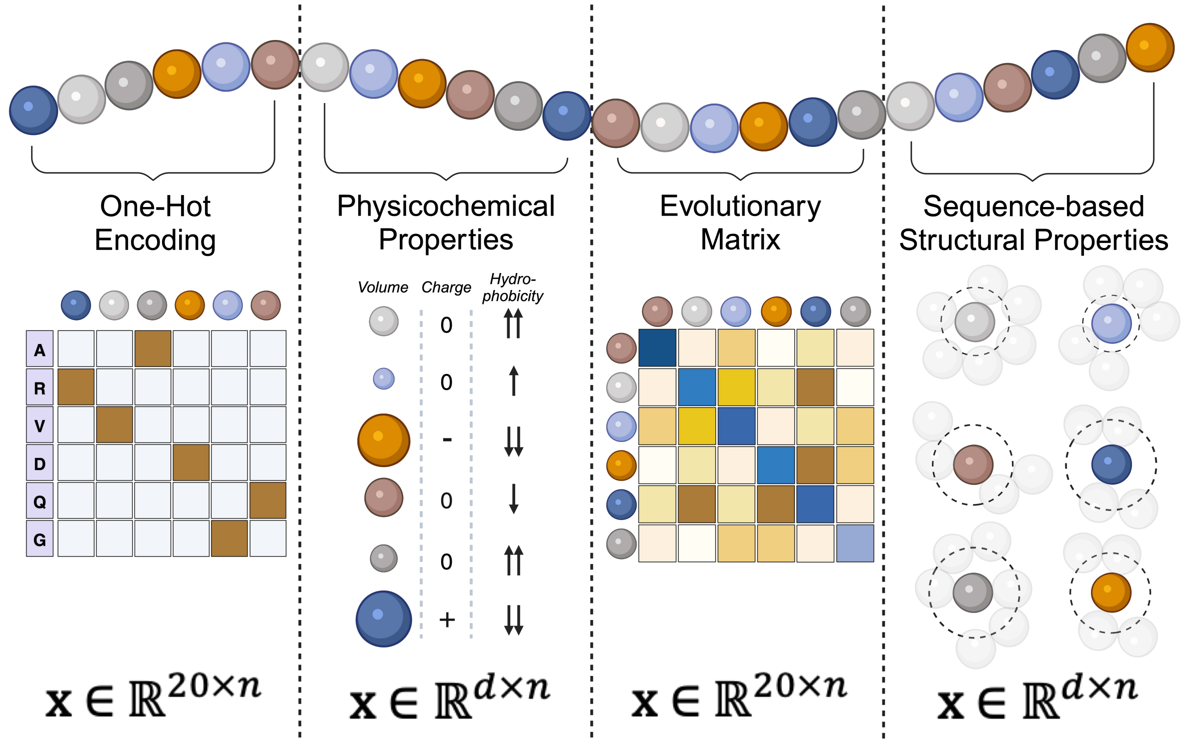

One-Hot Encoding
Physicochemical Properties
Evolutionary Matrix
Sequence-based Structural Properties
A
R
V
D
Q
G
Volume
Charge
Hydro-phobicity
0
0
-
0
0
+
$\mathbf{x} \in \mathbb{R}^{20\times n}$
$\mathbf{x} \in \mathbb{R}^{d\times n}$
$\mathbf{x} \in \mathbb{R}^{20\times n}$
$\mathbf{x} \in \mathbb{R}^{d\times n}$

Fig. 1. Fixed representations for encoding the protein sequence. OHE (left) is the simplest method and only uses the amino acid identity. Physicochemical properties (middle left) instead capture the nature of the amino acids by explicitly using their properties as features. Matrices such as the BLOSUM encoding introduce evolutionary information to the protein representation (middle-right). Lastly, the sequence can also be used to calculate structural properties such as SASA (right).

The simple nature and lack of inherent bias prevent OHE from capturing any relationships between amino acids before the training. Property-based encoding strategies emerge as a potential solution to instruct ML algorithms about the physicochemical nature of the sequences, either global protein descriptors or those at the residue level. The former captures the behavior of the entire protein chain through properties such as solubility or radius of gyration, while the latter instead enables the encoding of each amino acid using a set of properties such as charge, hydrophobicity, volume, or pK_a, imposing representation biases towards certain residue attributes and allowing the model to discern the similarities and differences between two residues. Various sets of physicochemical residue descriptors exist, such as the large database of amino acid indices, and AAindex (Kawashima and Kanehisa, 2000), containing over 500 matrices for encoding sequence information. Such a set of indices for charge, polarity, hydrophobicity, average accessible surface area, and side chain volume was used to model and predict the donor specificity of fold A glycosyltransferases by Taujale et al. (Taujale et al., 2020). Another example is the recent pre-print by Xu et al., where the authors employ physicochemical properties such as volume, hydrophobicity, and π-π interactions to model and improve enantioselectivity of carboxylesterase *Ac*Est1 from *Acinetobacter sp. JNU9335* (Xu et al., 2024).

Instead of manually choosing between the many similar indices, the inherent patterns of the physicochemical properties can be extracted through their principle components, such as the Vectors of Hydrophobic, Steric, and Electronic properties (VSHE) (Mei et al., 2005), z-scales (Hellberg et al., 1987; Jonsson et al., 1989; Sandberg et al., 1998; Wold et al., 2011), the DL-based amino acid parameter representations by Meiler et al. (Meiler et al., 2001), or the five factors described by Atchley et al. (Atchley et al., 2005). Using these principal components enables the incorporation of a wide range of different residue properties without drastically increasing the dimensionality of the vector representation due to the principal components containing information from multiple physicochemical properties. An example is Factor III by Atchley et al. which encompasses bulkiness, residue volume, average volume of a buried residue, side chain volume, and molecular weight (Atchley et al., 2005). Several ML models have employed these dimension-reduced physicochemical representations for different enzymes, including the thiolase activity and substrate specificity predictors (Robinson et al., 2020), the Sortase A mutagenesis model for ML-guided directed evolution (Saito et al., 2021), and DeepTM, a DL-based model for predicting the melting temperatures of proteins such as PET plastic-degrading enzymes (M. Li et al., 2023). Nevertheless, a potential issue with this approach is the “black box”-like nature, complicating the process of interpreting the results and discerning the actual residue property contributions when examining model feature importance.

Aside from introducing residue information and imposing an inherent bias to the protein representation through physicochemical properties, the encoding method can be based on the evolutionary information contained in the sequence. These biases force the model to learn evolutionary important patterns. One such technique, the BLOck SUbstitution Matrix (BLOSUM) encoding, is generated from alignments of protein sequences and focuses on evolutionary changes and conservation (Henikoff and Henikoff, 1992; Mount, 2008). Based on the frequency of amino acid substitutions in these alignments, each entry in a BLOSUM matrix represents the likelihood of substitution between amino acids, calculated based on observed substitutions in protein families. In BLOSUM encoding, each amino acid is replaced by a vector derived from the corresponding row in the BLOSUM matrix, $v_i = (x_A, x_G, \ldots, x_Y)$ where x_A is the likelihood score that the i^{th} residue is substituted with alanine, thus enabling the representation to capture the evolutionary history and functional similarities between amino acids. We employed this sequence representation in our model for predicting glycosyltransferase activity specificity (GASP), which allowed the model to use the evolutionary information to discern the wide array of different glycosyltransferases (Harding-Larsen et al., 2023). The evolutionary information can also be captured using a Position Specific Scoring Matrix (PSSM), a method that uses a Multiple Sequence Alignment (MSA) of a set of proteins to quantify the likelihood p_{ij} that an amino acid at a specific position j mutates into the i^{th} residue. These matrices can be constructed using a sequence similarity program such as PSI-BLAST (Altschul et al., 1997).

Finally, a fourth approach to extracting biological information from the protein sequences is to exploit the relationship between the primary sequence and the 3D structure. Secondary structure elements have long been possible to estimate purely through primary sequence (Y. Yang et al., 2018), and also structural properties such as Solvent Accessible Surface Area (SASA) (Lee and Richards, 1971) and the Half Sphere Exposure (HSE) (Hamelryck, 2005) can be predicted from sequence alone (Cheng et al., 2005; Fraczkiewicz and Braun, 1998; Heffernan et al., 2017; Song et al., 2008). Sequence-based structural properties have been used in tandem with metabolic network properties, reaction thermodynamics, and assay conditions to predict WT metabolic enzyme turnover numbers (Heckmann et al., 2020, 2018), exhibiting significant importance compared to the other model features. Sequence-based structural properties were also applied in the previously mentioned DeepTM (M. Li et al., 2023) algorithm, again as part of a larger feature set.

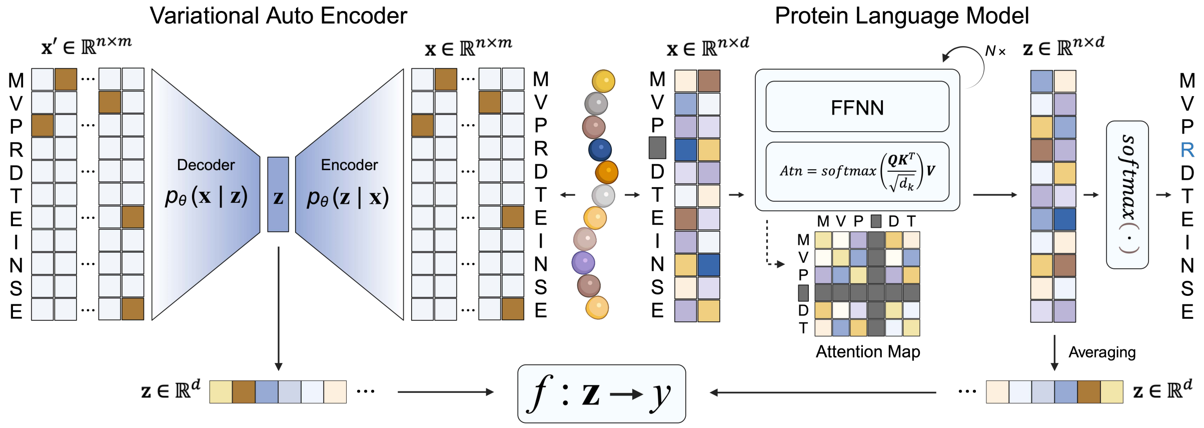
Lastly, it should be noted that the development of AlphaFold2 (Jumper et al., 2021) and similar sequence-to-structure tools (Ahdritz et al., 2022; Baek et al., 2021; Lin et al., 2023) has blurred the boundary between sequence- and structure-based protein representations, as these tools are capable of predicting the entire 3D structure using only the sequence. This ambiguity is necessary to consider, *e.g.*, for fair comparison of sequence-only encoding techniques and algorithms.

2.2 Representation learning

An alternative to manually extracting features from sequence information is to learn features or representations of sequences through machine learning from data (Iuchi et al., 2021; Sinai and Kelsic, 2020). The key idea is to learn general representations through a machine model by training on large data sets of unlabeled protein sequences. The obtained representations of the pre-trained embedding model are then used to train a task-specific (surrogate) model, requiring less labeled data. The following sections will describe two common approaches for learning sequence embeddings (Figure 2).

2.2.1 Variational Autoencoders

Variational Autoencoders (VAEs), introduced by Kingma and Welling in 2013 (Kingma and Welling, 2013), offer a framework for training deep latent variable models that learn meaningful representations by optimizing a lower bound on the likelihood of the data, essentially trying to maximize the probability of observing the training data under the model. This process involves a balance between accurately reconstructing data and enforcing a structured latent space, making it possible for VAEs to generate new data samples that resemble the original inputs. This allows VAEs to capture essential features of the data efficiently. The utility of VAEs is particularly evident in handling high-dimensional and sparse data, such as large sets of one-hot encoded (OHE) protein sequences, enabling the extraction of compact and meaningful representations (Detlefsen et al., 2022).



Variational Auto Encoder
Protein Language Model
$\mathbf{x}' \in \mathbb{R}^{n \times m}$
$\mathbf{x} \in \mathbb{R}^{n \times m}$
$\mathbf{x} \in \mathbb{R}^{n \times d}$
$\mathbf{z} \in \mathbb{R}^{n \times d}$
M V P R D T E I N S E
Decoder
$p_\theta(\mathbf{x} \mid \mathbf{z})$
$\mathbf{z}$
Encoder
$p_\theta(\mathbf{z} \mid \mathbf{x})$
FFNN
$Atn = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$
$N\times$
M V P D T
Attention Map
Averaging
$softmax(\cdot)$
$\mathbf{z} \in \mathbb{R}^d$
$f : \mathbf{z} \rightarrow y$

Fig. 2. Two common approaches for learning sequence embedding. Variational Autoencoders (left) are latent variable models that utilize an encoder-decoder setup to learn a latent space embedding, **z**. Protein Language Models (right) are also used to generate sequence representations but instead employ an attention mechanism that dynamically weighs the relevance of different parts of a protein and a Feedforward Neural Network (FFNN). A protein encoding can be obtained by averaging over the neural embeddings. The resulting representations from both techniques can then be used for fine-tuning task-specific predictions.

The foundation of VAEs is centered around the transformation of input data (*e.g.* OHE sequences), x, into a latent distribution, z, through an encoder, $q_\theta(z|x)$. The latent distribution, typically Gaussian, is characterized by parameters (mean and variance) derived from the input by a neural network. The decoder of the VAE then attempts to reconstruct the input data from the latent variables, following the distribution $p_\phi(x|z)$. The objective of training a VAE is to maximize the evidence lower bound (ELBO) on the log-likelihood, which is expressed as:

$$\mathcal{L}(\boldsymbol{x}; \theta, \phi) = \mathbb{E}_{q_\theta(\mathbf{z}|\boldsymbol{x})}\left[\log p_\phi(x|\mathbf{z})\right] - D_{KL}(q_\theta(z|\boldsymbol{x})||p(z))$$

The first term in the ELBO represents the reconstruction loss, promoting similarity between the decoded samples and the original inputs, and the second term is the Kullback-Leibler (KL) divergence, serving as a regularization term ensuring that the latent space is well-regularized and continuous, enabling efficient data representation and interpolation (Tschannen et al., 2018; Vincent et al., 2008).

In the context of protein sequences, VAEs leverage the manifold hypothesis, which suggests that high-dimensional data can be effectively modeled on a low-dimensional, non-linear manifold (Vincent et al., 2008). VAEs achieve two critical objectives: (i) reducing the dimensionality and sparsity to mitigate the curse of high dimensionality (Bellman, 1966) and (ii) incorporating domain-specific knowledge through the model architecture and sequence preprocessing and sequence alignment (Detlefsen et al., 2022). Choices made when building the architecture and constructing the MSA not only facilitate more efficient learning but also enhance the model's ability to support transfer learning by introducing inductive biases that align with the tree topology of the evolutionary history underlying the protein family (Ding et al., 2019). For these among other reasons, latent variable models such as VAEs have seen widespread adoption for predicting the mutational effect on protein fitness and in MLDE. Notable examples are the mutational effect predictor EVE by Frazer et al. (Frazer et al., 2021) or applications in MLDE studies conducted by Wittmann et al. (Bruce J Wittmann et al., 2021; Bruce J. Wittmann et al., 2021). Giessel et al. utilized Variational Autoencoders to engineer therapeutic enzyme variants with improved stability and activity, showcasing the model's ability to generate novel ornithine transcarbamylase sequences with enhanced therapeutic potential, marking a significant advancement in the application of VAEs for therapeutic enzyme engineering (Giessel et al., 2022). Hawkins-Hooker et al. successfully employed Variational Autoencoders to generate novel, functional variants of the luxA bacterial luciferase, demonstrating VAEs' capacity to explore protein sequence space and manipulate biophysical properties such as solubility, thereby presenting a valuable complement to traditional protein engineering methods (Hawkins-Hooker et al., 2021). Kohout et al. leverage VAEs to design novel variants of haloalkane dehalogenases for biocatalysis, demonstrating the applicability to generate sequences with stability and activity comparable to wild types while addressing challenges in maintaining protein solubility (Kohout et al., 2023). Finally, Hsu et al. highlighted the versatility of VAEs by augmenting evolutionary density scores extracted from the DeepSequence VAE model (Riesselman et al., 2018) with the simplistic OHE (Hsu et al., 2022). The augmentation approach achieved high performance across 19 different datasets — even models trained on as few data points as 42.

2.2.2 Protein Language Models

Another common method for generating protein sequence representations is Protein Language Models (PLMs), which nowadays increasingly employ the Transformer architecture (Vaswani et al., 2017). The Transformer is an ML architecture originally popularized in the domain of natural language processing to learn general patterns of language by predicting the missing words intentionally removed from sentences by their context. PLMs are trained on large protein sequence databases containing sequences sampled across different organisms. The training objective of PLMs is to reconstruct the sequence of a protein after it has been partially corrupted through the masked language modeling objective (Devlin et al., 2018). Similar to VAEs, PLMs can be used to extract latent representations of protein sequences, by forward passing sequences through the trained model and averaging the final layer output over the sequence length (Rao et al., 2020). A major difference between PLMs and VAEs is the attention mechanism at the core of PLMs, which allows the network to build up complex representations that incorporate context from across sequences (Rives et al., 2021):

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}}\right)\boldsymbol{V}$$

The attention mechanism used in Protein Language Models (PLMs) dynamically weighs the relevance of different parts of a protein sequence by calculating a weighted sum of values ($\boldsymbol{V}$). The weights are determined by the compatibility of queries ($\boldsymbol{Q}$) and keys ($\boldsymbol{K}$), which is scaled by a constant, the square root of the dimension of the keys (d_k) in the original transformer implementation (Vaswani et al., 2017), and normalized through a softmax function. Analysis of PLM representations has revealed that PLMs intrinsically learn biologically relevant features. For instance, their attention maps have been shown to bear a close resemblance to contact maps in proteins, indicating their capability to capture essential biological insights (Rives et al., 2021). PLM representations have demonstrated great flexibility in domain-specific tasks, such as function prediction, protein localization, and mutational effect prediction (Brandes et al., 2022; Elnaggar et al., 2021; Ferruz et al., 2022; Goldman et al., 2022; Rives et al., 2021; Thumuluri et al., 2022). PLMs offer a robust way to generate highly effective representations for domain-specific applications, making them a popular choice when creating ML models for biocatalysis. Examples of PLMs for biocatalysis include the study by Yu et al. utilizing contrastive learning for the precise annotation of enzyme functions by Enzyme Commission (EC) numbers, outperforming conventional tools in accuracy and capability to annotate underexplored and mislabeled enzymes (Yu et al., 2023). Hoffbauer and Strodel introduce TransMEP, a tool employing transfer learning from protein language models to accurately predict the effects of mutations on proteins, demonstrating the efficacy of leveraging pre-trained models like ESM-2 (Lin et al., 2023) for mutation effect prediction in protein engineering (Hoffbauer and Strodel, 2024). The pre-trained model of ESM-1b (Rives et al., 2021) has also seen extensive use in biocatalysis, either directly employed as protein representations for supervised tasks (Goldman et al., 2022; Hou et al., 2023; Bruce J. Wittmann et al., 2021; Xu et al., 2022), or in the form of a fine-tuned task-specific encodings (Kroll et al., 2023a, 2023b).

2.2.3 Comparing VAEs with PLMs

Both PLM and VAE representations frequently rank as the state of the art in task-specific application benchmarks, such as mutational effect prediction (Livesey and Marsh, 2023) or MLDE studies (Bruce J. Wittmann et al., 2021). When comparing VAEs to PLMs for applications in protein engineering, some general rules can be drawn. There are some indications that VAEs show greater performance for task-specific applications (Bruce J. Wittmann et al., 2021). VAEs are also smaller than PLMs, which makes them faster at inference and easier to run without large computational resources. Furthermore, VAEs are superior during sampling, due to their ability to easily sample from the latent distribution by passing latent variables through the decoder. VAEs can be highly customized, for example, allowing the creation of latent variables with fewer dimensions to facilitate data visualization or fine-tuning (Detlefsen et al., 2022). On the other hand, VAEs have to be trained individually for each protein family, whereas PLMs can be used across all protein families without further training, even generalizing beyond naturally observed proteins (Verkuil et al., 2022). Interestingly, nowadays ML developers are exploring the possibility of combining PLMs and VAEs (Sevgen et al., 2023).

2.3 Structure-Informed Sequence Representations

Some methods incorporate structural information when producing a sequence representation. Here, the protein structure is employed as a selection filter for the identification of important residues, delimiting the sequence encoding to a curated list of amino acids and circumventing the issue of information dilution where redundant features dominate the informative ones. For biocatalysis, these structure-informed sequence representations ensure that the focus is directed towards important parts of the enzyme, such as the active site, remote binding sites, or other areas believed to be important for the enzymatic property to be modeled (*e.g.*, dimer interfaces).

In structure-informed sequence representations, a 3D structure is combined with an MSA to identify and encode specific residues in every protein of interest. Generally, two different approaches exist for this identification: manual selection and spherical extraction. The former method entails examining the template structure and choosing the residues important for the area in focus such as the residues lining the active site as described by Röttig et al. in their Active Site Classification (ASC) strategy to model the protein families of kinases, nucleotidyl cyclases, trypsins, malate/lactate dehydrogenases, and decarboxylating dehydrogenases (Röttig et al., 2010). The list of manually curated residues is then mapped onto every protein in the MSA through the aligned positions of the identified residues. In the spherical extraction method, the list of important residues is instead acquired automatically by constructing a spherical boundary around the area in focus, *e.g.*, the catalytic residues, and then extracting all amino acids encompassed by this boundary using protein structure analysis programs such as MDTraj (McGibbon et al., 2015) or BioPython (Cock et al., 2009). This automated selection approach was employed by Robinson et al. to model and predict the substrate specificity of OleA thiolases; aligning all 73 sequences to the OleA thiolase from *Xanthomonas campestris* (Goblirsch et al., 2016) and extracting the active site residues from a crystal structure of the before-mentioned protein using a 12 Å sphere centered around the C_α of the active site cysteine (Robinson et al., 2020). Another example is Goldman et al. who examined the activity and substrate specificity of multiple protein families including glycosyltransferases and halogenases using spheres ranging from 3 to 30 Å (Goldman et al., 2022).

Both selection strategies have their merits and deficiencies: while manual selection ensures a significant degree of control over the choice of residues, it ultimately requires expert curation and is highly protein-specific. The spherical extraction technique sacrifices some of this control to alleviate these issues by only needing the centroid and radius to be defined, making the process faster than the manual selection.

Importantly, the structure-informed approach currently requires an MSA to map the identified residues to the entire set of proteins, which might cause problems for poor alignments with many gaps that offer minimal protein information. Furthermore, while the strategy can be used to bias the representation to focus on specific areas of the protein, discarding a significant portion of the sequence is also an inherent limitation of the method. If a distant part of the protein is important for a property, *e.g.*, due to allostery influencing protein activity (Calvó-Tusell et al., 2022a), this information will be lost when only focusing on a specific site. Furthermore, if an ML model targets global properties such as protein fitness scores (Fox, 2005; Michael et al., 2023; Bruce J. Wittmann et al., 2021; Wu et al., 2019) or melting temperature (M. Li et al., 2023), it is unlikely to benefit from focusing the protein representation on a particular part of the protein.

3. Structure Representations

The biological structure representation contains information about the relative 3D positions and chemical identities of every atom and bond of the protein, $\mathbf{x} = \mathbb{R}^{3\times N}$, with N being the length of the sequence. Increasing the information complexity from a 1D amino acid sequence to a 3D structure thus introduces additional challenges for the encoding, especially when working with simpler ML architectures requiring an abstraction of the protein structure into a one-dimensional representation vector. Encoding the protein structure can either be done by extracting fixed features directly from the structure or by converting the highly detailed 3D protein into a simpler representation for producing learned representations. Alternatively, it can be done by utilizing a novel structure alphabet.

3.1 Fixed Features Extracted from the Protein Structure

Similar to describing the sequence through a set of fixed properties, fixed structure representations can be constructed by quantifying different aspects of the protein structure. While the use of these structural features has been limited in ML for biocatalysis, several approaches exist for extracting features from the 3D structure of a protein. Many enzymes utilize a binding pocket to tailor the catalytic environment, which can be converted to numerical descriptors through tools such as Fpocket (Le Guilloux et al., 2009), a program for detecting and describing ligand-binding pockets. Features from Fpocket have seen use in allosteric site prediction (Xiao et al., 2022). Accurate van der Waals surface area descriptors, moments of inertia, electrostatics, and thermodynamic values can be calculated through programs such as ProtDCal (Ruiz-Blanco et al., 2015), and those features have seen use in models predicting the substrate specificity of nitrilases (Mou et al., 2021) or estimating the kinetic parameters of glycoside hydrolases (Carlin et al., 2016).

3.2 Simplification of the 3D Protein Structure for Representation Learning

Instead of distilling the structural information into a set of descriptors, the structural data can be converted into simplified representations that retain more information than fixed structure features. This can be done with a cubic grid (voxel), protein graph representations, or protein surface representations. These methods can then be employed in DL architectures to construct learned protein representations (Figure 3) (Isert et al., 2023).

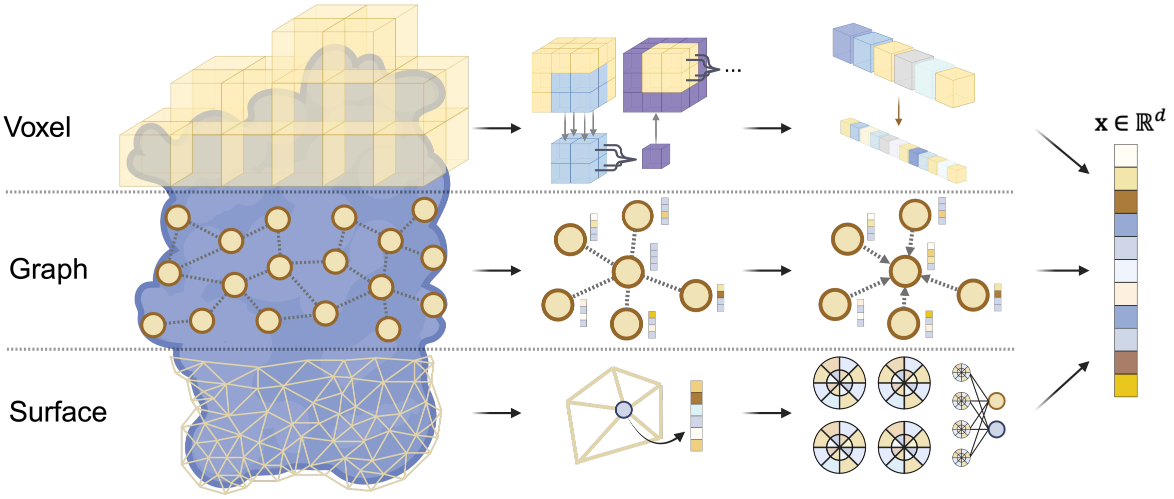


Fig. 3. Three common structure representations for DL architectures and their process towards a learned 1D vector representation x ∈ $\mathbb{R}^d$. Top: the protein structure is approximated using a 3D voxel grid representation. This grid is processed using a 3D CNN, where voxels are sequentially convoluted to reach the desired dimensions. Middle: the protein graph is a non-linear representation of the structure using nodes and edges. In the GNN, the properties of each node are passed through the edges to update the node information. Bottom: Triangulation creates a protein surface representation with each vertex containing physicochemical information. The mesh is usually deformed to a polar coordinate system and processed using a convolutional network.

3.2.1 Grid Representations

The continuous protein structure can be converted to a discrete representation by dividing the molecular space into individual grid sections. Volumetric cubes — so-called voxels — represent 3D data by an assembly of course-grained cubes, drastically reducing the dimensions of the encoding (Isert et al., 2023). This can either be implemented by dividing the structure into smaller “microenvironments” and then encoding each of these microenvironments individually (Paik et al., 2023; Shroff et al., 2020; Torng and Altman, 2017), or by encoding the entire protein into a single arrangement of cubes based on a regular 3D grid (Amidi et al., 2018).

MutCompute is a tool that utilizes the former strategy of microenvironments (Paik et al., 2023; Shroff et al., 2020). For every residue in each protein, a cubic 20Å microenvironment is represented by 1Å voxel cubes containing information about atom labels, partial charges, and solvent accessibility of each atom within the voxel cube. The microenvironment representation is then processed by a 3D convolutional neural network (CNN) and later a fully connected neural network (FCNN). This allows the authors to evaluate the chemical and steric suitability of each of the 20 natural amino acids. This can be used as the basis for mutagenesis, such as highlighted by the study achieving an improved thermostability of the *Bacillus stearothermphilus* DNA polymerase (Paik et al., 2023). Novel work has expanded upon the model of MutCompute, introducing information about phosphorus and grouped halogens and thereby facilitating the training on heterogeneous microenvironments (d'Oelsnitz et al., 2024). The new model, MutComputeX, was employed for the engineering of activity-enriched variants of methyltransferase.

Instead of dividing the protein structure into smaller segments, Amidi et al. employed the entire protein structure in their encoding strategy (Amidi et al., 2018). The protein backbone is converted into a binary voxel grid with a predefined resolution and processed by a 3D CNN. The model was trained to predict EC numbers, achieving an accuracy of 78.4%. The authors furthermore highlighted the versatility of this approach, as the model's binary voxel representation can be replaced by physicochemical properties such as hydrophobicity and isoelectric points. This allows future models to include inductive biases tailor-made for a specific task. It should be noted that while the voxel representation can directly capture the 3D nature of proteins, it is not without limitations. For example, it is sensitive to rotations and translations of a 3D structure in space and does not directly capture information about chemical bonds.

3.2.2 Protein Graphs

An alternative approach to grid representations is to collapse the 3D protein structure to a graph representation where the structural information of the protein is encoded as elements and connections, designated as "vertices"/"nodes" and "edges", respectively (Fasoulis et al., 2021). Different detail levels can be employed when creating protein graphs, *e.g.*, for atomistic resolution, features of each node consist of atom type and charge, while the edges represent the molecular bonds (Fasoulis et al., 2021). A more coarse-grained approach is the residue-level description where the nodes represent entire amino acids and the edges specify both the covalent and non-covalent interactions between the residues. For residue-level protein graphs, the node features can include physicochemical properties such as polarity and hydrophobicity (Fasoulis et al., 2021), or more advanced residue encodings such as evolutionary information or secondary structure (M. Li et al., 2023). Importantly, a graph is a non-linear data structure. The node connections can be represented using adjacency matrices where the i^{th} element in the j^{th} row describes the edge between the i^{th} and the j^{th} node, with the ordering of the nodes being arbitrary. The protein contact map is an example of an adjacency matrix.

Due to the non-linearity of graph representations, it is often infeasible to combine them with a classical ML architecture, such as logistic regression or tree-based models. This processing issue is solved by employing Graph Neural Networks (GNNs), a network architecture that directly implements the graph representation in model construction. In contrast to traditional neural networks where the information is passed through a series of hidden layers, GNNs utilize the edges as channels for information transfer between the individual nodes. This ensures that only information originating from neighboring nodes within a pre-defined proximity is used to update each node (Zhou et al., 2020).

An exciting example of a GNN-based enzyme predictor is DeepFRI, a model leveraging both sequence and structure representations to model Gene Ontology (GO) terms and EC numbers (Gligorijević et al., 2021). Here, the sequence embeddings of a pre-trained PLM are used as residue nodes while a protein contact map is utilized as graph edges. A recent pre-print also proposed to combine the ESM2 sequence embeddings with graph-based structure embeddings for downstream tasks, such as predicting EC numbers, introducing the Protein Structure Transformer (PST) architecture, outperforming previous state-of-the-art models (Chen et al., 2024).

It should be noted that while building GNNs requires a significant amount of data, pre-trained structure embeddings can be utilized as protein encodings, drawing a parallel to the pre-trained sequence embeddings. This was highlighted by the authors of PST, exhibiting high performance using pre-trained protein embeddings extracted from the model (Chen et al., 2024). Another example is the Masked Inverse Folding (MIF) model (K. K. Yang et al., 2022), a GNN trained on the sequences and structures of 19.000 proteins in the CATH4.2 dataset (Dawson et al., 2019, 2017) to reconstruct a corrupted protein sequence using backbone information. The MIF embeddings have seen use as a representation of the protein structure (Hou et al., 2023), where the power of GNNs is harnessed to process structural information without requiring either a large dataset or computationally costly model training.

3.2.3 Surface Encodings

Finally, the protein can be modeled using a mesh-based variant of the molecular surface, a continuous sheet describing the accessibility trace of the molecule using a probe of a given radius (Richards, 1977). An example is the surface used for calculating the previously mentioned SASA, where the contact surface is the parts of the atomic van der Waals spheres in contact with the probe. The continuous surface can be discretized using triangulation, where the curvature is converted into a protein polygon mesh using tools such as MSMS (Sanner et al., 1996). These surface meshes are often encoded with the physicochemical information of the residues or atoms, allowing them to function as protein representations in ML models.

Notable examples of models harnessing surface representations include molecular surface interaction fingerprinting MaSIF (Gainza et al., 2019). In this example, the surface is here segmented by assigning radial patches to every vertex in the protein mesh and generating an overlapping collection of surface vertices. Geometric features and chemical properties are calculated for each vertex within the patches, and the mesh is mapped to a polar coordinate system. This representation is passed through a convolutional architecture that produces learned fingerprint descriptors. The authors utilized these fingerprints to classify ligand-binding pockets, predict protein–protein interaction sites, and estimate the structural configurations of protein–protein complexes. While not inherently targeting biocatalysis, Gainza et al. consequentially highlight the advantage of surface presentation learning for understanding protein interactions.

In SURFMAP, the reduced surface generated by the MSMS tool (Sanner et al., 1996) is employed to generate a set of particles, each 3Å away from the protein surface (Schweke et al., 2022). After mapping the particles with a feature such as hydrophobicity or stickiness related to the closest residue, their spherical coordinates are projected onto a 2D map using the Sanson-Flamsteed 2D projection. The authors employed this simplified representation to construct a hierarchical clustering model of superoxidase dismutases. This allowed them to distinguish between enzymes with different oligomerization states and metal ion binding preferences. Lastly, the HoloProt model combined structure- and surface-based graphs in multi-scale graph representation to predict enzyme classifications and protein-ligand binding affinities (Somnath et al., 2021).

3.3 Alternative Structure Representations

While we have generally categorized protein structure representation as either fixed descriptors or geometrical simplifications for learned representations, some approaches fall outside of this division. Recently, a novel technique for representing the protein structure using a string of letters has emerged in Foldseek (van Kempen et al., 2023). Originally designed as a tool to efficiently align a query structure against large databases, Kempen et al. developed an intriguing structure encoding. An artificial alphabet — denoted 3Di — describing the tertiary interactions of the protein is generated using a VAE. Each protein is encoded using this 3Di alphabet, and the resulting sequences are parsed through the prefilter modules of MMseqs2 (Steinegger and Söding, 2017), a protein sequence searching tool, to use in alignment queries. The Foldseek structure-to-sequence approach facilitates the use of traditional sequence representation architecture to process structural information (Heinzinger et al., 2023; Sledzieski et al., 2023; Su et al., 2023; Waksman et al., 2024). While no enzyme models have been trained using these 3Di representations as of the writing of this review, we envision this to be an exciting area for future utilization of structural information.

4. Dynamics Representation

At the heart of enzymology lies the dynamic nature of enzymes (Henzler-Wildman and Kern, 2007), a realm where static structural protein models meet their limits (Lane, 2023). Enzyme dynamics are becoming a key ingredient to understanding and engineering enzyme function, yet the incorporation of dynamic representations in ML remains in its infancy. Enzyme dynamics is observed as the collective movements at time scales of femtosecond bond vibrations, nanosecond side-chain fluctuations, and millisecond domain motions. Together, these motions are termed conformational dynamics and are critical for understanding enzymes (Agarwal et al., 2020; Corbella et al., 2023; Henzler-Wildman and Kern, 2007).

4.1 Dynamics as a Tool to Understand, Predict, and Engineer Enzymatic Activity

Dynamics are important and offer explanations to why distal mutations accumulate during directed evolution campaigns (Osuna, 2021), why conformational changes such as lid opening/closing rates can be rate-limiting (Wolf-Watz et al., 2004), and how conformational heterogeneity is linked with evolvability of enzyme function (Campbell et al., 2016, 2018; Corbella et al., 2023; Kim and Porter, 2021). Enzyme dynamics form a foundation on which enzymes have been studied rationally, ranging from the canonical β-lactamase (Galdadas et al., 2021), to halogenases (Ainsley et al., 2018), transferases (Tian et al., 2024), lipases (Behera and Balasubramanian, 2023), luciferases(Schenkmayerova et al., 2021), dehalogenases (Vasina et al., 2022), and dehydrogenases (Acevedo-Rocha et al., 2021; Calzadiaz-Ramirez et al., 2020). Dynamics often explain the evolution of enzymes, as they seemingly evolve dynamic networks and freeze out unproductive motions to increase catalytic activity (Bunzel et al., 2021; Campbell et al., 2016).

Predictions of mutant effects on dynamics using statistical tools and algorithms are currently enabling the challenging task of conformationally driven enzyme design (Osuna, 2021). The approaches are, however, not limited to computational tools. Experimentally driven design of dynamics is also underway, enabled by advances in NMR, room-temperature and time-resolved X-ray crystallography, facilitating experimental studies of enzyme dynamics and elucidating its link to activity (Bhattacharya et al., 2022; Broom et al., 2020; Weinert et al., 2017).

What remains are ML/DL-driven end-to-end solutions for predicting changes in catalytic activity based on dynamic representations. This necessarily requires numerical representations that are well-suited for available architectures. The next frontier of computational biology is to predict the correlation between conformational dynamics and specific mutations, and their effect on activity, work which is well underway. This includes recent works on multi-state design, including simple dynamic representations to predict changes in activity, and ensemble-based enzyme design (Broom et al., 2020; Elia Venanzi et al., 2024; St-Jacques et al., 2023).

4.2 A Primer on Conformational Dynamics

Utilizing the temporal dimension of structural biology implies moving from a single structure parameterized computationally by Euclidean coordinates $\mathbf{x} \in \mathbb{R}^{3n}$ to a set of structures $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots\ldots, \mathbf{x}_n\}$. The temporal perspective ($\mathbb{R}^{3n}_{x,y,z} \times \mathbb{R}^{t}$) is challenging for biologists and computational scientists alike, as relevant collective movements must be extracted and correlated with enzymatic properties. It is a significant challenge for both communities to represent these movements efficiently. The task of dynamic representations is thus finding a map between the high-dimensional input using a collection of structures $\mathbf{X}$ to a lower-dimensional representation $f : \mathbf{X} \rightarrow \mathbb{R}^{m}$, without losing essential information.

Reflecting contemporary opinions (Vani et al., 2023), it is pertinent to clarify the dynamics of enzymes, which can be defined as a hierarchy of information (Figure 4). While the simplest protein dynamics examination is short-timescale sampling around one conformational state, for systems populated by multiple conformational states, *e.g.*, A, B, and C, conformational diversity is defined as all accessible conformations without any order {C, A, B}. Conformational ranking implies that the order of relative population is known {A, B, C}. Boltzmann diversity orders all conformational states with correct Boltzmann weights (relative populations). Lastly, conformational dynamics are all accessible conformational states with correct Boltzmann weights and inter-conversion timescales (arrows in Figure 4). Using these definitions, many approaches do not rigorously describe conformational dynamics, but only aspects on low rungs of the information hierarchy.

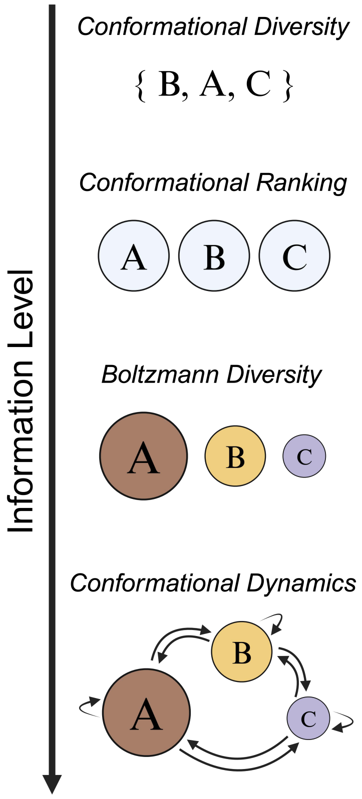


Fig. 4. The hierarchy of information for dynamics. Conformational Diversity is all accessible conformations without any order, while the order of the relative population is known in Conformational Ranking. Boltzmann Diversity orders all conformational states according to their Boltzmann weights. Lastly, Conformational Dynamics contains all accessible conformational states with correct Boltzmann weights and inter-conversion timescales (arrows).

4.3 Dimensionality Reduction of MD Simulations

Enzyme dynamics is typically studied computationally using long-duration molecular dynamics (MD) simulations *in silico*, based on Newtonian dynamics using small time steps to propagate a system forward a small unit in time (typically femtoseconds, 10^{-15} s). Often, this is carried out for millions of time steps resulting in a high-dimensional representation, and the challenge then lies in reducing dimensionality while conserving relevant dynamics information (Figure 5). These reductions are termed collective variables (Bhakat, 2022).

Collective variables were conventionally geometric measures between key catalytic residues and the ligand (Bhakat, 2022). These may represent the temporal fluctuation of distances, angles, or dihedral angles, thus summarising key interactions. The measures are selected based on domain knowledge of enzyme function and mechanism and have been successfully used to predict and engineer enzymes (A.Maria-Solano et al., 2018; Elia Venanzi et al., 2024).

Modern collective variables are learned, finding a collective coordinate system that retains crucial information of the dynamic system. Briefly, a linear/non-linear map (E) is estimated which projects the high-dimensional data $\mathbf{X}$ to a lower dimensional space $y = E(\mathbf{X})$ (See Figure 5) (Noé et al., 2020). Common examples include principal component analysis (PCA), and time-lagged independent component analysis (tICA) (Bhakat, 2022; Schultze and Grubmüller, 2021), or a more advanced variational approach for Markov processes

(VAMPnets) (Ghorbani et al., 2022; Mardt et al., 2018), These are frequently used to represent the dynamic enzyme system and can help with visualizing the relative population of conformational states (Acevedo-Rocha et al., 2021; Agarwal et al., 2020; Curado-Carballada et al., 2019; Romero-Rivera et al., 2017).

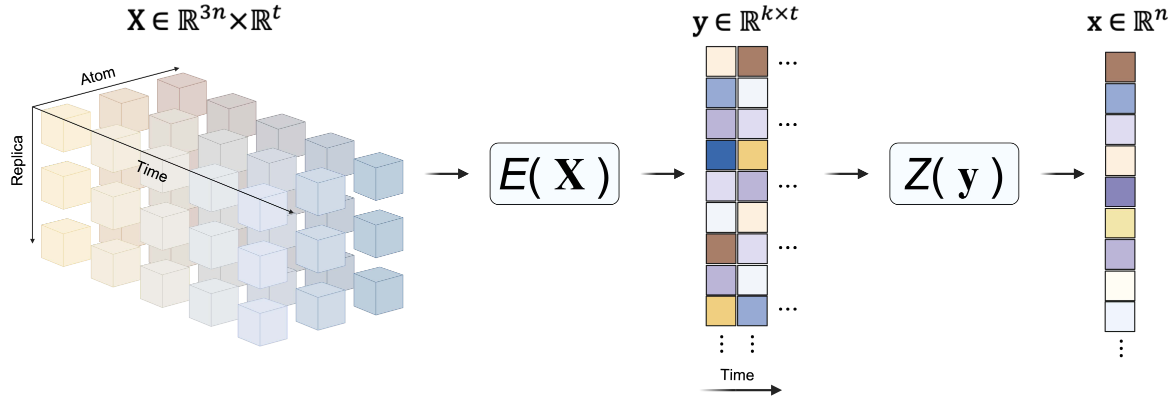

$\mathbf{X} \in \mathbb{R}^{3n} \times \mathbb{R}^{t}$
Atom
Replica
Time
$E(\mathbf{X})$
$\mathbf{y} \in \mathbb{R}^{k \times t}$
Time
$Z(\mathbf{y})$
$\mathbf{x} \in \mathbb{R}^{n}$

Fig. 5. Procuring protein representations from dynamics. Dynamics are often studied using high dimensional MD simulations, with **X** containing both multidimensional spatial and temporal information. Using a map, *E*, lower-dimensional collective variables that summarise the relevant dynamics of the system can be extracted. The dimensions can be further reduced by averaging over the temporal dimension, *Z*(**y**), obtaining time-averaged variables.

In analogy with collective variables, many dynamic representations often remain a function of time, and time-averaged measures are thus beneficial to further reduce the dimensionality (Z(y) in Figure 5). For example, root-mean-square deviation (RMSD, $\mathbb{R}^{n}(t)$) is a time-dependent measure, but root-mean-square fluctuation (RMSF, $\mathbb{R}^{n}$) is not. Time-averaged measures are popular as they can reduce geometric collective variables (*e.g.* distance fluctuations) to a single scalar value. While this summarises the entire time series, it is inherently coarse-grained, thus potentially losing the representation of key dynamic behavior. Nevertheless, the time-dependent and independent measures (RMSD and RMSF, respectively) and their variance remain key representations of rigid and mobile regions in enzymes as well as or indicators of whether catalytically conducive conformations are sampled. These features can be thought of in the context of the aforementioned map *f*, in this case, *Z*(*E*(**X**)), which produces a low-dimensional representation $\mathbb{R}^{n}$ by summarising the variability of a collection of structures **X** across a simulation (Ainsley et al., 2018; Audagnotto et al., 2022; Kamerlin and Warshel, 2010).

4.4 Multi-state Design

Another state-of-the-art strategy is to employ energy-centric methods. These methods cannot explain anything past the Boltzmann diversity on the conformational information hierarchy and assume that hinge motions or other major conformational states can be slightly perturbed in their stability by mutation to favor a desired conformation. These major conformational states may be contributing to substrate specificity and activity, thus a multi-state design accounts for the relevant $\Delta\Delta G$ of mutations with respect to the change in conformation (St-Jacques et al., 2023). This energy-centric representation associates an energy value with each mutant and conformational state, which may be used to assess the relative stability of conformational states. In terms of *f*, each structure **x** is assigned an energy which drastically reduces the dimensionality of the representation.

4.5 Shortest Path Map; A Dynamic Representation

At equilibrium, a more informative representation of dynamics may instead be derived from long-duration MD simulations. These representations elucidate allosteric networks (communication paths between distal residues and the active site) and can be obtained by considering the dynamic cross-correlation matrix made of elements

$$C_{ij} = \frac{\langle \Delta r_i \cdot \Delta r_j \rangle}{\sqrt{\langle r_i^2 \rangle \langle r_j^2 \rangle}}$$

where C_{ij} is the dynamic cross-correlation between residue *i* and *j*, $\langle \Delta r_i \cdot \Delta r_j \rangle$ is the time-averaged displacement from the mean coordinate of residue i and j, and $\sqrt{\langle r^2{}_i \rangle \langle r^2{}_j \rangle}$ is a normalization factor. This representation was developed by the group of Silvia Osuna and recently deployed as a web server (Casadevall et al., 2024), conferring accessibility of dynamic representations. The measure lies one rank above residue-independent measures such as RMSF, as it treats pairs of residues in a dynamic, but time-averaged, context (Morra et al., 2012). One obtains a representation of $\mathbb{R}^{n \times n}$, where *n* is the number of atoms, a square matrix with information about the covariance of residues. The allosteric networks derived from this representation have been strongly correlated with distal mutations and subsequent effects on catalytic activity. In fact, many directed evolution campaigns accumulate mutations along allosteric networks in retro-aldolase, tryptophan synthase, cytochrome P450 oxygenase, imidazole glycerol phosphate synthase, and protein tyrosine phosphatase (Acevedo-Rocha et al., 2021; Calvó-Tusell et al., 2022b; Crean et al., 2021; Gergel et al., 2023; Maria-Solano et al., 2021; Romero-Rivera et al., 2022, 2017). Alternatively, asymmetric measures have also become prevalent, describing the directionality in coupling and thus elucidating residues controlling dynamics (Kazan et al., 2023).

During catalytic transformation, non-equilibrium dynamics have been observed using advanced MD tools. This so-called D-NEMD method is an alternative but complimentary way of representing allosteric networks from which one obtains a time-dependent vector, $R^n(t)$, that carries information about communication pathways in the catalytic cycle (Castelli et al., 2024; Oliveira et al., 2021).

4.6 Learned Dynamic Representations and Future Directions

Finally, to address conformational transitions using a full description of conformational dynamics, Markov state models (MSM) are critical as they capture both relative populations and inter-conversion timescales between conformational states (Chodera and Noé, 2014). Despite their initial challenges (Konovalov et al., 2021), MSMs have successfully been applied to explain the dynamic behavior of many enzymes, *e.g*., polymerases, isomerase, glycosylases, and synthase (Gordon et al., 2016; Konovalov et al., 2021; Wapeesittipan et al., 2019). With subsequent advances in ML, the collective variables are learned and extracted to form a thermodynamic and kinetic basis for understanding the enzyme in question (Ghorbani et al., 2022; Mardt et al., 2018). They are typically represented by a transition probability matrix ($\mathbb{R}^{|S| \times |S|}$ where |S| is the number of discrete states) and a stationary distribution ($\pi = [\pi_1, \ldots, \pi_{|S|}]$) describing the relative population of states, which are obtained from long-duration MDs.

The representations above are often derived from long-duration MD simulations, and thus limit the use of dynamics data in ML due to their computational cost. This tension lies in the discrepancy between the femtosecond time step of MDs and the microsecond-millisecond timescales at which large conformational changes occur that are important for enzymatic catalysis.

In principle, however, MD is not the only approach for obtaining a collection of structures X. The field is currently addressing this through the use of ML tools and DL generative models, where X is considered as being derived from a probability distribution $p(x)$. Generating X is thus a question of sampling from $p(x)$. It has been shown that AlphaFold2 can be used to obtain various conformational states of proteins by feeding shallow MSAs (Casadevall et al., 2023; Sala et al., 2023; Wayment-Steele et al., 2024). These methods only obtain conformational diversity on the information hierarchy but have subsequently been extended toward Boltzmann diversity using seeded MD simulations (Audagnotto et al., 2022; Vani et al., 2023). Alternatively, a combination of AlphaFold2 and generative models has also been developed to enable the generation of conformational ensembles (Jing et al., 2024). Thus, a rapidly expanding toolkit with which conformational ensembles can be generated is being established (Arts et al., 2023; Bose et al., 2023; Mansoor et al., 2023; Noé et al., 2020), enabling dynamic representations to be used for in biocatalysis.

5. Protein-Substrate Representations

In previous sections, the emphasis has been on the featurisation of the protein. However, those strategies do not consider the possible interactions with the protein environments, *e.g.*, solvents, ligands, substrates, or cofactors. This is an integral part of biocatalysis and constitutes a treasure trove of information that could prove beneficial in the training of ML models. The inclusion of protein-substrate interactions would, in most cases, include molecular docking, but could also involve protein dynamics, QM/MM simulations, or even crystallized complexes (Bonk et al., 2019). This could, in turn, assist in addressing tasks such as predicting substrate specificity or elucidating the structure-function of enzymes (Berselli et al., 2021). Within the realm of ML, features extracted from substrate-docking have yet to be fully leveraged (Ao et al., 2024) and are possibly challenged by difficulties in translating protein-substrate complexes into a numerical and general representation. However, some studies have successfully included information harvested from protein-substrate complexes for ML models employing different strategies which will be introduced in this section (Figure 6).

5.1 Molecular Docking-based Descriptors and Binding Energies

One strategy to generate descriptors of the protein-substrate binding involves the use of scoring functions derived from the docking. For example, the scoring functions from Rosetta (Davis and Baker, 2009; Meiler and Baker, 2006) can be combined with physicochemical and active site descriptors to train a model that can predict the substrate scope of bacterial nitrilases (Mou et al., 2021). The scoring functions described interfacial interaction energy terms including full-atom van der Waals attraction, electrostatics, van der Waals repulsion, hydrogen bonding terms, and solvation energy. From all the features used to train the random forest model, the attractive part of the Lennard-Jones potential obtained from the molecular docking scoring functions was revealed to be the most consistently important variable for the model's performance. A similar approach has been employed to predict the site of metabolism for cytochrome P450 monooxygenases and their substrates in multiple instances (Feng et al., 2023; Huang et al., 2013; Zaretzki et al., 2013, 2011). One example included the use of substrate interaction-based descriptors derived from Autodock Vina (Eberhardt et al., 2021; Trott and Olson, 2010) along with chemical reactivity descriptors to train a multiple-instance ranking algorithm (Huang et al., 2013). The model was then used to predict the site of metabolism of the substrates of two cytochrome P450 enzymes, yielding an accuracy of the top two predicted rank positions of 86 % and 83 %, respectively for the two isoforms.

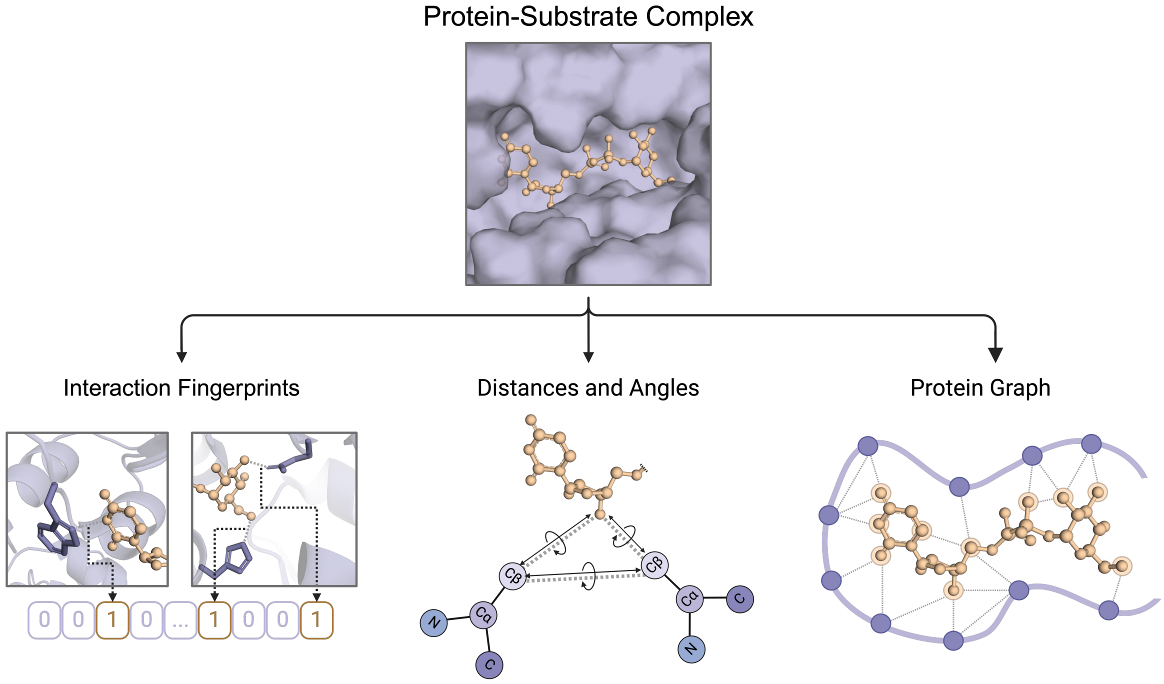


Fig. 6. Approaches for encoding the protein-substrate complexes. The protein-substrate complex can be encoded based on the intermolecular interactions into a binary string commonly denoted as a fingerprint (left). The complex can also be represented by the dihedral angles and distances between catalytic residues along with the angles and distances between catalytic residues and the substrate (middle). Lastly, the protein-substrate complex can be converted into a graph representation where the nodes represent the atoms and the edges represent the interaction between two atoms (right). Notably, while not shown, the complexes can also be represented using scoring functions.

A slightly different route was taken in a study of the bile acid specificity in a single bile acid hydrolase (WT and two mutational variants) (Karlov et al., 2023). Here, a previously published complex of the bile acid hydrolase and a bile acid was used as a template to model the complex with other bile acid substrates with MD simulations. The last nanosecond of a 100 ns simulation was used for binding energy calculations employing molecular mechanics Poisson-Boltzmann surface area and molecular mechanics generalized Born surface area methods implemented in AmberTools (Case et al., 2023). The calculated binding energies were then correlated with the corresponding activity data using linear regression which led to the identification of structural determinants of substrate binding and specificity.

5.2 Interaction fingerprinting

Another way of representing protein-substrate interactions is through interaction fingerprinting which captures the protein-substrate interactions in one-dimensional binary representations (Figure 6) (Desaphy et al., 2013). This method was utilized for predicting kinase inhibitors by comparing models trained on ligand-interaction fingerprints with models trained on molecular fingerprints of the substrates (Witek et al., 2014). Here, the models trained on the interaction fingerprints outperformed the models trained on molecular fingerprints in discriminating between active and inactive compounds. The use of interaction fingerprints was also explored in a model trained to predict the ligand affinity of HIV-1 protease inhibitors (Leidner et al., 2019). The authors extracted interaction fingerprints from crystallized protein-substrate complexes harvested from the Protein Data Bank (Berman et al., 2000), adapting the binary encoding into continuous features describing selected non-covalent interactions. These interaction fingerprints were used to train a gradient-boosting model achieving an RMSE of 1.48 kcal/mol. The study also demonstrated the interpretability of the model using Shapley values which elucidated that van der Waals interactions were critical for model performance.

5.3 Distance and Angle-based Representations

An alternative encoding strategy for protein-substrate complexes is the use of distances and angles between the substrate and surrounding residues (Figure 6). This was leveraged in a study of hydrolases for the breakdown of several classes of substrates (Ran et al., 2023). Here, the authors aimed to construct a model that could predict the hydrolytic activation free energy for the reactive complexes of hydrolase-catalyzed reactions along with the favored enantiomer of the product. The ability to predict the enantiomeric outcome was enabled by including an atomic distance map consisting of atomic distances between a docked substrate and the Cα atoms of the surrounding catalytic residues transformed into a tensor by a single-layer CNN. This map was concatenated with the dihedral angles of the docked substrate converted into sine and cosine values. Combined with sequence-based representations and substrate SMILES, this model could classify reactive and unreactive poses achieving an AUC of 0.87 and a good Pearson R value of 0.72. The model predicted the enantiomeric preference with an accuracy of 55 %. Distances and angles between substrate and enzyme were also employed in a study of ketol-acid reductoisomerases (Bonk et al., 2019). The 68 generated features, consisting of distances and angles between catalytic residues, substrate, cofactor, and active site waters, and magnesium ions, were regularised using LASSO regression, fed to a logistic classifier, and subsequently clustered. The trained model could differentiate between reactive and almost-reactive trajectories with >85 % accuracy. Furthermore, ranking the features from LASSO enabled the identification of a subpart of the reactive site to be particularly important in describing the activity of the enzyme.

5.4 Graph Neural Networks for Protein-Substrate Interactions

Lately, GNNs have been readily employed to capture detailed information from the protein-substrate complex by converting the docking pose into a graph representation where the nodes represent the atoms and the edges represent their interaction (Yang et al., 2023). This could include the interaction between protein and substrate, between protein and protein, and between substrate and substrate (Figure 6) (Lu et al., 2023; Xia et al., 2023). While not in the realm of biocatalysis, this technique has been used to improve the accuracy of scoring functions of molecular docking (Wang et al., 2022; L. Yang et al., 2021) and to predict protein-ligand affinities (Mastropietro et al., 2023; Wang et al., 2023), especially within drug discovery (Z. Yang et al., 2022). Since enzymes do not solely rely on binding affinity for their functionality, one cannot draw direct parallels between the use of GNNs in these cases and in the case of predicting/understanding the substrate scope of enzymes. However, one study used a GNN-based model to predict and interpret the substrate specificity of multiple mutational variants of two model proteases (Lu et al., 2023). This was achieved by developing a protein graph convolutional network that could model protein structures and their complexes as fully connected graphs where each node corresponded to an amino acid from either the protein or the peptide-substrate while the edges represent the pairwise residue interactions between the nodes. The generated model could ultimately predict protease activity with a given substrate achieving an accuracy >85 % across protease variants. In addition, the authors also displayed how node and edge ablation tests provided insights into the feature importance of the models. In a model that only included sequence-based features, the edges did not affect the model accuracy, and the peptide nodes played a leading role. However, when energy-based features were included, ablating edge-based features significantly impacted the model accuracy with the intermolecular edges being particularly important.

Overall, the use of protein-substrate complexes to generate representations holds great promise within ML for biocatalytic systems. Many of the described methods capture interpretable information which is useful in cases where explainability is an important factor. However, one should still keep in mind that obtaining protein-substrate complexes is computationally demanding when using molecular docking, making the method realistic for smaller datasets, at least until the ML-based docking methods significantly accelerate the process (Buttenschoen et al., 2024). In addition, molecular docking is not an accurate method, especially without manual inspection of poses, which could directly impact the accuracy of the model.

6. Choosing a Suitable Representation

Selecting the most appropriate representation approach when constructing models can be a challenging task, and although several attempts have been made to examine the efficacies of different encoding techniques (Elabd et al., 2020; Goldman et al., 2022; Michael et al., 2023; Bruce J. Wittmann et al., 2021), no consensus exists for determining the best representation for a new protein ML model. Consequently, finding a suitable protein representation remains case-dependent. To address this issue, we propose two general factors to consider (Figure 7). The first factor is the model setup, determining the overall design of the predictive tool. This includes the size of the training dataset, defining the ease of discovering hidden patterns, and the choice of ML architecture, imposing requirements for the input representation. The second factor is the model objective, describing the type of task envisioned for the resulting model. Linking the choice of representation with project objectives such as the assayed property, wild type vs. mutational predictor, and explainability may eventually increase the chances of achieving these objectives. We expect that these two factors can be used as a source of inspiration and guidance when creating new ML models for biocatalysis.

6.1 Model Setup

When developing an ML model, design decisions are often made based on element harmony, where the size of the dataset matches the model architecture. This is also applicable to the choice of a suitable protein representation, and selecting a harmonious encoding strategy based on the model setup is extremely important. In this section, we will discuss how model design can influence the appropriate representation approach.

6.1.1 Size of Dataset

An important feature of the model setup is the size of the dataset. Here, a protein representation approach that produces a large feature set might be problematic when encoding smaller data sets due to a poor data-to-feature ratio, as the high dimensionality introduces sparsity and higher chances of finding patterns in feature noise. This can lead to significant overfitting, thus hindering the identification of hidden patterns and trends in the data which is crucial for an efficient and accurate predictive model (Bellman, 1961; Theodoridis and Koutroumbas, 2008). The low-to-medium-throughput nature of experiments is a common issue in biocatalysis, which imposes significant restrictions on the choice of suitable representations for ML to ensure only informative features are incorporated.

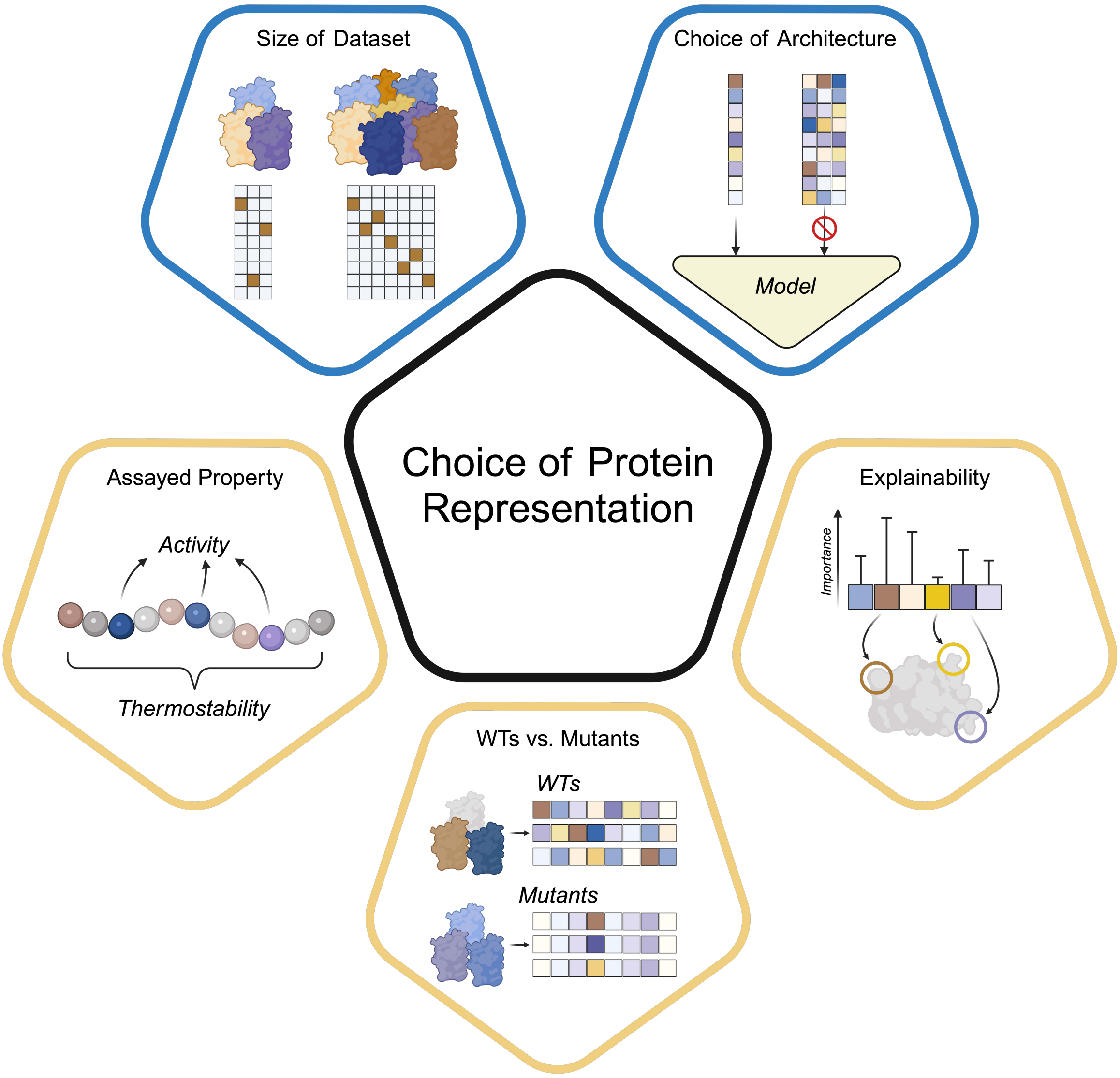


Fig. 7. Factors influencing the choice of a suitable protein representation. The first main factor is “model setup” (colored blue), which concerns the size of the dataset due to small datasets potentially preventing the discovery of patterns contained in sparse representations. The choice of ML architecture might instead impede the use of certain representations due to incompatibility. The second main factor is “model objective” (colored beige), as specialized representation might enhance models for predicting assayed enzyme properties such as activity, while full representations will likely better suit global properties, *e.g.*, thermostability. Furthermore, WT models impose different requirements on the encoding strategy than mutant predictors due to the disparity in representation similarity. Finally, any explainability task will benefit from a clear connection between the model features and protein features.

A promising strategy to circumvent this problem is to leverage the large pre-trained models for self-supervised representation learning (Ferruz and Höcker, 2022; Notin et al., 2023; Qiu and Wei, 2023). A notable example of this is the approach introduced by Biswas et al., which involved fine-tuning the deep neural network UniRep by using the sequences evolutionarily related to their protein of interest, GFP, thus adapting the resulting latent vector embeddings to better encode protein information crucial to the evolution of GFP (Biswas et al., 2021). The resulting ML models were capable of identifying mutants with increased fluorescence using as few as 24 mutants as training data. Biswas et al. observed a large sequence diversity in the new model-based variants, suggesting that the increased density of evolutionary important information contained in the protein representation due to the fine-tuning procedure allowed for a greater exploration of the sequence-to-function space.

Related to utilizing knowledge from pre-trained embeddings, insights obtained from a mutational study of a single enzyme can be transferred to homologues with little characterization. This is known as transfer learning which entails training models on large datasets to study scarce datasets (Yosinski et al., 2014). This could eliminate the requirement of conducting a thorough mutational assay every time a new enzyme is examined and facilitate Low-N modeling, though this is yet to be explored for biocatalysis.

Alleviating the issue of a low amount of data can be done with the previously mentioned approach of augmenting a VAE-based evolutionary density score with a simple OHE (Hsu et al., 2022). Models trained on as few as 48 proteins exhibited good performance when utilizing this augmentation technique. This finding highlights how combining representations containing different protein information can be beneficial.

Notably, while a low amount of data is a significant hindrance for most encoding strategies, a large dataset might instead hinder the use of representations requiring significant processing power. This includes methods for QM calculations or MD simulations, as their computational demands make them infeasible for datasets with a large selection of proteins. This might be especially relevant for predictive models trained on dynamics representations, as the acquisition of such protein encodings is often computationally expensive, introducing a question of balance between a larger dataset and an increased usage of computational resources.

Lastly, while the size of the training dataset is extremely influential for the choice of suitable representation, another important related step is the split between test and training data. Here, the choice of representation influences the preferred approach for cross-validation due to the different types of information bias(Corso et al., 2024; Kanakala et al., 2022; Kroll and Lercher, 2023; J. Li et al., 2023). It is important to harmonize the dataset validation strategy with the protein representation.

6.1.2 Choice of Architecture

Even though the choice of model architecture is often related to the amount of training data available due to how the performance of ML algorithms often depends on the size of the dataset (Beleites et al., 2012; Raudys and Jain, 1991), the architecture imposes different requirements to the representation than those described in the previous section. While innumerable ML architectures have been developed, researchers are more likely to build models inside of their field of expertise. Therefore, the model architecture is often determined before the encoding approach, and the choice of protein representation is therefore strongly influenced by the model architecture. Classical ML methods, such as logistic regression, KNN, and random forest, usually require a 1D vector with numerical values. Consequently, any multidimensional information must either be flattened or reduced in dimensions before use in these models, potentially losing the important data structure contained in the representation. Employing a representation with a large feature set together with the simplest of architectures might also cause problems due to their limited capacity to discover the patterns in the feature set.

Some protein representations might require the use of advanced DL architectures such as GNNs and CNNs as highlighted in the description of structure representations. If a researcher's field of expertise is mainly CNNs, combining these ML architectures with a protein voxel representation is likely more beneficial than attempting to employ protein graphs and GNNs. Consequently, the generalisability of fixed descriptors is quite advantageous.

Finally, some ML models have shown dispositions towards memorization instead of generalization (Buttenschoen et al., 2024; Corso et al., 2024; Kroll and Lercher, 2023; Wallach and Heifets, 2018). Rather than learning a fundamental relationship between the proteins and their function through the model features, they memorize all individual representations in the training set which leads to a high degree of overfitting. If the chosen architecture tends to achieve high validation accuracy due to such memorization, we propose to employ fixed encoding strategies instead of learned representation. This is due to the latter often behaving as a fingerprint with few similarities between two representations, while a set of proteins encoded with fixed representations often has the same values across different descriptors. In consequence, the model will be less likely to turn towards memorization when these fixed features are used.

6.2 Model Objective

The second factor that influences the choice of suitable protein representation is the objective envisioned for the ML model. Certain enzyme properties might benefit from using specialized representation methods. Another important distinction comes from the contrast between training models on WT and mutational data. Finally, we will discuss tasks in which explainability is essential.

6.2.1 Assayed Property

If the objective of the model is to examine the activity or specificity of the enzymes, it is crucial to encode the active site — potentially only focusing on the area of the protein containing this site. In our recent model for glycosyltransferase acceptor specificity predictions, we limited the representation to contain only the N-terminal domain which contains the acceptor binding site (Harding-Larsen et al., 2023). The structure-informed ASC method also allowed Röttig et al. to focus the representation on the active site (Röttig et al., 2010). Other examples of the representations targeting task-specific parts of the protein include the domain embeddings of Domain-PFP for predicting Gene Ontology (GO) annotations (Ibtehaz et al., 2023), the site embeddings and encoding of neighbouring regions N-linked glycosylation site predictions in EMNGly (Hou et al., 2023), and the microenvironments of MutCompute used for identifying position where mutations can stabilize the local environment (Paik et al., 2023; Shroff et al., 2020).

However, as previously described, limiting the representation to specific areas of the protein can potentially remove important information, such as for allostery or protein fitness. To capture this information, a more general protein encoding will be more suitable to allow the resulting ML model to explore the entire sequence and structure landscape.

6.2.2 Wild Type vs Mutational Data

Aside from predicted property, the type of enzymes, be it mutants or wild-type (WT) proteins, will also significantly influence the choice of representation as two variants of the same enzyme are inherently more similar than two WT proteins from the same family. An ML model trained on mutant data can thus utilize more specialized protein representations than a model trained on WT data due to a significant portion of the sequence being constant across every variant. This strategy was employed by Saito et al. to encode variants of Sortase A for use in MLDE by only encoding five positions known to result in a high-activity variant, ultimately achieving an improved variant of the enzyme (Saito et al., 2021). Such an approach will not be possible for a WT predictor, as not only will large portions of the proteins potentially differ, but the length of each protein is unlikely to be equal.

Due to the limited variance contained in the sequences of mutant datasets, the representation strategies require higher sensitivity to the minute changes between each variant. Otherwise, the resulting ML model will be unable to discern top-performing variants from those of poor nature. Unfortunately, no gold standard has been established for the sensitivity of encoding techniques, and it is therefore difficult to determine the best representation strategy in this endeavour. Wittmann et al. proposed that learned embeddings obtained from models trained on MSAs will result in representations containing a higher density of information important for mutational tasks due to highlighting which mutations are evolutionarily feasible (Bruce J. Wittmann et al., 2021). Nevertheless, they only observed small performance increases when using embeddings from MSA Transformer (Rao et al., 2021), highlighting how a suitable representation can be highly case-dependent. Consequently, new representation learning models should be benchmarked through large collections of diverse datasets such as the deep mutational scans collected in ProteinGym (Notin et al., 2023).

WT models do not have the same sensitivity issue due to the larger variance between the training sequences. This is of course by design, as WT models often remove proteins within a preset similarity cutoff. Instead, the representation of WT proteins introduces a question of compatibility across all proteins in both the training and test data. Methods requiring sequence alignments, such as OHE, BLOSUM encodings, or structure-informed approaches, will not work with sequences of low similarity. Here, graph models trained on structurally heterogeneous enzymes might be superior.

6.2.3 Explaining Protein Representations

In some studies, the model objective is mainly to produce a predictive model that can be utilized for future *in silico* scoring of potential variants or WT enzymes for a given reaction. In that case, the representation strategy producing the highest accuracy is likely desired. However, if the purpose of the model is instead to obtain a fundamental understanding of the forces governing the protein function and the modeled process, the explainability of the model is crucial.

Recently, the notion of Explainable AI (XAI) has gained momentum, with terms such as explainability, interpretability, and justification being regarded as increasingly valuable for new models (Novakovsky et al., 2022; Vilone and Longo, 2020; Wellawatte et al., 2023; Wojciech Samek et al., 2019). In ML for biocatalysis, the ability to explain model decisions actively allows a more thorough understanding of enzyme features and phenotypes. However, as XAI mainly addresses the *model* features, the accuracy of said explanations depends on the connection between model features and protein properties — a connection, that is defined by the encoding strategy.

If the model features represent inherent amino acid characteristics such as physicochemical properties, incorporation of XAI can help pinpoint which of these residue features are important for model predictions. This knowledge may lead to novel insights as well as potentially assist in choosing targets for the rational design of new variants with enhanced enzymatic properties. XAI was utilized by Robinson et al. to elucidate the essential residues for the activity of thiolase members of the OleA enzyme family (Robinson et al., 2020) and by Taujale et al. to discover a buried residue important for the donor specificity of fold A glycosyltransferases (Taujale et al., 2020).

If coarse-grained protein properties are implemented in the model features, the ability to identify important amino acid attributes is reduced. Here, the implementation of XAI can instead be utilized to compare the influence of the different protein characteristics, an approach taken by Heckman et al. to highlight the importance of structural properties for the activity of metabolic enzymes at the genome scale (Heckmann et al., 2020, 2018), as well as by Mou et al. (Mou et al., 2021) and Carlin et al. (Carlin et al., 2016) to identify key ligand binding-related features for nitrilase substrate specificity and glycoside hydrolase kinetics, respectively.

Finally, encoding the protein using learned embeddings introduces some interesting challenges in XAI, as the abstract representation often does not translate directly to specific properties in the protein. Consequently, explaining the protein properties based on the importance of the model features is even more complicated than for the coarse-grained representations. One solution is to use an attention mechanism when constructing the protein embeddings, as implemented by Li et al. when examining the positional importance with regard to the k_{cat} of WT metabolic enzymes (Li et al., 2022). Due to the DL nature of their model architecture, they would have been unable to directly extract the feature importance of their model (Wellawatte et al., 2023; Wojciech Samek et al., 2019). Here, the authors incorporated an additional sub-architecture, the attention mechanism, that allows the model to "remember" the connection between input properties and embedding features (Bahdanau et al., 2014; Li et al., 2022; Wellawatte et al., 2023).

Instead of changing the architecture, the model decisions can also be elucidated using input perturbation such as *in silico* mutagenesis, where the input sequence is perturbed by changing a single amino acid and then examining the difference between the model prediction of the original and new sequence (Novakovsky et al., 2022; Zhou and Troyanskaya, 2015). This difference, also known as the attribution score (Novakovsky et al., 2022), can then be calculated for a large number of perturbations, ideally, all possible ones, resulting in a thorough sequence-function landscape of the ML model. This landscape can be examined to determine the key residue properties, thus introducing explainability to an inherently abstract protein representation and modeling approach.

7. Summary & Outlook

In this review, we have presented a diverse selection of the most prominent strategies for encoding enzyme information for ML modeling. The representation approaches are capable of utilizing varying levels of protein information, from primary sequence to temporal dynamics, and their complexities range from fixed descriptors with little inherent bias to learned presentations extracted from complex DL models. To navigate this ever-growing field, we introduced two main factors for choosing the most suitable encoding strategy: “model setup”, especially concerning the training dataset size and ML architecture, and “model objective”, relating to the assayed enzyme property, the differences between a WT model and mutant predictor, and explainability of the model. We believe that this review serves as both a source of information and a guide for future researchers in biocatalysis when determining a suitable encoding strategy for their own ML models. The field is rapidly expanding, and we envision a promising future for the development and use of more sophisticated protein encodings. Solving the Low-N objective is a pressing objective, and future approaches should build on the pioneering work of fine-tuning pre-trained PLM embeddings or the combination of representations containing distinct information and inherent bias. Another vital task is to efficiently incorporate protein dynamics representations due to their ability to capture crucial aspects of enzymatic behavior. Lastly, we hope that future ML projects for biocatalysis will ensure a better alignment between the choice of protein representation and model design.

Acknowledgments

We thank The Novo Nordisk Foundation for supporting this work through grant NNF20CC0035580. This work was also supported by Czech Ministry of Education, Youth and Sports [CETOCOEN Excellence CZ.02.1.01/0.0/0.0/17_043/0009632, ESFRI RECETOX RI LM2023069, ESFRI ELIXIR LM2023055] and the CETOCOEN EXCELLENCE Teaming 2 project supported by Horizon2020 of the European Union [857560].

Figures were created with BioRender.com

Declaration of Competing Interest

The authors declare no competing interests.

References

Acevedo-Rocha, C.G., Gamble, C.G., Lonsdale, R., Li, A., Nett, N., Hoebenreich, S., Lingnau, J.B., Wirtz, C., Fares, C., Hinrichs, H., Deege, A., Mulholland, A.J., Nov, Y., Leys, D., McLean, K.J., Munro, A.W., Reetz, M.T., 2018. P450-Catalyzed regio- and diastereoselective steroid hydroxylation: Efficient directed evolution enabled by mutability landscaping. ACS Catal 8, 3395–3410. https://doi.org/10.1021/ACSCATAL.8B00389/ASSET/IMAGES/LARGE/CS-2018-003898_0005.JPEG

Acevedo-Rocha, C.G., Li, A., D'Amore, L., Hoebenreich, S., Sanchis, J., Lubrano, P., Ferla, M.P., Garcia-Borràs, M., Osuna, S., Reetz, M.T., 2021. Pervasive cooperative mutational effects on multiple catalytic enzyme traits emerge via long-range conformational dynamics. Nat Commun 12. https://doi.org/10.1038/s41467-021-21833-w

Agarwal, P.K., Bernard, D.N., Bafna, K., Doucet, N., 2020. Enzyme dynamics: Looking beyond a single structure. ChemCatChem 12, 4704–4720. https://doi.org/10.1002/cctc.202000665

Ahdritz, G., Bouatta, N., Kadyan, S., Xia, Q., Gerecke, W., O, T.J., Berenberg, D., Fisk, I., Zanichelli, N., Zhang, B., Nowaczynski, A., Wang, B., Stepniewska-Dziubinska, M.M., Zhang, S., Ojewole, A., Efe Guney, M., Biderman, S., Watkins, A.M., Ra, S., Ribalta Lorenzo, P., Nivon, L., Weitzner, B., Andrew Ban, Y.-E., Sorger, P.K., Mostaque, E., Zhang, Z., Bonneau, R., AlQuraishi, M., Allen Hamilton, B., Bio, C., 2022. OpenFold: Retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization. bioRxiv 2022.11.20.517210. https://doi.org/10.1101/2022.11.20.517210

Ainsley, J., Mulholland, A.J., Black, G.W., Sparagano, O., Christov, C.Z., Karabencheva-Christova, T.G., 2018. Structural Insights from Molecular Dynamics Simulations of Tryptophan 7-Halogenase and Tryptophan 5-Halogenase. ACS Omega 3, 4847–4859. https://doi.org/10.1021/acsomega.8b00385

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25, 3389–3402. https://doi.org/10.1093/NAR/25.17.3389

A.Maria-Solano, M., Serrano-Hervás, E., Romero-Rivera, A., Iglesias-Fernández, J., Osuna, S., 2018. Role of conformational dynamics in the evolution of novel enzyme function. Chemical Communications 54, 6622–6634. https://doi.org/10.1039/C8CC02426J

Amidi, A., Amidi, S., Vlachakis, D., Megalooikonomou, V., Paragios, N., Zacharaki, E.I., 2018. EnzyNet: enzyme classification using 3D convolutional neural networks on spatial representation. PeerJ 6. https://doi.org/10.7717/PEERJ.4750

Ao, Y.F., Dörr, M., Menke, M.J., Born, S., Heuson, E., Bornscheuer, U.T., 2024. Data-Driven Protein Engineering for Improving Catalytic Activity and Selectivity. Chembiochem 25. https://doi.org/10.1002/CBIC.202300754

Arnold, F.H., 2018. Directed Evolution: Bringing New Chemistry to Life. Angew Chem Int Ed Engl 57, 4143. https://doi.org/10.1002/ANIE.201708408

Arnold, F.H., 1998. Design by Directed Evolution. Acc Chem Res 31, 125–131. https://doi.org/10.1021/AR960017F/ASSET/IMAGES/LARGE/AR960017FF00005.JPEG

Arnold, F.H., 1996. Directed evolution: Creating biocatalysts for the future. Chem Eng Sci 51, 5091–5102. https://doi.org/10.1016/S0009-2509(96)00288-6

Arts, M., Frellsen, J., Boomsma, W., 2023. Internal-Coordinate Density Modelling of Protein Structure: Covariance Matters. ArXiv.

Atchley, W.R., Zhao, J., Fernandes, A.D., Drüke, T., 2005. Solving the protein sequence metric problem. Proc Natl Acad Sci U S A 102, 6395–6400. https://doi.org/10.1073/PNAS.0408677102/SUPPL_FILE/08677TABLE5.XLS

Audagnotto, M., Czechtizky, W., De Maria, L., Käck, H., Papoian, G., Tornberg, L., Tyrchan, C., Ulander, J., 2022. Machine learning/molecular dynamic protein structure prediction approach to investigate the protein conformational ensemble. Sci Rep 12, 10018. https://doi.org/10.1038/s41598-022-13714-z

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G.R., Wang, J., Cong, Q., Kinch, L.N., Dustin Schaeffer, R., Millán, C., Park, H., Adams, C., Glassman, C.R., DeGiovanni, A., Pereira, J.H., Rodrigues, A. V., Van Dijk, A.A., Ebrecht, A.C., Opperman, D.J., Sagmeister, T., Buhlheller, C., Pavkov-Keller, T., Rathinaswamy, M.K., Dalwadi, U., Yip, C.K., Burke, J.E., Christopher Garcia, K., Grishin, N. V., Adams, P.D., Read, R.J., Baker, D., 2021. Accurate prediction of protein structures and interactions using a three-track neural network. Science (1979) 373, 871–876. https://doi.org/10.1126/SCIENCE.ABJ8754/SUPPL_FILE/ABJ8754_MDAR_REPRODUCIBILITY_CHECKLIST.PDF

Bahdanau, D., Cho, K.H., Bengio, Y., 2014. Neural Machine Translation by Jointly Learning to Align and Translate. 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings.

Baxter, J., 2000. A model of inductive bias learning. Journal of artificial intelligence research 12, 149–198.

Behera, S., Balasubramanian, S., 2023. Lipase A from Bacillus subtilis: Substrate Binding, Conformational Dynamics, and Signatures of a Lid. J. Chem. Inf. Model. https://doi.org/10.1021/acs.jcim.3c01681

Beleites, C., Neugebauer, U., Bocklitz, T., Krafft, C., Popp, J., 2012. Sample Size Planning for Classification Models. Anal Chim Acta 760, 25–33. https://doi.org/10.1016/j.aca.2012.11.007

Bellman, R., 1966. Dynamic programming. Science (1979) 153, 34–37.

Bellman, R., 1961. Adaptive control processes : a guided tour. Princeton University Press.

Bengio, Y., Courville, A., Vincent, P., 2013. Representation learning: A review and new perspectives. IEEE Trans Pattern Anal Mach Intell 35, 1798–1828.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E., 2000. The Protein Data Bank. Nucleic Acids Res 28, 235–242. https://doi.org/10.1093/NAR/28.1.235

Berselli, A., Ramos, M.J., Menziani, M.C., 2021. Novel Pet-Degrading Enzymes: Structure-Function from a Computational Perspective. Chembiochem 22, 2032–2050. https://doi.org/10.1002/CBIC.202000841

Bhakat, S., 2022. Collective variable discovery in the age of machine learning: reality, hype and everything in between. RSC Adv 12, 25010. https://doi.org/10.1039/D2RA03660F

Bhattacharya, S., Margheritis, E.G., Takahashi, K., Kulesha, A., D’Souza, A., Kim, I., Yoon, J.H., Tame, J.R.H., Volkov, A.N., Makhlynets, O. V, Korendovych, I. V, 2022. NMR-guided directed evolution. Nature 610, 389–393. https://doi.org/10.1038/s41586-022-05278-9

Bilal, M., Adeel, M., Rasheed, T., Zhao, Y., Iqbal, H.M.N., 2019. Emerging contaminants of high concern and their enzyme-assisted biodegradation – A review. Environ Int 124, 336–353. https://doi.org/10.1016/J.ENVINT.2019.01.011

Biswas, S., Khimulya, G., Alley, E.C., Esvelt, K.M., Church, G.M., 2021. Low-N protein engineering with data-efficient deep learning. Nature Methods 2021 18:4 18, 389–396. https://doi.org/10.1038/s41592-021-01100-y

Blaabjerg, L.M., Kassem, M.M., Good, L.L., Jonsson, N., Cagiada, M., Johansson, K.E., Boomsma, W., Stein, A., Lindorff-Larsen, K., 2023. Rapid protein stability prediction using deep learning representations. Elife 12. https://doi.org/10.7554/ELIFE.82593

Bonk, B.M., Weis, J.W., Tidor, B., 2019. Machine Learning Identifies Chemical Characteristics That Promote Enzyme Catalysis. J Am Chem Soc 141, 4108–4118. https://doi.org/10.1021/JACS.8B13879/ASSET/IMAGES/LARGE/JA-2018-138797_0003.JPEG

Bornscheuer, U.T., Pohl, M., 2001. Improved biocatalysts by directed evolution and rational protein design. Curr Opin Chem Biol 5, 137–143. https://doi.org/10.1016/S1367-5931(00)00182-4

Bose, A.J., Akhound-Sadegh, T., Fatras, K., Huguet, G., Rector-Brooks, J., Liu, C.-H., Nica, A.C., Korablyov, M., Bronstein, M., Tong, A., 2023. SE(3)-Stochastic Flow Matching for Protein Backbone Generation. ArXiv.

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

Broom, A., Rakotoharisoa, R. V, Thompson, M.C., Zarifi, N., Nguyen, E., Mukhametzhanov, N., Liu, L., Fraser, J.S., Chica, R.A., 2020. Ensemble-based enzyme design can recapitulate the effects of laboratory directed evolution in silico. Nat Commun 11, 4808. https://doi.org/10.1038/s41467-020-18619-x

Buller, R., Lutz, S., Kazlauskas, R.J., Snajdrova, R., Moore, J.C., Bornscheuer, U.T., 2023. From nature to industry: Harnessing enzymes for biocatalysis. Science 382, eadh8615. https://doi.org/10.1126/SCIENCE.ADH8615/ASSET/28A24AA3-9B29-4667-82B4-E629A9BD74F3/ASSETS/IMAGES/LARGE/SCIENCE.ADH8615-F5.JPG

Bunzel, H.A., Anderson, J.L.R., Hilvert, D., Arcus, V.L., van der Kamp, M.W., Mulholland, A.J., 2021. Evolution of dynamical networks enhances catalysis in a designer enzyme. Nat. Chem. 13, 1017–1022. https://doi.org/10.1038/s41557-021-00763-6

Buttenschoen, M., Morris, G.M., Deane, C.M., 2024. PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. Chem Sci 15, 3130–3139. https://doi.org/10.1039/D3SC04185A

Cadet, F., Fontaine, N., Li, G., Sanchis, J., Ng Fuk Chong, M., Pandjaitan, R., Vetrivel, I., Offmann, B., Reetz, M.T., 2018. A machine learning approach for reliable prediction of amino acid interactions and its application in the directed evolution of enantioselective enzymes. Sci Rep 8. https://doi.org/10.1038/S41598-018-35033-Y

Cadet, X.F., Gelly, J.C., van Noord, A., Cadet, F., Acevedo-Rocha, C.G., 2022. Learning Strategies in Protein Directed Evolution. Methods Mol Biol 2461, 225–275. https://doi.org/10.1007/978-1-0716-2152-3_15

Calvó-Tusell, C., Maria-Solano, M.A., Osuna, S., Feixas, F., 2022a. Time Evolution of the Millisecond Allosteric Activation of Imidazole Glycerol Phosphate Synthase. J Am Chem Soc 144, 7146–7159. https://doi.org/10.1021/JACS.1C12629/SUPPL_FILE/JA1C12629_SI_003.MP4

Calvó-Tusell, C., Maria-Solano, M.A., Osuna, S., Feixas, F., 2022b. Time Evolution of the Millisecond Allosteric Activation of Imidazole Glycerol Phosphate Synthase. J Am Chem Soc 144, 7146–7159. https://doi.org/10.1021/JACS.1C12629/SUPPL_FILE/JA1C12629_SI_003.MP4

Calzadiaz-Ramirez, L., Calvó-Tusell, C., Stoffel, G.M.M., Lindner, S.N., Osuna, S., Erb, T.J., Garcia-Borràs, M., Bar-Even, A., Acevedo-Rocha, C.G., 2020. In Vivo Selection for Formate Dehydrogenases with High Efficiency and Specificity toward NADP+. ACS Catal 10, 7512–7525. https://doi.org/10.1021/ACSCATAL.0C01487/ASSET/IMAGES/LARGE/CS0C01487_0008.JPEG

Campbell, E., Kaltenbach, M., Correy, G.J., Carr, P.D., Porebski, B.T., Livingstone, E.K., Afriat-Jurnou, L., Buckle, A.M., Weik, M., Hollfelder, F., Tokuriki, N., Jackson, C.J., 2016. The role of protein dynamics in the evolution of new enzyme function. Nat Chem Biol 12, 944–950. https://doi.org/10.1038/nchembio.2175

Campbell, E.C., Correy, G.J., Mabbitt, P.D., Buckle, A.M., Tokuriki, N., Jackson, C.J., 2018. Laboratory evolution of protein conformational dynamics. Curr Opin Struct Biol 50, 49–57. https://doi.org/10.1016/j.sbi.2017.09.005

Carlin, D.A., Caster, R.W., Wang, X., Betzenderfer, S.A., Chen, C.X., Duong, V.M., Ryklansky, C. V., Alpekin, A., Beaumont, N., Kapoor, H., Kim, N., Mohabbot, H., Pang, B., Teel, R., Whithaus, L., Tagkopoulos, I., Siegel, J.B., 2016. Kinetic Characterization of 100 Glycoside Hydrolase Mutants Enables the Discovery of Structural Features Correlated with Kinetic Constants. PLoS One 11, e0147596. https://doi.org/10.1371/JOURNAL.PONE.0147596

Casadevall, G., Casadevall, J., Duran, C., Osuna, S., 2024. The shortest path method (SPM) webserver for computational enzyme design. Protein Eng Des Sel 37, gzae005. https://doi.org/10.1093/protein/gzae005

Casadevall, G., Duran, C., Osuna, S., 2023. AlphaFold2 and Deep Learning for Elucidating Enzyme Conformational Flexibility and Its Application for Design. JACS Au 3, 1554–1562. https://doi.org/10.1021/jacsau.3c00188

Case, D.A., Aktulga, H.M., Belfon, K., Cerutti, D.S., Cisneros, G.A., Cruzeiro, V.W.D., Forouzesh, N., Giese, T.J., Götz, A.W., Gohlke, H., Izadi, S., Kasavajhala, K., Kaymak, M.C., King, E., Kurtzman, T., Lee, T.S., Li, P., Liu, J., Luchko, T., Luo, R., Manathunga, M., Machado, M.R., Nguyen, H.M., O'Hearn, K.A., Onufriev, A. V., Pan, F., Pantano, S., Qi, R., Rahnamoun, A., Risheh, A., Schott-Verdugo, S., Shajan, A., Swails, J., Wang, J., Wei, H., Wu, X., Wu, Y., Zhang, S., Zhao, S., Zhu, Q., Cheatham, T.E., Roe, D.R., Roitberg, A., Simmerling, C., York, D.M., Nagan, M.C., Merz, K.M., 2023. AmberTools. J Chem Inf Model 63, 6183–6191. https://doi.org/10.1021/ACS.JCIM.3C01153/ASSET/IMAGES/LARGE/CI3C01153_0002.JPEG

Castelli, M., Marchetti, F., Osuna, S., F. Oliveira, A.S., Mulholland, A.J., Serapian, S.A., Colombo, G., 2024. Decrypting Allostery in Membrane-Bound K-Ras4B Using Complementary In Silico Approaches Based on Unbiased Molecular Dynamics Simulations. J. Am. Chem. Soc. 146, 901–919. https://doi.org/10.1021/jacs.3c11396

Chen, D., Hartout, P., Pellizzoni, P., Oliver, C., Borgwardt, K., 2024. Endowing Protein Language Models with Structural Knowledge.

Cheng, J., Randall, A.Z., Sweredoski, M.J., Baldi, P., 2005. SCRATCH: a protein structure and structural feature prediction server. Nucleic Acids Res 33. https://doi.org/10.1093/NAR/GKI396

Cherry, J.R., Fidantsef, A.L., 2003. Directed evolution of industrial enzymes: an update. Curr Opin Biotechnol 14, 438–443. https://doi.org/10.1016/S0958-1669(03)00099-5

Cherry, J.R., Lamsa, M.H., Schneider, P., Vind, J., Svendsen, A., Jones, A., Pedersen, A.H., 1999. Directed evolution of a fungal peroxidase. Nature Biotechnology 1999 17:4 17, 379–384. https://doi.org/10.1038/7939

Chodera, J.D., Noé, F., 2014. Markov state models of biomolecular conformational dynamics. Curr Opin Struct Biol, Theory and simulation / Macromolecular machines 25, 135–144. https://doi.org/10.1016/j.sbi.2014.04.002

Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., De Hoon, M.J.L., 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25, 1422–1423. https://doi.org/10.1093/BIOINFORMATICS/BTP163

Corbella, M., Pinto, G.P., Kamerlin, S.C.L., 2023. Loop dynamics and the evolution of enzyme activity. Nat Rev Chem 7, 536–547. https://doi.org/10.1038/s41570-023-00495-w

Corso, G., Deng, A., Fry, B., Polizzi, N., Barzilay, R., Jaakkola, T., 2024. Deep Confident Steps to New Pockets: Strategies for Docking Generalization.

Crean, R.M., Biler, M., Van Der Kamp, M.W., Hengge, A.C., Kamerlin, S.C.L., 2021. Loop Dynamics and Enzyme Catalysis in Protein Tyrosine Phosphatases. J Am Chem Soc 143, 3830–3845. https://doi.org/10.1021/JACS.0C11806/ASSET/IMAGES/LARGE/JA0C11806_0009.JPEG

Curado-Carballada, C., Feixas, F., Osuna, S., 2019. Molecular Dynamics Simulations on Aspergillus niger Monoamine Oxidase: Conformational Dynamics and Inter-monomer Communication Essential for Its Efficient Catalysis. Adv Synth Catal 361, 2718–2726. https://doi.org/10.1002/ADSC.201900158

Davis, I.W., Baker, D., 2009. RosettaLigand docking with full ligand and receptor flexibility. J Mol Biol 385, 381–392. https://doi.org/10.1016/J.JMB.2008.11.010

Dawson, N.L., Lewis, T.E., Das, S., Lees, J.G., Lee, D., Ashford, P., Orengo, C.A., Sillitoe, I., 2019. CATH protein domain classification (version 4.2) [WWW Document]. University College London. https://doi.org/https://doi.org/10.5522/04/7937330.v1

Dawson, N.L., Lewis, T.E., Das, S., Lees, J.G., Lee, D., Ashford, P., Orengo, C.A., Sillitoe, I., 2017. CATH: An expanded resource to predict protein function through structure and sequence. Nucleic Acids Res 45, D289–D295. https://doi.org/10.1093/NAR/GKW1098

Desaphy, J., Raimbaud, E., Ducrot, P., Rognan, D., 2013. Encoding protein-ligand interaction patterns in fingerprints and graphs. J Chem Inf Model 53, 623–637. https://doi.org/10.1021/CI300566N/SUPPL_FILE/CI300566N_SI_001.PDF

Detlefsen, N.S., Hauberg, S., Boomsma, W., 2022. Learning meaningful representations of protein sequences. Nat Commun 13, 1914.

Devine, P.N., Howard, R.M., Kumar, R., Thompson, M.P., Truppo, M.D., Turner, N.J., 2018. Extending the application of biocatalysis to meet the challenges of drug development. Nature Reviews Chemistry 2018 2:12 2, 409–421. https://doi.org/10.1038/s41570-018-0055-1

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Ding, X., Zou, Z., Brooks III, C.L., 2019. Deciphering protein evolution and fitness landscapes with latent space models. Nat Commun 10, 5644.

d'Oelsnitz, S., Diaz, D.J., Kim, W., Acosta, D.J., Dangerfield, T.L., Schechter, M.W., Minus, M.B., Howard, J.R., Do, H., Loy, J.M., Alper, H.S., Zhang, Y.J., Ellington, A.D., 2024. Biosensor and machine learning-aided engineering of an amaryllidaceae enzyme. Nature Communications 2024 15:1 15, 1–14. https://doi.org/10.1038/s41467-024-46356-y

Eberhardt, J., Santos-Martins, D., Tillack, A.F., Forli, S., 2021. AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. J Chem Inf Model 61, 3891–3898. https://doi.org/10.1021/ACS.JCIM.1C00203/SUPPL_FILE/CI1C00203_SI_002.ZIP

Elabd, H., Bromberg, Y., Hoarfrost, A., Lenz, T., Franke, A., Wendorff, M., 2020. Amino acid encoding for deep learning applications. BMC Bioinformatics 21, 1–14. https://doi.org/10.1186/S12859-020-03546-X/FIGURES/4

Elia Venanzi, N.A., Basciu, A., Vargiu, A.V., Kiparissides, A., Dalby, P.A., Dikicioglu, D., 2024. Machine Learning Integrating Protein Structure, Sequence, and Dynamics to Predict the Enzyme Activity of Bovine Enterokinase Variants. J. Chem. Inf. Model. https://doi.org/10.1021/acs.jcim.3c00999

Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., others, 2021. Prottrans: Toward understanding the language of life through self-supervised learning. IEEE Trans Pattern Anal Mach Intell 44, 7112–7127.

Fasoulis, R., Paliouras, G., Kavraki, L.E., 2021. Graph representation learning for structural proteomics. Emerg Top Life Sci 5, 789. https://doi.org/10.1042/ETLS20210225

Feng, Y., Gong, C., Zhu, J., Liu, G., Tang, Y., Li, W., 2023. Prediction of Sites of Metabolism of CYP3A4 Substrates Utilizing Docking-Derived Geometric Features. J Chem Inf Model 63, 4158–4169. https://doi.org/10.1021/ACS.JCIM.3C00549/SUPPL_FILE/CI3C00549_SI_002.XLSX

Ferruz, N., Höcker, B., 2022. Controllable protein design with language models. Nature Machine Intelligence 2022 4:6 4, 521–532. https://doi.org/10.1038/s42256-022-00499-z

Ferruz, N., Schmidt, S., Höcker, B., 2022. ProtGPT2 is a deep unsupervised language model for protein design. Nat Commun 13, 4348.

Folkman, L., Stantic, B., Sattar, A., Zhou, Y., 2016. EASE-MM: Sequence-Based Prediction of Mutation-Induced Stability Changes with Feature-Based Multiple Models. J Mol Biol 428, 1394–1405. https://doi.org/10.1016/J.JMB.2016.01.012

Fox, R., 2005. Directed molecular evolution by machine learning and the influence of nonlinear interactions. J Theor Biol 234, 187–199. https://doi.org/10.1016/J.JTBI.2004.11.031

Fraczkiewicz, R., Braun, W., 1998. Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules. J Comput Chem 19, 319–333. https://doi.org/10.1002/(sici)1096-987x(199802)19:3<319::aid-jcc6>3.0.co

France, S.P., Hepworth, L.J., Turner, N.J., Flitsch, S.L., 2017. Constructing Biocatalytic Cascades: In Vitro and in Vivo Approaches to de Novo Multi-Enzyme Pathways. ACS Catal 7, 710–724. https://doi.org/10.1021/ACSCATAL.6B02979/ASSET/IMAGES/LARGE/CS-2016-02979U_0018.JPEG

Frazer, J., Notin, P., Dias, M., Gomez, A., Min, J.K., Brock, K., Gal, Y., Marks, D.S., 2021. Disease variant prediction with deep generative models of evolutionary data. Nature 599, 91–95.

Gainza, P., Sverrisson, F., Monti, F., Rodolà, E., Boscaini, D., Bronstein, M.M., Correia, B.E., 2019. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. Nature Methods 2019 17:2 17, 184–192. https://doi.org/10.1038/s41592-019-0666-6

Galanie, S., Entwistle, D., Lalonde, J., 2020. Engineering biosynthetic enzymes for industrial natural product synthesis. Nat Prod Rep 37, 1122–1143. https://doi.org/10.1039/C9NP00071B

Galdadas, I., Qu, S., Oliveira, A.S.F., Olehnovics, E., Mack, A.R., Mojica, M.F., Agarwal, P.K., Tooke, C.L., Gervasio, F.L., Spencer, J., Bonomo, R.A., Mulholland, A.J., Haider, S., 2021. Allosteric communication in class A β-lactamases occurs via cooperative coupling of loop dynamics. Elife 10.

Gandomkar, S., Żądło-Dobrowolska, A., Kroutil, W., 2019. Extending Designed Linear Biocatalytic Cascades for Organic Synthesis. ChemCatChem 11, 225–243. https://doi.org/10.1002/CCTC.201801063

Gergel, S., Soler, J., Klein, A., Schülke, K.H., Hauer, B., Garcia-Borràs, M., Hammer, S.C., 2023. Engineered cytochrome P450 for direct arylalkene-to-ketone oxidation via highly reactive carbocation intermediates. Nature Catalysis 2023 6:7 6, 606–617. https://doi.org/10.1038/s41929-023-00979-4

Ghorbani, M., Prasad, S., Klauda, J.B., Brooks, B.R., 2022. GraphVAMPNet, using graph neural networks and variational approach to Markov processes for dynamical modeling of biomolecules. J Chem Phys 156, 184103. https://doi.org/10.1063/5.0085607

Giessel, A., Dousis, A., Ravichandran, K., Smith, K., Sur, S., McFadyen, I., Zheng, W., Licht, S., 2022. Therapeutic enzyme engineering using a generative neural network. Sci Rep 12, 1536.

Giver, L., Gershenson, A., Freskgard, P.O., Arnold, F.H., 1998. Directed evolution of a thermostable esterase. Proceedings of the National Academy of Sciences 95, 12809–12813. https://doi.org/10.1073/PNAS.95.22.12809

Gligorijević, V., Renfrew, P.D., Kosciolek, T., Leman, J.K., Berenberg, D., Vatanen, T., Chandler, C., Taylor, B.C., Fisk, I.M., Vlamakis, H., Xavier, R.J., Knight, R., Cho, K., Bonneau, R., 2021. Structure-based protein function prediction using graph convolutional networks. Nature Communications 2021 12:1 12, 1–14. https://doi.org/10.1038/s41467-021-23303-9

Goblirsch, B.R., Jensen, M.R., Mohamed, F.A., Wackett, L.P., Wilmot, C.M., 2016. Substrate trapping in crystals of the thiolase olea identifies three channels that enable long chain olefin biosynthesis. Journal of Biological Chemistry 291, 26698–26706. https://doi.org/10.1074/JBC.M116.760892

Goldman, S., Das, R., Yang, K.K., Coley, C.W., 2022. Machine learning modeling of family wide enzyme-substrate specificity screens. PLoS Comput Biol 18, e1009853. https://doi.org/10.1371/JOURNAL.PCBI.1009853

Gordon, S.E., Weber, D.K., Downton, M.T., Wagner, J., Perugini, M.A., 2016. Dynamic Modelling Reveals 'Hotspots' on the Pathway to Enzyme-Substrate Complex Formation. PLoS Comput Biol 12, e1004811. https://doi.org/10.1371/journal.pcbi.1004811

Greenhalgh, J.C., Fahlberg, S.A., Pfleger, B.F., Romero, P.A., 2021. Machine learning-guided acyl-ACP reductase engineering for improved in vivo fatty alcohol production. Nature Communications 2021 12:1 12, 1–10. https://doi.org/10.1038/s41467-021-25831-w

Hamelryck, T., 2005. An amino acid has two sides: a new 2D measure provides a different view of solvent exposure. Proteins 59, 38–48. https://doi.org/10.1002/PROT.20379

Harding-Larsen, D., Madsen, C.D., Teze, D., Kittilä, T., Langhorn, M.R., Gharabli, H., Hobusch, M., Otalvaro, F.M., Kırtel, O., Bidart, G.N., Mazurenko, S., Travnik, E., Welner, D.H., 2023. GASP: A pan-specific predictor of family 1 glycosyltransferase specificity enabled by a pipeline for substrate feature generation and large-scale experimental screening. https://doi.org/10.26434/CHEMRXIV-2023-PR9CK

Hauer, B., 2020. Embracing Nature's Catalysts: A Viewpoint on the Future of Biocatalysis. ACS Catal 10, 8418–8427. https://doi.org/10.1021/ACSCATAL.0C01708/ASSET/IMAGES/LARGE/CS0C01708_0003.JPEG

Hawkins-Hooker, A., Depardieu, F., Baur, S., Couairon, G., Chen, A., Bikard, D., 2021. Generating functional protein variants with variational autoencoders. PLoS Comput Biol 17, e1008736.

Heath, R.S., Ruscoe, R.E., Turner, N.J., 2022. The beauty of biocatalysis: sustainable synthesis of ingredients in cosmetics. Nat Prod Rep 39, 335–388. https://doi.org/10.1039/D1NP00027F

Heckmann, D., Campeau, A., Lloyd, C.J., Phaneuf, P. V., Hefner, Y., Carrillo-Terrazas, M., Feist, A.M., Gonzalez, D.J., Palsson, B.O., 2020. Kinetic profiling of metabolic specialists demonstrates stability and consistency of in vivo enzyme turnover numbers. Proc Natl Acad Sci U S A 117, 23182–23190.

https://doi.org/10.1073/PNAS.2001562117/SUPPL_FILE/PNAS.2001562117.SD01.XLSX

Heckmann, D., Lloyd, C.J., Mih, N., Ha, Y., Zielinski, D.C., Haiman, Z.B., Desouki, A.A., Lercher, M.J., Palsson, B.O., 2018. Machine learning applied to enzyme turnover numbers reveals protein structural correlates and improves metabolic models. Nature Communications 2018 9:1 9, 1–10. https://doi.org/10.1038/s41467-018-07652-6

Heffernan, R., Yang, Y., Paliwal, K., Zhou, Y., 2017. Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility. Bioinformatics 33, 2842–2849. https://doi.org/10.1093/BIOINFORMATICS/BTX218

Heinzinger, Michael, Weissenow, Konstantin, Gomez Sanchez, Joaquin, Henkel, Adrian, Steinegger, Martin, Rost, B., Heinzinger, M, Weissenow, K, Gomez Sanchez, J, Henkel, A, Steinegger, M, Prostt5, R., 2023. ProstT5: Bilingual Language Model for Protein Sequence and Structure. bioRxiv 2023.07.23.550085. https://doi.org/10.1101/2023.07.23.550085

Hellberg, S., Sjöström, M., Skagerberg, B., Wold, S., 1987. Peptide Quantitative Structure-Activity Relationships, a Multivariate Approach. J Med Chem 30, 1126–1135. https://doi.org/10.1021/JM00390A003/SUPPL_FILE/JM00390A003_SI_001.PDF

Henikoff, S., Henikoff, J.G., 1992. Amino acid substitution matrices from protein blocks. Proceedings of the National Academy of Sciences 89, 10915–10919. https://doi.org/10.1073/PNAS.89.22.10915

Henzler-Wildman, K., Kern, D., 2007. Dynamic personalities of proteins. Nature 450.

Hoffbauer, T., Strodel, B., 2024. TransMEP: Transfer learning on large protein language models to predict mutation effects of proteins from a small known dataset. bioRxiv 2021–2024.

Hon, J., Borko, S., Stourac, J., Prokop, Z., Zendulka, J., Bednar, D., Martinek, T., Damborsky, J., 2020. EnzymeMiner: automated mining of soluble enzymes with diverse structures, catalytic properties and stabilities. Nucleic Acids Res 48, W104–W109. https://doi.org/10.1093/NAR/GKAA372

Hou, X., Wang, Yu, Bu, D., Wang, Yaojun, Sun, S., 2023. EMNGly: predicting N-linked glycosylation sites using the language models for feature extraction. Bioinformatics 39. https://doi.org/10.1093/BIOINFORMATICS/BTAD650

Hsu, C., Nisonoff, H., Fannjiang, C., Listgarten, J., 2022. Learning protein fitness models from evolutionary and assay-labeled data. Nat Biotechnol 40, 1114–1122. https://doi.org/10.1038/S41587-021-01146-5

Huang, T.W., Zaretzki, J., Bergeron, C., Bennett, K.P., Breneman, C.M., 2013. DR-Predictor: Incorporating flexible docking with specialized electronic reactivity and machine learning techniques to predict CYP-mediated sites of metabolism. J Chem Inf Model 53, 3352–3366. https://doi.org/10.1021/CI4004688/SUPPL_FILE/CI4004688_SI_001.ZIP

Huffman, M.A., Fryszkowska, A., Alvizo, O., Borra-Garske, M., Campos, K.R., Canada, K.A., Devine, P.N., Duan, D., Forstater, J.H., Grosser, S.T., Halsey, H.M., Hughes, G.J., Jo, J., Joyce, L.A., Kolev, J.N., Liang, J., Maloney, K.M., Mann, B.F., Marshall, N.M., McLaughlin, M., Moore, J.C., Murphy, G.S., Nawrat, C.C., Nazor, J., Novick, S., Patel, N.R., Rodriguez-Granillo, A., Robaire, S.A., Sherer, E.C., Truppo, M.D., Whittaker, A.M., Verma, D., Xiao, L., Xu, Y., Yang, H., 2019. Design of an in vitro biocatalytic cascade for the manufacture of islatravir. Science (1979) 366, 1255–1259. https://doi.org/10.1126/SCIENCE.AAY8484/SUPPL_FILE/AAY8484-HUFFMAN-SM.PDF

Ibtehaz, N., Kagaya, Y., Kihara, D., 2023. Domain-PFP allows protein function prediction using function-aware domain embedding representations. Communications Biology 2023 6:1 6, 1–14. https://doi.org/10.1038/s42003-023-05476-9

Iqbal, S., Ge, F., Li, F., Akutsu, T., Zheng, Y., Gasser, R.B., Yu, D.J., Webb, G.I., Song, J., 2022. PROST: AlphaFold2-aware Sequence-Based Predictor to Estimate Protein Stability Changes upon Missense Mutations. J Chem Inf Model. https://doi.org/10.1021/ACS.JCIM.2C00799/SUPPL_FILE/CI2C00799_SI_001.PDF

Isert, C., Atz, K., Schneider, G., 2023. Structure-based drug design with geometric deep learning. Curr Opin Struct Biol 79, 102548. https://doi.org/10.1016/J.SBI.2023.102548

Iuchi, H., Matsutani, T., Yamada, K., Iwano, N., Sumi, S., Hosoda, S., Zhao, S., Fukunaga, T., Hamada, M., 2021. Representation learning applications in biological sequence analysis. Comput Struct Biotechnol J 19, 3198–3208.

Jing, B., Berger, B., Jaakkola, T., 2024. AlphaFold Meets Flow Matching for Generating Protein Ensembles.

Jonsson, J., Eriksson, L., Hellberg, S., Sjöström, M., Wold, S., 1989. Multivariate Parametrization of 55 Coded and Non-Coded Amino Acids. Quantitative Structure-Activity Relationships 8, 204–209. https://doi.org/10.1002/QSAR.19890080303

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S.A.A., Ballard, A.J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A.W., Kavukcuoglu, K., Kohli, P., Hassabis, D., 2021. Highly accurate protein structure prediction with AlphaFold. Nature 2021 596:7873 596, 583–589. https://doi.org/10.1038/s41586-021-03819-2

Kamerlin, S.C.L., Warshel, A., 2010. At the dawn of the 21st century: Is dynamics the missing link for understanding enzyme catalysis? Proteins: Structure, Function, and Bioinformatics 78, 1339–1375. https://doi.org/10.1002/prot.22654

Kanakala, G.C., Aggarwal, R., Nayar, D., Priyakumar, U.D., 2022. Latent Biases in Machine Learning Models for Predicting Binding Affinities Using Popular Data Sets. ACS Omega. https://doi.org/10.1021/ACSOMEGA.2C06781/ASSET/IMAGES/LARGE/AO2C06781_0004.JPEG

Karlov, D.S., Long, S.L., Zeng, X., Xu, F., Lal, K., Cao, L., Hayoun, K., Lin, J., Joyce, S.A., Tikhonova, I.G., 2023. Characterization of the mechanism of bile salt hydrolase substrate specificity by experimental and computational analyses. Structure 31, 629-638.e5. https://doi.org/10.1016/J.STR.2023.02.014

Kawashima, S., Kanehisa, M., 2000. AAindex: Amino Acid index database. Nucleic Acids Res 28, 374–374. https://doi.org/10.1093/NAR/28.1.374

Kazan, I.C., Mills, J.H., Ozkan, S.B., 2023. Allosteric regulatory control in dihydrofolate reductase is revealed by dynamic asymmetry. Protein Science 32, e4700. https://doi.org/10.1002/pro.4700

Khan, N.R., Rathod, V.K., 2015. Enzyme catalyzed synthesis of cosmetic esters and its intensification: A review. Process Biochemistry 50, 1793–1806. https://doi.org/10.1016/J.PROCBIO.2015.07.014

Kim, A.K., Porter, L.L., 2021. Functional and Regulatory Roles of Fold-Switching Proteins. Structure 29, 6–14. https://doi.org/10.1016/J.STR.2020.10.006

Kingma, D.P., Welling, M., 2013. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.

Kirk, O., Borchert, T.V., Fuglsang, C.C., 2002. Industrial enzyme applications. Curr Opin Biotechnol 13, 345–351. https://doi.org/10.1016/S0958-1669(02)00328-2

Kohout, P., Vasina, M., Majerova, M., Novakova, V., Damborsky, J., Bednar, D., Marek, M., Prokop, Z., Mazurenko, S., 2023. Design of Enzymes for Biocatalysis, Bioremediation, and Biosensing using Variational Autoencoder-Generated Latent Spaces.

Konovalov, K.A., Unarta, I.C., Cao, S., Goonetilleke, E.C., Huang, X., 2021. Markov State Models to Study the Functional Dynamics of Proteins in the Wake of Machine Learning. JACS Au 1, 1330–1341. https://doi.org/10.1021/jacsau.1c00254

Kouba, P., Kohout, P., Haddadi, F., Bushuiev, A., Samusevich, R., Sedlar, J., Damborsky, J., Pluskal, T., Sivic, J., Mazurenko, S., 2023. Machine Learning-Guided Protein Engineering. ACS Catal 13, 13863–13895. https://doi.org/10.1021/ACSCATAL.3C02743

Kroll, A., Lercher, M.J., 2023. Machine learning models for the prediction of enzyme properties should be tested on proteins not used for model training. bioRxiv 2023.02.06.526991. https://doi.org/10.1101/2023.02.06.526991

Kroll, A., Ranjan, S., Engqvist, M.K.M., Lercher, M.J., 2023a. A general model to predict small molecule substrates of enzymes based on machine and deep learning. Nature Communications 2023 14:1 14, 1–13. https://doi.org/10.1038/s41467-023-38347-2

Kroll, A., Rousset, Y., Hu, X.P., Liebrand, N.A., Lercher, M.J., 2023b. Turnover number predictions for kinetically uncharacterized enzymes using machine and deep learning. Nature Communications 2023 14:1 14, 1–14. https://doi.org/10.1038/s41467-023-39840-4

Kunka, A., Marques, S.M., Havlasek, M., Vasina, M., Velatova, N., Cengelova, L., Kovar, D., Damborsky, J., Marek, M., Bednar, D., Prokop, Z., 2023. Advancing Enzyme's Stability and Catalytic Efficiency through Synergy of Force-Field Calculations, Evolutionary Analysis, and Machine Learning. ACS Catal 13, 12506–12518. https://doi.org/10.1021/ACSCATAL.3C02575/SUPPL_FILE/CS3C02575_SI_005.XLSX

Lane, T.J., 2023. Protein structure prediction has reached the single-structure frontier. Nat Methods 20, 170–173. https://doi.org/10.1038/s41592-022-01760-4

Le Guilloux, V., Schmidtke, P., Tuffery, P., 2009. Fpocket: An open source platform for ligand pocket detection. BMC Bioinformatics 10, 1–11. https://doi.org/10.1186/1471-2105-10-168/TABLES/1

Lee, B., Richards, F.M., 1971. The interpretation of protein structures: Estimation of static accessibility. J Mol Biol 55, 379-IN4. https://doi.org/10.1016/0022-2836(71)90324-X

Leidner, F., Kurt Yilmaz, N., Schiffer, C.A., 2019. Target-Specific Prediction of Ligand Affinity with Structure-Based Interaction Fingerprints. J Chem Inf Model 59, 3679–3691. https://doi.org/10.1021/ACS.JCIM.9B00457/SUPPL_FILE/CI9B00457_SI_001.PDF

Li, B., Yang, Y.T., Capra, J.A., Gerstein, M.B., 2020. Predicting changes in protein thermodynamic stability upon point mutation with deep 3D convolutional neural networks. PLoS Comput Biol 16. https://doi.org/10.1371/JOURNAL.PCBI.1008291

Li, F., Yuan, L., Lu, H., Li, G., Chen, Y., Engqvist, M.K.M., Kerkhoven, E.J., Nielsen, J., 2022. Deep learning-based kcat prediction enables improved enzyme-constrained model reconstruction. Nature Catalysis 2022 5:8 5, 662–672. https://doi.org/10.1038/s41929-022-00798-z

Li, G., Qin, Y., Fontaine, N.T., Ng Fuk Chong, M., Maria-Solano, M.A., Feixas, F., Cadet, X.F., Pandjaitan, R., Garcia-Borràs, M., Cadet, F., Reetz, M.T., 2021. Machine Learning Enables Selection of Epistatic Enzyme Mutants for Stability Against Unfolding and

Detrimental Aggregation. Chembiochem 22, 904–914. https://doi.org/10.1002/CBIC.202000612

Li, J., Guan, X., Zhang, O., Sun, K., Wang, Y., Bagni, D., Head-Gordon, T., Pitzer, †, 2023. Leak Proof PDBBind: A Reorganized Dataset of Protein-Ligand Complexes for More Generalizable Binding Affinity Prediction. ArXiv.

Li, M., Wang, H., Yang, Z., Zhang, L., Zhu, Y., 2023. DeepTM: A deep learning algorithm for prediction of melting temperature of thermophilic proteins directly from sequences. Comput Struct Biotechnol J 21, 5544–5560. https://doi.org/10.1016/j.csbj.2023.11.006

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., Rives, A., 2023. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science (1979) 379, 1123–1130. https://doi.org/10.1126/SCIENCE.ADE2574/SUPPL_FILE/SCIENCE.ADE2574_SM.PDF

Livesey, B.J., Marsh, J.A., 2023. Updated benchmarking of variant effect predictors using deep mutational scanning. Mol Syst Biol e11474.

Lu, C., Lubin, J.H., Sarma, V. V., Stentz, S.Z., Wang, G., Wang, S., Khare, S.D., 2023. Prediction and design of protease enzyme specificity using a structure-aware graph convolutional network. Proc Natl Acad Sci U S A 120, e2303590120. https://doi.org/10.1073/PNAS.2303590120/SUPPL_FILE/PNAS.2303590120.SD07.XLSX

Ma, E.J., Siirola, E., Moore, C., Kummer, A., Stoeckli, M., Faller, M., Bouquet, C., Eggimann, F., Ligibel, M., Huynh, D., Cutler, G., Siegrist, L., Lewis, R.A., Acker, A.C., Freund, E., Koch, E., Vogel, M., Schlingensiepen, H., Oakeley, E.J., Snajdrova, R., 2021. Machine-Directed Evolution of an Imine Reductase for Activity and Stereoselectivity. ACS Catal 11, 12433–12445. https://doi.org/10.1021/ACSCATAL.1C02786/SUPPL_FILE/CS1C02786_SI_003.CSV

Mansoor, S., Baek, M., Park, H., Lee, G.R., Baker, D., 2023. Protein Ensemble Generation through Variational Autoencoder Latent Space Sampling. bioRxiv. https://doi.org/10.1101/2023.08.01.551540

Mardt, A., Pasquali, L., Wu, H., Noé, F., 2018. VAMPnets for deep learning of molecular kinetics. Nat Commun 9, 5. https://doi.org/10.1038/s41467-017-02388-1

Maria-Solano, M.A., Kinateder, T., Iglesias-Fernández, J., Sterner, R., Osuna, S., 2021. In Silico identification and experimental validation of distal activity-enhancing mutations in tryptophan synthase. ACS Catal 11, 13733–13743. https://doi.org/10.1021/ACSCATAL.1C03950/SUPPL_FILE/CS1C03950_SI_001.PDF

Markus, B., Christian C, G., Andreas, K., Arkadij, K., Stefan, L., Gustav, O., Elina, S., Radka, S., 2023. Accelerating Biocatalysis Discovery with Machine Learning: A Paradigm Shift in Enzyme Engineering, Discovery, and Design. ACS Catal 13, 14454–14469. https://doi.org/10.1021/ACSCATAL.3C03417

Mastropietro, A., Pasculli, G., Bajorath, J., 2023. Learning characteristics of graph neural networks predicting protein–ligand affinities. Nature Machine Intelligence 2023 5:12 5, 1427–1436. https://doi.org/10.1038/s42256-023-00756-9

Mazurenko, S., Prokop, Z., Damborsky, J., 2020. Machine Learning in Enzyme Engineering. ACS Catal 10, 1210–1223. https://doi.org/10.1021/ACSCATAL.9B04321/ASSET/IMAGES/LARGE/CS9B04321_0004.JPEG

McGibbon, R.T., Beauchamp, K.A., Harrigan, M.P., Klein, C., Swails, J.M., Hernández, C.X., Schwantes, C.R., Wang, L.P., Lane, T.J., Pande, V.S., 2015. MDTraj: A Modern

Open Library for the Analysis of Molecular Dynamics Trajectories. Biophys J 109, 1528. https://doi.org/10.1016/J.BPJ.2015.08.015

Mei, H., Liao, Z.H., Zhou, Y., Li, S.Z., 2005. A new set of amino acid descriptors and its application in peptide QSARs. Peptide Science 80, 775–786. https://doi.org/10.1002/BIP.20296

Meiler, J., Baker, D., 2006. ROSETTALIGAND: Protein–small molecule docking with full side-chain flexibility. Proteins: Structure, Function, and Bioinformatics 65, 538–548. https://doi.org/10.1002/PROT.21086

Meiler, J., Müller, M., Zeidler, A., Schmäschke, F., 2001. Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks. J Mol Model 7, 360–369. https://doi.org/10.1007/S008940100038/METRICS

Michael, R., Kæstel-Hansen, J., Groth, P.M., Bartels, S., Salomon, J., Tian, P., Hatzakis, N.S., Boomsma, W.K., 2023. Assessing the performance of protein regression models. bioRxiv 2023.06.18.545472. https://doi.org/10.1101/2023.06.18.545472

Mohanan, N., Montazer, Z., Sharma, P.K., Levin, D.B., 2020. Microbial and Enzymatic Degradation of Synthetic Plastics. Front Microbiol 11, 580709. https://doi.org/10.3389/FMICB.2020.580709/BIBTEX

Morra, G., Potestio, R., Micheletti, C., Colombo, G., 2012. Corresponding Functional Dynamics across the Hsp90 Chaperone Family: Insights from a Multiscale Analysis of MD Simulations. PLoS Comput Biol 8, e1002433. https://doi.org/10.1371/JOURNAL.PCBI.1002433

Mou, Z., Eakes, J., Cooper, C.J., Foster, C.M., Standaert, R.F., Podar, M., Doktycz, M.J., Parks, J.M., 2021. Machine learning-based prediction of enzyme substrate scope: Application to bacterial nitrilases. Proteins: Structure, Function, and Bioinformatics 89, 336–347. https://doi.org/10.1002/PROT.26019

Mount, D.W., 2008. Using BLOSUM in sequence alignments. Cold Spring Harb Protoc 3. https://doi.org/10.1101/PDB.TOP39

Nazor, J., Liu, J., Huisman, G., 2021. Enzyme evolution for industrial biocatalytic cascades. Curr Opin Biotechnol 69, 182–190. https://doi.org/10.1016/J.COPBIO.2020.12.013

Noé, F., Tkatchenko, A., Müller, K.-R., Clementi, C., 2020. Machine Learning for Molecular Simulation. Annu Rev Phys Chem 71, 361–390. https://doi.org/10.1146/annurev-physchem-042018-052331

Notin, P., Kollasch, A.W., Ritter, D., Niekerk, L. Van, Paul, S., Spinner, H., Rollins, N.J., Shaw, A., Weitzman, R., Frazer, J., Dias, M., Franceschi, D., Orenbuch, R., Gal, Y., Marks, D.S., 2023. ProteinGym: Large-Scale Benchmarks for Protein Fitness Prediction and Design.

Novakovsky, G., Dexter, N., Libbrecht, M.W., Wasserman, W.W., Mostafavi, S., 2022. Obtaining genetics insights from deep learning via explainable artificial intelligence. Nature Reviews Genetics 2022 24:2 24, 125–137. https://doi.org/10.1038/s41576-022-00532-2

Oberg, N., Zallot, R., Gerlt, J.A., 2023. EFI-EST, EFI-GNT, and EFI-CGFP: Enzyme Function Initiative (EFI) Web Resource for Genomic Enzymology Tools. J Mol Biol 435, 168018. https://doi.org/10.1016/J.JMB.2023.168018

Oliveira, A.S.F., Ciccotti, G., Haider, S., Mulholland, A.J., 2021. Dynamical nonequilibrium molecular dynamics reveals the structural basis for allostery and signal propagation in biomolecular systems. Eur. Phys. J. B 94.

Osuna, S., 2021. The challenge of predicting distal active site mutations in computational enzyme design. WIREs Computational Molecular Science 11. https://doi.org/10.1002/wcms.1502

Paik, I., Ngo, P.H.T., Shroff, R., Diaz, D.J., Maranhao, A.C., Walker, D.J.F., Bhadra, S., Ellington, A.D., 2023. Improved Bst DNA Polymerase Variants Derived via a Machine Learning Approach. Biochemistry 62, 410–418. https://doi.org/10.1021/ACS.BIOCHEM.1C00451/ASSET/IMAGES/LARGE/BI1C00451_0006.JPEG

Qiu, Y., Wei, G.W., 2023. Artificial intelligence-aided protein engineering: from topological data analysis to deep protein language models. Brief Bioinform 24, 1–13. https://doi.org/10.1093/BIB/BBAD289

Qu, G., Li, A., Acevedo-Rocha, C.G., Sun, Z., Reetz, M.T., 2020. The Crucial Role of Methodology Development in Directed Evolution of Selective Enzymes. Angewandte Chemie International Edition 59, 13204–13231. https://doi.org/10.1002/ANIE.201901491

Radley, E., Davidson, J., Foster, J., Obexer, R., Bell, E.L., Green, A.P., 2023. Engineering Enzymes for Environmental Sustainability. Angewandte Chemie International Edition 62, e202309305. https://doi.org/10.1002/ANIE.202309305

Raimondi, D., Orlando, G., Vranken, W.F., Moreau, Y., 2019. Exploring the limitations of biophysical propensity scales coupled with machine learning for protein sequence analysis. Scientific Reports 2019 9:1 9, 1–11. https://doi.org/10.1038/s41598-019-53324-w

Ran, X., Jiang, Y., Shao, Q., Yang, Z.J., 2023. EnzyKR: a chirality-aware deep learning model for predicting the outcomes of the hydrolase-catalyzed kinetic resolution. Chem Sci 14, 12073–12082. https://doi.org/10.1039/D3SC02752J

Rao, R., Meier, J., Sercu, T., Ovchinnikov, S., Rives, A., 2020. Transformer protein language models are unsupervised structure learners. Biorxiv 2012–2020.

Rao, R.M., Liu, J., Verkuil, R., Meier, J., Canny, J., Abbeel, P., Sercu, T., Rives, A., 2021. MSA Transformer. PMLR, pp. 8844–8856.

Raudys, S.J., Jain, A.K., 1991. Small Sample Size Effects in Statistical Pattern Recognition: Recommendations for Practitioners. IEEE Trans Pattern Anal Mach Intell 13, 252–264. https://doi.org/10.1109/34.75512

Reetz, M.T., Qu, G., Sun, Z., 2024. Engineered enzymes for the synthesis of pharmaceuticals and other high-value products. Nature Synthesis 2024 3:1 3, 19–32. https://doi.org/10.1038/s44160-023-00417-0

Renata, H., Wang, Z.J., Arnold, F.H., 2015. Expanding the Enzyme Universe: Accessing Non-Natural Reactions by Mechanism-Guided Directed Evolution. Angewandte Chemie International Edition 54, 3351–3367. https://doi.org/10.1002/ANIE.201409470

Richards, F.M., 1977. Areas, volumes, packing and protein structure. Annu Rev Biophys Bioeng 6, 151–176. https://doi.org/10.1146/ANNUREV.BB.06.060177.001055

Riesselman, A.J., Ingraham, J.B., Marks, D.S., 2018. Deep generative models of genetic variation capture the effects of mutations. Nature Methods 2018 15:10 15, 816–822. https://doi.org/10.1038/s41592-018-0138-4

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C.L., Ma, J., Fergus, R., 2021. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc Natl Acad Sci U S A 118, e2016239118. https://doi.org/10.1073/PNAS.2016239118/SUPPL_FILE/PNAS.2016239118.SAPP.PDF

Robinson, S.L., Smith, M.D., Richman, J.E., Aukema, K.G., Wackett, L.P., 2020. Machine learning-based prediction of activity and substrate specificity for OleA enzymes in the thiolase superfamily. Synth Biol 5. https://doi.org/10.1093/SYNBIO/YSAA004

Romero-Rivera, A., Corbella, M., Parracino, A., Patrick, W.M., Kamerlin, S.C.L., 2022. Complex Loop Dynamics Underpin Activity, Specificity, and Evolvability in the (βα) _8

Barrel Enzymes of Histidine and Tryptophan Biosynthesis. JACS Au 2, 943–960. https://doi.org/10.1021/jacsau.2c00063

Romero-Rivera, A., Garcia-Borràs, M., Osuna, S., 2017. Role of Conformational Dynamics in the Evolution of Retro-Aldolase Activity. ACS Catal 7, 8524–8532. https://doi.org/10.1021/acscatal.7b02954

Röttig, M., Rausch, C., Kohlbacher, O., 2010. Combining Structure and Sequence Information Allows Automated Prediction of Substrate Specificities within Enzyme Families. PLoS Comput Biol 6, e1000636. https://doi.org/10.1371/JOURNAL.PCBI.1000636

Ruiz-Blanco, Y.B., Paz, W., Green, J., Marrero-Ponce, Y., 2015. ProtDCal: A program to compute general-purpose-numerical descriptors for sequences and 3D-structures of proteins. BMC Bioinformatics 16, 1–15. https://doi.org/10.1186/S12859-015-0586-0/TABLES/4

Saito, Y., Oikawa, M., Sato, T., Nakazawa, H., Ito, T., Kameda, T., Tsuda, K., Umetsu, M., 2021. Machine-Learning-Guided Library Design Cycle for Directed Evolution of Enzymes: The Effects of Training Data Composition on Sequence Space Exploration. ACS Catal 11, 14615–14624. https://doi.org/10.1021/ACSCATAL.1C03753/SUPPL_FILE/CS1C03753_SI_007.XLSX

Sala, D., Engelberger, F., Mchaourab, H.S., Meiler, J., 2023. Modeling conformational states of proteins with AlphaFold. Curr Opin Struct Biol 81, 102645. https://doi.org/10.1016/j.sbi.2023.102645

Sandberg, M., Eriksson, L., Jonsson, J., Sjöström, M., Wold, S., 1998. New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. J Med Chem 41, 2481–2491. https://doi.org/10.1021/JM9700575/SUPPL_FILE/JM2481.PDF

Sanner, M., Olson, A., Spehner, J., 1996. Reduced surface: an efficient way to compute molecular surfaces. Biopolymers. https://doi.org/10.1002/(SICI)1097-0282(199603)38:3

Santacoloma, P.A., Sin, G., Gernaey, K. V., Woodley, J.M., 2011. Multienzyme-catalyzed processes: Next-generation biocatalysis. Org Process Res Dev 15, 203–212. https://doi.org/10.1021/OP1002159/ASSET/IMAGES/MEDIUM/OP-2010-002159_0011.GIF

Savile, C.K., Janey, J.M., Mundorff, E.C., Moore, J.C., Tam, S., Jarvis, W.R., Colbeck, J.C., Krebber, A., Fleitz, F.J., Brands, J., Devine, P.N., Huisman, G.W., Hughes, G.J., 2010. Biocatalytic asymmetric synthesis of chiral amines from ketones applied to sitagliptin manufacture. Science (1979) 329, 305–309. https://doi.org/10.1126/SCIENCE.1188934/SUPPL_FILE/SAVILE.SOM.PDF

Schenkmayerova, A., Pinto, G.P., Toul, M., Marek, M., Hernychova, L., Planas-Iglesias, J., Daniel Liskova, V., Pluskal, D., Vasina, M., Emond, S., Dörr, M., Chaloupkova, R., Bednar, D., Prokop, Z., Hollfelder, F., Bornscheuer, U.T., Damborsky, J., 2021. Engineering the protein dynamics of an ancestral luciferase. Nature Communications 2021 12:1 12, 1–16. https://doi.org/10.1038/s41467-021-23450-z

Schultze, S., Grubmüller, H., 2021. Time-Lagged Independent Component Analysis of Random Walks and Protein Dynamics. J. Chem. Theory Comput. 17, 5766–5776. https://doi.org/10.1021/acs.jctc.1c00273

Schweke, H., Mucchielli, M.H., Chevrollier, N., Gosset, S., Lopes, A., 2022. SURFMAP: A Software for Mapping in Two Dimensions Protein Surface Features. J Chem Inf Model 62, 1595–1601. https://doi.org/10.1021/ACS.JCIM.1C01269/ASSET/IMAGES/LARGE/CI1C01269_0004.JPEG

Sevgen, E., Moller, J., Lange, A., Parker, J., Quigley, S., Mayer, J., Srivastava, P., Gayatri, S., Hosfield, D., Korshunova, M., others, 2023. ProT-VAE: Protein Transformer Variational AutoEncoder for Functional Protein Design. bioRxiv 2021–2023.

Sheldon, R.A., Woodley, J.M., 2018. Role of Biocatalysis in Sustainable Chemistry. Chem Rev 118, 801–838. https://doi.org/10.1021/ACS.CHEMREV.7B00203/ASSET/IMAGES/LARGE/CR-2017-002034_0025.JPEG

Shroff, R., Cole, A.W., Diaz, D.J., Morrow, B.R., Donnell, I., Annapareddy, A., Gollihar, J., Ellington, A.D., Thyer, R., 2020. Discovery of novel gain-of-function mutations guided by structure-based deep learning. ACS Synth Biol 9, 2927–2935. https://doi.org/10.1021/ACSSYNBIO.0C00345/ASSET/IMAGES/LARGE/SB0C00345_0003.JPEG

Sinai, S., Kelsic, E.D., 2020. A primer on model-guided exploration of fitness landscapes for biological sequence design. arXiv preprint arXiv:2010.10614.

Sledzieski, S., Devkota, K., Singh, R., Cowen, L., Berger, B., 2023. TT3D: Leveraging precomputed protein 3D sequence models to predict protein–protein interactions. Bioinformatics 39. https://doi.org/10.1093/BIOINFORMATICS/BTAD663

Somnath, V.R., Bunne, C., Krause, A., 2021. Multi-Scale Representation Learning on Proteins. Adv Neural Inf Process Syst 34, 25244–25255.

Song, J., Tan, H., Takemoto, K., Akutsu, T., 2008. HSEpred: predict half-sphere exposure from protein sequences. Bioinformatics 24, 1489–1497. https://doi.org/10.1093/BIOINFORMATICS/BTN222

Sperl, J.M., Sieber, V., 2018. Multienzyme Cascade Reactions - Status and Recent Advances. ACS Catal 8, 2385–2396. https://doi.org/10.1021/ACSCATAL.7B03440/ASSET/IMAGES/MEDIUM/CS-2017-03440Y_0021.GIF

Steinegger, M., Söding, J., 2017. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nature Biotechnology 2017 35:11 35, 1026–1028. https://doi.org/10.1038/nbt.3988

Stimple, S.D., Smith, M.D., Tessier, P.M., 2020. Directed evolution methods for overcoming trade-offs between protein activity and stability. AIChE J 66. https://doi.org/10.1002/AIC.16814

St-Jacques, A.D., Rodriguez, J.M., Eason, M.G., Foster, S.M., Khan, S.T., Damry, A.M., Goto, N.K., Thompson, M.C., Chica, R.A., 2023. Computational remodeling of an enzyme conformational landscape for altered substrate selectivity. Nat Commun 14. https://doi.org/10.1038/s41467-023-41762-0

Su, J., Han, C., Zhou, Y., Shan, J., Zhou, X., Yuan, F., 2023. SaProt: Protein Language Modeling with Structure-aware Vocabulary. bioRxiv 2023.10.01.560349. https://doi.org/10.1101/2023.10.01.560349

Taujale, R., Venkat, A., Huang, L.C., Zhou, Z., Yeung, W., Rasheed, K.M., Li, S., Edison, A.S., Moremen, K.W., Kannan, N., 2020. Deep evolutionary analysis reveals the design principles of fold a glycosyltransferases. Elife 9. https://doi.org/10.7554/ELIFE.54532

Teng, S., Srivastava, A.K., Wang, L., 2010. Sequence feature-based prediction of protein stability changes upon amino acid substitutions. BMC Genomics 11, 1–8. https://doi.org/10.1186/1471-2164-11-S2-S5/FIGURES/4

Theodoridis, S., Koutroumbas, K., 2008. Pattern Recognition, Fourth Edition. Pattern Recognition, Fourth Edition 1–961. https://doi.org/10.1016/B978-1-59749-272-0.X0001-2

Thumuluri, V., Almagro Armenteros, J.J., Johansen, A.R., Nielsen, H., Winther, O., 2022. DeepLoc 2.0: multi-label subcellular localization prediction using protein language models. Nucleic Acids Res 50, W228–W234.

Tian, J., Dong, X., Wu, T., Wen, P., Liu, X., Zhang, M., An, X., Shi, D., 2024. Revealing the conformational dynamics of UDP-GlcNAc recognition by O-GlcNAc transferase via Markov state model. Int J Biol Macromol 256, 128405. https://doi.org/10.1016/j.ijbiomac.2023.128405

Tokuriki, N., Jackson, C.J., Afriat-Jurnou, L., Wyganowski, K.T., Tang, R., Tawfik, D.S., 2012. Diminishing returns and tradeoffs constrain the laboratory optimization of an enzyme. Nature Communications 2012 3:1 3, 1–10. https://doi.org/10.1038/ncomms2246

Torng, W., Altman, R.B., 2017. 3D deep convolutional neural networks for amino acid environment similarity analysis. BMC Bioinformatics 18, 1–23. https://doi.org/10.1186/S12859-017-1702-0/FIGURES/8

Trott, O., Olson, A.J., 2010. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem 31, 455–461. https://doi.org/10.1002/JCC.21334

Tschannen, M., Bachem, O., Lucic, M., 2018. Recent advances in autoencoder-based representation learning. arXiv preprint arXiv:1812.05069.

Turner, N.J., 2009. Directed evolution drives the next generation of biocatalysts. Nature Chemical Biology 2009 5:8 5, 567–573. https://doi.org/10.1038/nchembio.203

van Kempen, M., Kim, S.S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C.L.M., Söding, J., Steinegger, M., 2023. Fast and accurate protein structure search with Foldseek. Nature Biotechnology 2023 42:2 42, 243–246. https://doi.org/10.1038/s41587-023-01773-0

Vani, B.P., Aranganathan, A., Wang, D., Tiwary, P., 2023. AlphaFold2-RAVE: From Sequence to Boltzmann Ranking. J. Chem. Theory Comput. 19, 4351–4354. https://doi.org/10.1021/acs.jctc.3c00290

Vasina, M., Vanacek, P., Hon, J., Kovar, D., Faldynova, H., Kunka, A., Buryska, T., Badenhorst, C.P.S., Mazurenko, S., Bednar, D., Stavrakis, S., Bornscheuer, U.T., deMello, A., Damborsky, J., Prokop, Z., 2022. Advanced database mining of efficient haloalkane dehalogenases by sequence and structure bioinformatics and microfluidics. Chem Catalysis 2, 2704–2725. https://doi.org/10.1016/J.CHECAT.2022.09.011

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv Neural Inf Process Syst 30.

Verkuil, R., Kabeli, O., Du, Y., Wicky, B.I.M., Milles, L.F., Dauparas, J., Baker, D., Ovchinnikov, S., Sercu, T., Rives, A., 2022. Language models generalize beyond natural proteins. bioRxiv 2012–2022.

Vilone, G., Longo, L., 2020. Explainable Artificial Intelligence: a Systematic Review.

Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.-A., 2008. Extracting and composing robust features with denoising autoencoders, in: Proceedings of the 25th International Conference on Machine Learning. pp. 1096–1103.

Waksman, T., Astin, E., Fisher, S.R., Hunter, W., Bos, J., 2024. Computational prediction of structure, function and interaction of Myzus persicae (green peach aphid) salivary effector proteins. Mol Plant Microbe Interact. https://doi.org/10.1094/MPMI-10-23-0154-FI

Wallach, I., Heifets, A., 2018. Most Ligand-Based Classification Benchmarks Reward Memorization Rather than Generalization. J Chem Inf Model 58, 916–932. https://doi.org/10.1021/ACS.JCIM.7B00403/SUPPL_FILE/CI7B00403_SI_002.PDF

Wang, K., Zhou, R., Tang, J., Li, M., 2023. GraphscoreDTA: optimized graph neural network for protein–ligand binding affinity prediction. Bioinformatics 39. https://doi.org/10.1093/BIOINFORMATICS/BTAD340

Wang, Y., Wei, Z., Xi, L., 2022. Sfcnn: a novel scoring function based on 3D convolutional neural network for accurate and stable protein–ligand affinity prediction. BMC Bioinformatics 23, 1–18. https://doi.org/10.1186/S12859-022-04762-3/FIGURES/5

Wapeesittipan, P., Mey, A.S.J.S., Walkinshaw, M.D., Michel, J., 2019. Allosteric effects in cyclophilin mutants may be explained by changes in nano-microsecond time scale motions. Commun Chem 2, 1–9. https://doi.org/10.1038/s42004-019-0136-1

Wayment-Steele, H.K., Ojoawo, A., Otten, R., Apitz, J.M., Pitsawong, W., Hömberger, M., Ovchinnikov, S., Colwell, L., Kern, D., 2024. Predicting multiple conformations via sequence clustering and AlphaFold2. Nature 625, 832–839. https://doi.org/10.1038/s41586-023-06832-9

Weinert, T., Olieric, N., Cheng, R., Brünle, S., James, D., Ozerov, D., Gashi, D., Vera, L., Marsh, M., Jaeger, K., Dworkowski, F., Panepucci, E., Basu, S., Skopintsev, P., Doré, A.S., Geng, T., Cooke, R.M., Liang, M., Prota, A.E., Panneels, V., Nogly, P., Ermler, U., Schertler, G., Hennig, M., Steinmetz, M.O., Wang, M., Standfuss, J., 2017. Serial millisecond crystallography for routine room-temperature structure determination at synchrotrons. Nat Commun 8, 542. https://doi.org/10.1038/s41467-017-00630-4

Wellawatte, G.P., Gandhi, H.A., Seshadri, A., White, A.D., 2023. A Perspective on Explanations of Molecular Prediction Models. J Chem Theory Comput 19, 2149–2160. https://doi.org/10.1021/ACS.JCTC.2C01235/ASSET/IMAGES/LARGE/CT2C01235_0005.JPEG

Witek, J., Smusz, S., Rataj, K., Mordalski, S., Bojarski, A.J., 2014. An application of machine learning methods to structural interaction fingerprints—a case study of kinase inhibitors. Bioorg Med Chem Lett 24, 580–585. https://doi.org/10.1016/j.bmcl.2013.12.017

Wittmann, Bruce J, Johnston, K.E., Wu, Z., Arnold, F.H., 2021. Advances in machine learning for directed evolution. Curr Opin Struct Biol 69, 11–18.

Wittmann, Bruce J., Yue, Y., Arnold, F.H., 2021. Informed training set design enables efficient machine learning-assisted directed protein evolution. Cell Syst 12, 1026-1045.e7. https://doi.org/10.1016/J.CELS.2021.07.008

Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, Klaus-Robert Müller, 2019. Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Lecture Notes in Computer Science. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-030-28954-6

Wold, S., Eriksson, L., Hellberg, S., Jonsson, J., Sjostrom, M., Skagerberg, B., Wikstrom, C., 2011. Principal property values for six non-natural amino acids and their application to a structure–activity relationship for oxytocin peptide analogues. https://doi.org/10.1139/v87-305 65, 1814–1820. https://doi.org/10.1139/V87-305

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

Woodley, J.M., 2022. Ensuring the Sustainability of Biocatalysis. ChemSusChem 15, e202102683. https://doi.org/10.1002/CSSC.202102683

Wu, S., Snajdrova, R., Moore, J.C., Baldenius, K., Bornscheuer, U.T., 2021. Biocatalysis: Enzymatic Synthesis for Industrial Applications. Angewandte Chemie International Edition 60, 88–119. https://doi.org/10.1002/ANIE.202006648

Wu, Z., Jennifer Kan, S.B., Lewis, R.D., Wittmann, B.J., Arnold, F.H., 2019. Machine learning-assisted directed protein evolution with combinatorial libraries. Proc Natl Acad

Sci U S A 116, 8852–8858. https://doi.org/10.1073/PNAS.1901979116/SUPPL_FILE/PNAS.1901979116.SAPP.PDF

Xia, C., Feng, S.H., Xia, Y., Pan, X., Shen, H. Bin, 2023. Leveraging scaffold information to predict protein-ligand binding affinity with an empirical graph neural network. Brief Bioinform 24. https://doi.org/10.1093/BIB/BBAC603

Xiao, S., Tian, H., Tao, P., 2022. PASSer2.0: Accurate Prediction of Protein Allosteric Sites Through Automated Machine Learning. Front Mol Biosci 9, 879251. https://doi.org/10.3389/FMOLB.2022.879251/BIBTEX

Xu, G., Dou, Z., Chen, Xuanzao, Zhu, L., Zheng, X., Chen, Xiaoyu, Xue, J., Niwayama, S., Ni, Y., 2024. Enhanced stereodivergent evolution of carboxylesterase for efficient kinetic resolution of near-symmetric esters through machine learning. https://doi.org/10.21203/RS.3.RS-3897762/V1

Xu, Y., Verma, D., Sheridan, R.P., Liaw, A., Ma, J., Marshall, N.M., McIntosh, J., Sherer, E.C., Svetnik, V., Johnston, J.M., 2020. Deep Dive into Machine Learning Models for Protein Engineering. J Chem Inf Model 60, 2773–2790. https://doi.org/10.1021/ACS.JCIM.0C00073/ASSET/IMAGES/LARGE/CI0C00073_0008.JPEG

Xu, Z., Wu, J., Song, Y.S., Mahadevan, R., 2022. Enzyme Activity Prediction of Sequence Variants on Novel Substrates using Improved Substrate Encodings and Convolutional Pooling.

Yamada, H., Kobayashi, M., 1996. Nitrile hydratase and its application to industrial production of acrylamide. Biosci Biotechnol Biochem 60, 1391–1400. https://doi.org/10.1271/BBB.60.1391

Yang, J., Li, F.-Z., Arnold, F.H., 2024. Opportunities and Challenges for Machine Learning-Assisted Enzyme Engineering. ACS Cent Sci. https://doi.org/10.1021/ACSCENTSCI.3C01275

Yang, K.K., Eleutherai, N.Z., Yeh, H., 2022. Masked inverse folding with sequence transfer for protein representation learning. bioRxiv 2022.05.25.493516. https://doi.org/10.1101/2022.05.25.493516

Yang, K.K., Wu, Z., Arnold, F.H., 2019. Machine-learning-guided directed evolution for protein engineering. Nature Methods 2019 16:8 16, 687–694. https://doi.org/10.1038/s41592-019-0496-6

Yang, L., Yang, G., Chen, X., Yang, Q., Yao, X., Bing, Z., Niu, Y., Huang, L., Yang, Lei, 2021. Deep Scoring Neural Network Replacing the Scoring Function Components to Improve the Performance of Structure-Based Molecular Docking. ACS Chem Neurosci 12, 2133–2142. https://doi.org/10.1021/ACSCHEMNEURO.1C00110/SUPPL_FILE/CN1C00110_SI_001.PDF

Yang, M., Fehl, C., Lees, K. V., Lim, E.K., Offen, W.A., Davies, G.J., Bowles, D.J., Davidson, M.G., Roberts, S.J., Davis, B.G., 2018. Functional and informatics analysis enables glycosyltransferase activity prediction. Nat Chem Biol 14, 1109–1117. https://doi.org/10.1038/S41589-018-0154-9

Yang, Y., Gao, J., Wang, J., Heffernan, R., Hanson, J., Paliwal, K., Zhou, Y., 2018. Sixty-five years of the long march in protein secondary structure prediction: the final stretch? Brief Bioinform 19, 482–494. https://doi.org/10.1093/BIB/BBW129

Yang, Y., Niroula, A., Shen, B., Vihinen, M., 2016. PON-Sol: prediction of effects of amino acid substitutions on protein solubility. Bioinformatics 32, 2032–2034. https://doi.org/10.1093/BIOINFORMATICS/BTW066

Yang, Y., Zeng, L., Vihinen, M., 2021. PON-Sol2: Prediction of Effects of Variants on Protein Solubility. Int J Mol Sci 22. https://doi.org/10.3390/IJMS22158027

Yang, Z., Zhong, W., Lv, Q., Dong, T., Yu-Chian Chen, C., 2023. Geometric Interaction Graph Neural Network for Predicting Protein-Ligand Binding Affinities from 3D Structures (GIGN). Journal of Physical Chemistry Letters 14, 2020–2033. https://doi.org/10.1021/ACS.JPCLETT.2C03906/SUPPL_FILE/JZ2C03906_SI_001.PDF

Yang, Z., Zhong, W., Zhao, L., Yu-Chian Chen, C., 2022. MGraphDTA: deep multiscale graph neural network for explainable drug–target binding affinity prediction. Chem Sci 13, 816–833. https://doi.org/10.1039/D1SC05180F

Yeh, A.H.W., Norn, C., Kipnis, Y., Tischer, D., Pellock, S.J., Evans, D., Ma, P., Lee, G.R., Zhang, J.Z., Anishchenko, I., Coventry, B., Cao, L., Dauparas, J., Halabiya, S., DeWitt, M., Carter, L., Houk, K.N., Baker, D., 2023. De novo design of luciferases using deep learning. Nature 2023 614:7949 614, 774–780. https://doi.org/10.1038/s41586-023-05696-3

Yosinski, J., Clune, J., Bengio, Y., Lipson, H., 2014. How transferable are features in deep neural networks? Adv Neural Inf Process Syst 27.

Yu, T., Cui, H., Li, J.C., Luo, Y., Jiang, G., Zhao, H., 2023. Enzyme function prediction using contrastive learning. Science (1979) 379, 1358–1363.

Zaretzki, J., Bergeron, C., Rydberg, P., Huang, T.W., Bennett, K.P., Breneman, C.M., 2011. RS-predictor: A new tool for predicting sites of cytochrome P450-mediated metabolism applied to CYP 3A4. J Chem Inf Model 51, 1667–1689. https://doi.org/10.1021/CI2000488/SUPPL_FILE/CI2000488_SI_001.ZIP

Zaretzki, J., Matlock, M., Swamidass, S.J., 2013. XenoSite: Accurately predicting cyp-mediated sites of metabolism with neural networks. J Chem Inf Model 53, 3373–3383. https://doi.org/10.1021/CI400518G/SUPPL_FILE/CI400518G_SI_002.ZIP

Zhao, H., Arnold, F.H., 1999. Directed evolution converts subtilisin E into a functional equivalent of thermitase. Protein Engineering, Design and Selection 12, 47–53. https://doi.org/10.1093/PROTEIN/12.1.47

Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., Sun, M., 2020. Graph neural networks: A review of methods and applications. AI Open 1, 57–81. https://doi.org/10.1016/J.AIOPEN.2021.01.001

Zhou, J., Troyanskaya, O.G., 2015. Predicting effects of noncoding variants with deep learning–based sequence model. Nature Methods 2015 12:10 12, 931–934. https://doi.org/10.1038/nmeth.3547