Enzyme Substrate Prediction from Three-Dimensional Feature Representations Using Space-Filling Curves

Dmitrij Rappoport[1] *and Adrian Jinich*[2]

[1] Department of Chemistry, University of California, Irvine, 1102 Natural Sciences 2, Irvine CA, 92697.

Email: dmitrij@rappoport.org

[2] Weill Cornell Medicine, 1300 York Ave, Box 65, New York, NY 10065. Email:

adj2010@med.cornell.edu

Abstract

Compact and interpretable structural feature representations are required for accurately predicting properties and function of proteins. In this work, we construct and evaluate three-dimensional feature representations of protein structures based on space-filling curves. We focus on the problem of enzyme substrate prediction, using two ubiquitous enzyme families as case studies: the short-chain dehydrogenase/reductases (SDRs) and the S-adenosylmethionine dependent methyltransferases (SAM-MTases). Space-filling curves such as the Hilbert curve and the Morton curve generate a reversible mapping from discretized three-dimensional to one-dimensional representations and thus help to encode three-dimensional molecular structures in a system-independent way and with only a few adjustable parameters. Using three-dimensional structures of SDRs and SAM-MTases generated using AlphaFold2, we assess the performance of the SFC-based feature representations in predictions on a new benchmark database of enzyme classification tasks including their cofactor and substrate selectivity. Gradient-boosted tree classifiers yield binary prediction accuracy of 0.77–0.91 and AUC (area under curve) characteristics of 0.83–0.92 for the classification tasks. We investigate the effects of amino acid encoding, spatial orientation, and (the few) parameters of SFC-based encodings on the accuracy of the predictions. Our results suggest that geometry-based approaches such as SFCs are promising for generating protein structural representations and are complementary to the existing protein feature representations such as ESM sequence embeddings.

Introduction

The function of the vast majority of proteins in general, and enzymes in particular, remains unknown. For example, according to the UniProt Knowledgebase (UniProtKB) [1], ≈ 60% of human proteins have the lowest annotation category (1-out-of-5), while only ≈ 3% have functional annotations with experimentally evidence that merit the highest annotation score [2]. While recent advances in protein structure prediction [3,4] will accelerate protein functional annotation, how best to map structure to function remains a challenging and open problem.

Different representations of protein structure can be used to map structure to function. One common approach uses 3D convolutional neural networks [5–7]. Concentric shells surrounding specific protein sites can also be used to represent the spatial distribution of biochemical properties [8]. Other methods use graph representations of enzymes and their active sites [9,10]. For example, Gligorijević et al. [9] generate amino contact maps from protein structures, and use the resulting adjacency matrix as input for a graph convolutional network. The distribution of torsion angles and pairwise distances, extracted for each amino acid type separately, can also generate feature maps for downstream protein or enzyme functional prediction [7,11].

A specific challenge in predicting function from protein structure is incorporating the full three-dimensional structure in addition to the amino acid sequence and the topological properties, as expressed, for example, by the residue contact maps [12–17]. Two possible paths towards this goal are using raw or minimally processed three-dimensional structural data together with specialized machine learning models capable of processing multidimensional data, such as convolutional neural networks [18,19], or using one-dimensional ("flattened") feature representations in combination with predictive models operating on conventional vector inputs—a much broader class of machine learning models that includes such popular and versatile methods as gradient-boosted trees [20–22] and support vector machines (SVM) [23–25]. Here we focus on the second alternative, which allows to decouple the task of obtaining features out of the raw structural data from that of learning the functional relationship between the extracted features and the target properties, while the

convolutional methods must perform both tasks at once. If the generation of one-dimensional feature representations is performed in a separate step and is independently optimized, one may expect predictions that generalize well, are easier to interpret, and possibly require smaller training sets. A variety of approaches for "flattening" multidimensional data have been developed for efficient storage and retrieval in databases, for example, random projections [26,27], embeddings [28], and fractal geometry-based methods [29,30]. All these methods aim to generate a linear ordering of objects in multidimensional space while approximately preserving near-neighbor relationships. The linear ordering serves as an index for similarity and range queries. The fractal geometry-based methods using space-filling curves (SFCs) [30,31] are particularly attractive for our purposes because their constructions are reversible and deterministic. Moreover, they can incorporate the invariances of protein structures such as rigid translations, rotations, and atom renumbering in a straightforward way.

In this paper we explore the generation of SFC-based feature representations of protein structures and their applications for enzyme substrate predictions. SFCs denote a class of curves in d-dimensional Euclidean space $\mathbb{R}^d$ ($d \geq 2$) with the property that they pass through every point of a square (for $d = 2$) or cube (for $d = 3$) [29,30]. Specifically, a SFC is defined by a continuous mapping of the (one-dimensional) unit interval [0, 1] onto an n-dimensional object with non-zero area (for $d = 2$), volume (for $d = 3$), or its higher-dimensional equivalent (for $d > 3$). The SFC thus "linearizes" the set of points in $\mathbb{R}^d$ ($d \geq 2$) by prescribing a well-defined order of their traversal. The first SFC was described by Peano in 1890 [32]. In the following year, Hilbert published the definition and the geometric procedure for generating the eponymous curve [33]. The Morton curve (also named Z-order curve) first appeared in a technical report in 1966 [34]. SFCs are constructed by recursive geometric procedure, which is schematically shown in Figure 1 for the Hilbert and Morton curves in the $d = 2$ case. The construction of the Hilbert curve proceeds by successively partitioning the square into subsquares and arranging them such that the curve passes through subsquares that share an edge. The same procedure is iteratively repeated for each subsquare. The Morton curve differs in the order of the traversal of the subsquares. The approximations to the two-dimensional SFC at finite resolution are the piecewise linear curves with 2^{2w} segments (approximating polygons), where $w = 1, 2, \ldots$ is the resolution

index. The SFC results in the limit $w \to \infty$. The construction can be generalized to d dimensions ($d > 2$), in which case the approximating polygon at resolution index w has 2^{dw} segments and traverses the entire (hyper)cube in $\mathbb{R}^d$. By varying the relative orientations of the basic patterns and the approximating polygons in the recursion, a family of related d-dimensional SFCs of each curve type results for $d > 2$ [31]. The SFC encoding is reversible, that is, for every segment of the SFC, we can determine the corresponding location in the d-dimensional space.

The self-similar construction of SFCs approximately preserves proximity, that is, points that are close in the linear SFC ordering are mapped to near points in $\mathbb{R}^d$. Conversely, most pairs of near points in $\mathbb{R}^d$ are also close to each other in the SFC ordering. SFCs have found applications in optimization [35–38], multidimensional indexing [39,40], numerical simulations [41,42], and parallel algorithms [43,44]. The locality preservation of the inverse mapping $\mathbb{R}^d \to [0, 1]$ is necessarily imperfect—one can always find a pair of points in $\mathbb{R}^d$ that are far apart in the SFC ordering [30,31]. Consider, for example, the turning points on the opposite sides of the two-dimensional Hilbert curve in Figure 1. Locality measures corresponding to the average and worst-case distance in $\mathbb{R}^d$ for neighboring points in the SFC ordering have been evaluated for different curve types and dimensionalities [31,45,46]. The Hilbert curve is optimal with respect to the worst-case locality measure in the two-dimensional case and has been shown to perform well in practice. SFCs thus offer a straightforward and reversible scheme of encoding three-dimensional protein structures as one-dimensional vectors with good information fidelity. At the same time, as we show below, the construction of SFC-based encodings is controlled by only a few system-independent adjustable parameters, which helps avoiding overfitting and improving interpretability.

As protein function can be described in multiple ways [47], we focus here on the enzyme-to-substrate mapping problem as a specific instance of the larger protein functional annotation challenge. Predicting enzyme function is of practical value in the context of metabolic engineering, where it can extend the set of enzymes to support novel pathways [48–52]. Several structure-based [53–55] and/or sequence-based [7,56–91] approaches to enzyme function prediction have been reported. The large majority of these methods focus only on predicting the enzyme family, typically in the form of its Enzyme Commission (EC) number, a four-digit

number that maps enzymes to a hierarchical classification scheme. The information obtained from predictions at this level of classification is often already encoded in the protein domains found within a sequence [92], which are readily accessible in protein databases such as UniProt [1,92], even for orphan, poorly annotated proteins. As an illustration, querying a recent enzymatic function prediction software [56] with the sequence for an orphan oxidoreductase in *Mycobacterium tuberculosis* (Rv3502c) as input, the output is the predicted EC class 1.1.1.-. This EC number corresponds (hierarchically) to: oxidoreductases (EC 1); acting on the CH–OH group of donors (EC 1.1); with nicotinamide adenine dinucleotide (NAD) or nicotinamide adenine dinucleotide phosphate (NADP) as acceptor (EC 1.1.1). The two predictions reported by the first and third EC number digits (EC 1 and EC 1.1.1) are encoded in the fact that the protein belongs to the "Short-chain dehydrogenase/reductase SDR" family (IPR002347), as per its "Family and Domains" annotation in UniProt and InterPro. However, the categorization of the orphan protein in question as a member of the EC class 1.1.1.- leaves its native substrate (secondary alcohol) and electron donor (NAD or NADP) unspecified and is too loose to determine its role in metabolism. In particular, enzymes of the short-chain dehydrogenase/reductases (SDRs) family are involved in metabolism of fatty acids, steroids, amino acids, xenobiotics, chromophores of visual perception, among others. [1] As this example shows, functional prediction must often be extended to include more granular predictions of enzyme substrates and cofactors within a protein family, as these can help generate testable hypotheses and guide downstream experimental efforts. Furthermore, despite the importance of functional predictions in enzymes, structured training and test data for machine learning models of enzymatic function are difficult to obtain.

In this work, we introduce a standardized dataset of annotated enzymes from two important protein families —SDRs and S-adenosylmethionine-dependent methyltransferases (SAM-MTases)—and a set of labels that classify them according to their cofactor and substrate structural class preference. We chose these enzyme families because of their size and of the chemical variability of their substrates, which necessitates more specific function predictions than enzyme class. At the same time, these enzyme families are structurally very different, thus providing us with a test of transferability of SFC-based feature representations. This dataset can be used to assess different protein representations and their use as input for machine learning

methods that predict enzyme-substrate associations. We then show that a representation of protein structure using SFCs can be used to accurately predict enzyme cofactor and substrate structural class preference across a wide range of tasks. We find that orientational sampling of protein structural conformations increases the accuracy of both cofactor and substrate structural class predictions. Comparisons to ESM sequence embeddings of SDRs and SAM-MTases [93,94] show that ESM feature representations have better performance than SFC-based structure encodings within the logistic regression classification model, however, SFC-based feature representations still provide very good predictions in combination with gradient-boosted trees [20–22].

Methods

Enzyme Selection and Benchmark Set

Enzymes belonging to the SDR and SAM-dependent methyltransferases were obtained from the UniProt database [1]. Specifically, we performed queries using InterPro protein family/domain identifiers corresponding to each family: IPR002347 for short-chain dehydrogenase/reductase (SDRs) and IPR029063 for S-adenosylmethionine-dependent methyltransferases (SAM-MTases). The resulting tables contain each enzyme's UniProt entry name, amino acid sequence and UniProt annotation score, a heuristic measure of annotation favoring literature-curated entries with experimental evidence [95]. In addition for experimentally well-annotated enzymes, it includes Rhea [96] and ChEBI [97] database identifiers mapping each enzyme to the catalyzed biochemical reaction and its substrate and products structure represented as SMILES strings [98]. We filtered the full set of enzymes in each family, keeping only those with UniProt annotation scores above "4-out-of-5", to obtain the subset of well-annotated enzymes with known substrates for inclusion in the benchmark set. The lists of protein structures were de-duplicated by their amino acid sequences, leaving 358 distinct SDR and 953 distinct SAM-MTase structures. The compiled benchmark dataset is available from Zenodo online repository under the Creative Commons 4.0 license [99].

Labeling Enzymes Using Substrate Clustering

Our first approach to classify enzymes according to the structural properties of the compounds they act on is based on structural clustering of the substrates and products of all annotated enzymes within a family. Broadly, we reasoned that one way to frame a machine learning classification task is to predict whether a given enzyme can catalytically act on substrates that belong to a structurally-related group of compounds, with the relevant clusters of compounds defined in an unsupervised manner. To implement this approach, we first removed all cofactors (NAD(H), NADP(H) for SDRs and S-adenosylmethionine for SAM-MTases), which appear in every reaction within each enzyme family, from the list of reaction components. Then, using RDKit [100], we took as input the SMILES string representations of all enzymatic substrates and products and obtained their corresponding Morgan fingerprints [101] as bit vectors with radius = 3. We then generated a 2-dimensional projection of the Morgan fingerprints using the UMAP algorithm [102], with the Jaccard metric for binary vectors. Since we eventually trained separate machine learning classifiers for each enzyme family, we note that the SDR and SAM-MTase substrates and products were processed separately at this stage. Finally, using *k*-means clustering [103,104] we clustered the compounds according to their 2D UMAP projection [102], choosing the optimal number of clusters that maximized the silhouette score [105]. This clustering procedure generated 9 structure clusters for SDR substrates (numbered 0–8, as shown in Figure 2) and 13 structure clusters for SAM-MTase substrates (numbered 0–12). The substrate structures and enzyme clusters are included in the Supporting Information (SI).

Labeling Enzymes Using Substructure Search on Substrates and Products

Our second approach to label enzymes according to the type of compound they act on is based on searching for broad classes of substrates represented by molecular substructures. We used RDKit to search for molecular patterns (phenol, sterol, or acyl-CoA) encoded as SMARTS strings [106]. The presence or absence of these substructures in the substrates or products was treated directly as a classification label. The number of SAM-MTases acting on sterols was too small for making statistically significant predictions. Thus we did not consider this classification task further.

Three-Dimensional Structure Generation

We constructed three-dimensional feature representations of 358 unique SDR and 953 unique SAM-MTase structures and assessed their performance on a set of binary classification tasks related to their cofactor and substrate specificity. The initial all-atom structures were generated by the public version of the AlphaFold2 protein folding approach [4] via the ColabFold notebook [107]. MMSeq2 multiple sequence alignment algorithm [108,109] was used instead of the original AlphaFold2 homology search due to its faster execution times [107]. The protein sequences and generated protein structures in PDB format are available from Zenodo [99].

Structural Encoding

To generate feature representations of three-dimensional protein structures, the all-atom input coordinates were first preprocessed to remove translational and rotational degrees of freedom and coarse-grained at the amino acid residue level. Each residue was represented by its center of mass coordinates (computed using non-hydrogen atoms) and its type. The preprocessed coordinates were converted to the SFC-based feature representation by discretizing the space coordinates, SFC-based encoding, and subsequent binning. The resulting representation is a sparse fixed-length binary vector that is easy to use in machine learning tasks, for example, classification, regression, or similarity calculations. The structural encoding procedure is summarized in Table 1. All coordinates are in atomic units (au).

Table 1. *Scheme of protein three-dimensional structure encoding algorithm.*

1	**Preprocessing**	
1A	Translation/Rotation	$\{X_A^{at}, Y_A^{at}, Z_A^{at}\} \in \mathbb{R}, A = 1, \ldots, N$
1B	Coarse Graining	$\{x_i, y_i, z_i\} \in \mathbb{R}, i = 1, \ldots, n$
1C	Orientation Sampling	$\{x'^j_i, y'^j_i, z'^j_i\} \in \mathbb{R}, i = 1, \ldots, n, j = 1, \ldots, s$
2	**Discretization**	

2A	Space Coordinate Binning	$\{\underline{x}_i, \underline{y}_i, \underline{z}_i\} \in [0, 2^w-1]$, $i = 1, \ldots, n$
2B	Amino Acid Encoding	$\{\alpha_i\} \in [0, 2^w-1]$, $i = 1, \ldots, n$
3	**SFC Encoding**	$\{\zeta_i\} \in [0, 2^{4w}-1]$, $i = 1, \ldots, n$ $\equiv \{Z_k\} \in [0, 1]$, $k = 1, \ldots, 2^{4w}$, n 1's
4	**Binning**	$\{b_k\} \in \mathbb{N}$, $k = 1, \ldots, 2^{4w/b}$

The translation/rotation preprocessing step (1A) removes the dependence of the input coordinates on arbitrary translations and rotations of the coordinate axes by transforming them to the structure's inertial frame coordinate system. The coordinate origin is moved to the center of mass (computed using non-hydrogen atoms), and the coordinate axes are oriented along principal axes of inertia in the order of decreasing moment of inertia. The resulting all-atom coordinates $\{X_A^{\text{at}}, Y_A^{\text{at}}, Z_A^{\text{at}}\}$, $A = 1, \ldots, N$, where N is the number of atoms, only depend on the internal degrees of freedom. We note that for a series of similar proteins, the use of inertial-frame coordinates improves the superposition of their three-dimensional structures but does not guarantee optimal alignment. The use of orientation sampling (see step 1C below) alleviates this issue, however, dedicated methods for multiple three-dimensional alignment should also be considered [110].

The following coarse-graining step (1B) represents each amino acid residue by its center-of-mass coordinates (based on non-hydrogen atoms) $\{x_i, y_i, z_i\}$, $i = 1, \ldots, n$, where n is the number of amino acids, and its residue type. Coarse graining is convenient in protein studies [111] because it encodes the internal structure of each residue as a single discrete variable (see 2B below) but all-atom approaches are more appropriate in other classes of molecules.

The exact SFC encoding can be approximated at different finite resolutions, which we identify by their bit width w. For a given value of w, each input coordinate (both spatial and residue type) is represented on a discrete grid of 2^w points. The continuous spatial coordinates are first discretized as integer indices $\{\underline{x}_i, \underline{y}_i,$

$\underline{z}_i\}$, $i = 1, \ldots, n$, by placing the protein structure on a uniform cubic grid with 2^w cells per Cartesian direction and side length L (step 2A). The amino acid residue type is also encoded by an integer index $\{\alpha_i\}$, $i = 1, \ldots, n$, in the range $[0, 2^w - 1]$ using a procedure described in the next section (step 2B). The SFC encoding algorithm requires all coordinates to have the same number of grid points. The range of values for the residue type is thus rescaled if necessary to the same number of grid cells as the spatial coordinates.

The SFC-based encoding (step 3) converts each quadruple $\{\alpha_i, \underline{x}_i, \underline{y}_i, \underline{z}_i\}$ of w-bit integers losslessly to a $4w$-bit integer coordinate ζ_i that gives its location index along the space-filling curve in the four-dimensional space. In this work, we use both the Hilbert SFC [30,33] and the Morton curve [34] for structure encoding. We note that the encoding is not invariant with the respect to the ordering of the Cartesian coordinates: the change in the first coordinate α_i is weighted more strongly than the changes in the subsequent coordinates $\underline{x}_i$, $\underline{y}_i$, and $\underline{z}_i$ (in order of decreasing weight).

The SFC coordinate ζ_i indicates the presence of an amino acid residue of type α_i at the location given by the integer indices $\{\underline{x}_i, \underline{y}_i, \underline{z}_i\}$. It follows that we can alternatively view the vector of n integer coordinates of $4w$ bit length as an extremely sparse binary vector $\{Z_k\}$, $k = 1, \ldots, 2^{4w}$, which contains only n non-zero elements. This alternative view represents the protein structure as an unordered set of n locations (out of 2^{4w} possible) that are occupied by amino acid residues. This representation has several useful properties. It is that it is invariant with respect to a renumbering of the residues. Moreover, it enables comparisons between proteins of different length n. Finally, it can be used to construct feature representations at different resolutions by binning (see below). The mapping of the ordered vector $\{\zeta_i\}$, $i = 1, \ldots, n$, to the binary vector (unordered set) $\{Z_k\}$ may seem to lose information about the order of amino acid residues in the protein sequence. However, this is not the case because the residue locations (and types) are encoded in the indices ζ_i and can be reconstructed when needed. Additionally, two different residues are extremely unlikely to share an index ζ_i unless the box size L is too small. In the absence of these improbable collisions, the vector $\{\zeta_i\}$ and the binary vector $\{Z_k\}$ are equivalent.

Finally, the binning (step 4) compresses the vector Z by aggregating its entries consecutively in bins of width 2^b. The resulting integer vector $\{b_k\}$ is of fixed length $2^{4w/b}$ with at most n non-zero elements. Due to

the locality-preserving property of SFCs, each bin maps to a contiguous subset of the four-dimensional space, and consecutive bins correspond to neighboring subsets.

Amino Acid Encoding

Several numerical encodings of natural amino acids (AA) have been proposed, of which the most common ones are based on substitution frequencies in proteins [112–118] or the physical or conformational properties of amino acids [119–127].The first group of encodings is based on the evolutionary notion that amino acids are more likely to be homologous with similar amino acids. The amino acid encoding from the substitution frequency matrix [112,113,115,128] by embedding in low-dimensional Euclidean spaces [114,116,117]. The second group of encodings starts from an empirical set of physical and chemical properties of AAs and reduces them to low-dimensional vectors by principal component analysis (PCA) [120,125,127].

In this work, we used a discretized version of the five-dimensional encoding of Li and Koehl (LK), which was obtained from the BLOSUM62 similarity matrix by including the 5 largest principal components [117]. To obtain an integer representation compatible with the discretized spatial coordinates, the LK feature vectors for the 20 natural AAs were arranged in the order of their decreasing PCA eigenvalues and encoded as integers $\alpha' = (a_0, a_1, a_2, a_3, a_4)$ in the range $[0, 2^5–1 = 31]$ with the j-th bit $a_j = 0$ if the j-th PCA component in the LK encoding was negative and 1 if it was positive. This approach generated several duplicates for similar amino acids, which were resolved manually, giving the encoding shown in Table 2. To generate a compatible grid to the spatial coordinates for SFC encoding, the 5-bit encodings α' were uniformly scaled as $\alpha = 2^{w-5}\,\alpha'$.

Table 2. *Modified Li–Koehl AA encoding (5-bit).*

AA	α′	AA	α′	AA	α′	AA	α′
Ala	21	Gln	25	Leu	0	Ser	22
Arg	26	Glu	24	Lys	17	Thr	23
Asn	31	Gly	29	Met	1	Trp	13

Asp	30	His	27	Phe	11	Tyr	10
Cys	7	Ile	2	Pro	20	Val	3

Additionally, we investigated an encoding based on the five-dimensional z-scales of Wold and co-workers [124,125]. The results are described in the SI.

Orientation Sampling

For a given pair of three-dimensional structures of the same composition, optimal alignment in terms of the root-mean-squared deviation can be efficiently computed [129,130]. However, extensions to multiple alignment and different compositions are non-trivial. Instead of trying to find a single best alignment, we replicated the input protein structure under a fixed set of orientations $\{x'^j_i, y'^j_i, z'^j_i\}$, $i = 1, \ldots, n, j = 1, \ldots, s$, which are chosen to uniformly sample the space $SO(3)$ of rotations in $\mathbb{R}^3$. The underlying assumption in our approach is that, for any given pair of structures, one or more sampled orientations will be reasonably close to the optimal superposition. A deterministic and computationally simple approach to uniform sampling of rotations is the method of successive orthogonal images (SOI) [131]. The base grid generated by the SOI method consists of $s = 72$ orientations, which can be progressively refined. In this work, we limited ourselves to considering only the base SOI grid. The SFC encoding was carried out independently for each orientation to generate the binary vectors $\{Z'^j_k\}$, $j = 1, \ldots, s$, which were subsequently combined into sparse binary vector $\{Z_k\}$, $k = 1, \ldots, 2^{4w}$ by binary OR operations. The resulting vector contains ns non-zero elements, barring collisions, and can be further aggregated to give the fixed-width integer vector $\{b_k\}$, $k = 1, \ldots, 2^{4w/b}$, as in the base case.

Results

Benchmark Datasets

Our standardized enzymes-substrate dataset contains reference data for 358 unique short-chain dehydrogenase reductases (SDRs) and 953 unique S-adenosylmethionine dependent methyltransferases (SAM-MTases) structures. For SDRs, the reference data includes the preference for the redox cofactor

(NAD(H) or NADP(H)) and the specificity with respect to a broad substrate class (e.g., sterols, and acyl-CoA substrates). For SAM-MTases, the enzymes acting on biopolymers (proteins or nucleic acids) or on small molecules are distinguished. The enzyme specificity with respect to two substrate classes (phenols and N-heterocycles) is also classified. Additionally, the binary classifications of substrate specificities of SDRs and SAM-MTases by substrate clusters from unsupervised clustering are included in the benchmark set [99].

The performance of the three-dimensional feature representations was assessed in enzymatic function predictions using our benchmark set. For SDRs, the benchmark set contains binary classification tasks with respect to the redox cofactor (NAD(H) or NADP(H)) and to substrate class (phenols, sterols, and acyl-CoA). For SAM-MTases, the classification of enzymes acting on biopolymers (proteins or nucleic acids) or on small molecules and substrate specificity with respect to phenols and N-heterocycles are included in the benchmark set. Additionally, binary classifications of enzyme specificity of SDRs and SMT-MTases based on substrate clusters from unsupervised clustering are part of the benchmark set.

The protein structures were preprocessed and converted to three-dimensional feature representations as described in the Methods section. Both Hilbert and Morton SFCs were studied for the SFC encoding. The discretization box was of side length $L = 200$ au (100 au for SAM-MTases) and of $w = 8$ bit width resolution. The residue types were encoded using a discretized version of the Li–Koehl (LK) amino acid encoding [117]. The binary vectors were binned using $2^b = 256$, 4096, and 65536 bins. Orientation sampling using the SOI method at the base grid level ($s = 72$ orientations) was applied. In the following, we evaluate the performance of three-dimensional feature representations as a function of the classification task, SFC type, and orientation sampling. We compare the full encodings containing the three-dimensional protein structure representations including the residue types with simplified encodings containing only amino acid types (bag-of-amino acids representation) or only the residue positions, regardless of amino acid type (structure-only representation).

The classifications were performed using the XGBoost binary classifier with GBTree booster, maximum depth 8, learning rate 0.2, and $L_2 = 1$ regularization [21]. For each classification task, we report the 5-fold

cross validated accuracy (percentage of correct predictions) and the receiver operator characteristic (ROC) area under curve (AUC). All classifications were performed using the scikit-learn library, version 1.0.2 [132]. The results of hyperparameter optimization using Bayesian optimization methods are included in the SI. We find that the results are not strongly dependent on the parameter choice. Classifications using logistic regression, support vector machine classification (SVMC) [23,24] and LightGBM methods [22] produced similar results and are reported in the SI.

Classification Task

The accuracy and AUC characteristics of 4 binary classification tasks for SDRs and 3 binary classification tasks for SAM-MTases using Hilbert space-filling curve are given in Table 3. Figure 3 (left panel) shows the ROC curve for the NAD/NADP classification with and without random shuffling of the labels. We obtain accuracy values (percentage of correctly predicted classes) between 0.77–0.91 for the classification tasks. The AUC characteristics for these binary classification tasks are between 0.83–0.92. The classification performed using randomly shuffled labels shown in Figure 3 (right panel) serves as a null hypothesis and predicts the AUC close to the theoretically expected value of 0.5 for a random binary classifier.

Table 3. *Accuracy and ROC area under curve (AUC) of structure encodings of SDRs and SAM-MTases using 8-bit Hilbert SFC, modified LK encoding, 4096 bins, orientation sampling using SOI (s = 72) for binary classification tasks (5-fold cross validated).*

SDR	Accuracy	AUC	SAM	Accuracy	AUC
NAD/NADP	0.77	0.84	Biopolymer/ small molecule	0.85	0.92
Phenols	0.81	0.83	Phenols	0.83	0.88
Sterols	0.83	0.83	N-Heterocycles	0.79	0.84
Acyl-CoA	0.91	0.90			

The results of binary classification tasks for substrate specificity to substrate clusters obtained from unsupervised clustering are shown in Table 4. Only clusters including 5 or more molecules (6 clusters for SDRs and 7 clusters for SAM-MTases) were included in classification tasks. The accuracy values of classification tasks with respect to clusters are in the range of 0.73–0.94 and the AUC values are 0.72–0.98. We note that the classification results are similar for SDRs and SAM-MTases.

Table 4. *Accuracy and ROC area under curve (AUC) of structure encodings of SDRs and SAM-MTases using 8-bit Hilbert SFC, modified LK encoding (Table 2), 4096 bins, orientation sampling using SOI (s = 72) for binary classification tasks from unsupervised clustering (5-fold cross validated). Only tasks with at least 5 examples per class were considered.*

SDR	Accuracy	AUC	SAM	Accuracy	AUC
Cluster 0	0.93	0.75	Cluster 0	0.86	0.88
Cluster 1	0.90	0.85	Cluster 1	0.94	0.91
Cluster 2	0.82	0.84	Cluster 5	0.88	0.72
Cluster 4	0.86	0.84	Cluster 6	0.92	0.89
Cluster 5	0.95	0.98	Cluster 7	0.89	0.84
Cluster 7	0.73	0.76	Cluster 8	0.93	0.96
			Cluster 11	0.88	0.92

Space-Filling Curve Type

Using the Morton curve instead of Hilbert SFC for the encoding gave very similar binary classification results, which are shown in Table 5. The accuracy values were in the range of 0.76–0.90, while the AUC values were between 0.83–0.92, almost all of them within one percent point of the results obtained with Hilbert SFC.

Table 5. *Accuracy and ROC area under curve (AUC) of structure encodings of SDRs and SAM-MTases using 8-bit Morton SFC, modified LK encoding (Table 2), 4096 bins, orientation sampling using SOI (s = 72) for binary classification tasks (5-fold cross validated).*

SDR	Accuracy	AUC	SAM	Accuracy	AUC
NAD/NADP	0.76	0.84	Biopolymer/ small molecule	0.85	0.92
Phenols	0.83	0.83	Phenols	0.83	0.88
Sterols	0.83	0.83	N-Heterocycles	0.79	0.85
Acyl-CoA	0.90	0.90			

Bag-of-Amino Acids and Structure-Only Representations

As described above, the SFC encoding imposes an ordering in the *d*-dimensional representation in that a change in the first coordinate of the *d*-dimensional representation induces a larger difference in the SFC ordering than a change in the second coordinate, etc. Since we apply the SFC encoding to the tuple $\{\alpha_i, \underline{x}_i, \underline{y}_i, \underline{z}_i\}$, the residue type α_i has by design the largest weight in the encoded vector. In fact, in the limit of an infinitely large discretization box (side length $L \rightarrow \infty$), all atomic positions are mapped to the origin and only the residue type information remains. The SFC encoded vector approaches a very simple model of the protein structure, which can be denoted as a bag-of-amino acids. This representation contains only the counts of residue types and no spatial or sequence information. Interestingly, this feature representation produces quite similar results to the full SFC encoding, as shown in Table 6. The accuracy values for the binary classification tasks were between 0.81–0.91 and the AUC values were found to be in the range 0.79–0.92.

Table 6. *Accuracy and ROC area under curve (AUC) of bag-of-amino acids encodings of SDRs and SAM-MTases using modified LK encoding (Table 2), 32 bins for binary classification tasks (5-fold cross validated).*

SDR	Accuracy	AUC	SAM	Accuracy	AUC

NAD/NADP	0.81	0.88	Biopolymer/ small molecule	0.84	0.92
Phenols	0.82	0.86	Phenols	0.81	0.88
Sterols	0.86	0.86	N-Heterocycles	0.75	0.79
Acyl-CoA	0.91	0.86			

To test the importance of the residue type and spatial information, we also considered the converse case of the structure-only encoding, which contains only the three-dimensional coordinates of the residue but no residue type information. The results are shown in Table 7 and again perform similarly to those obtained with the full SFC encoding. The accuracy values were between 0.77–0.86 and the AUC values were 0.73–0.92.

Table 7. *Accuracy and ROC area under curve (AUC) of structure-only encodings of SDRs and SAM-MTases using 8-bit Hilbert SFC, 4096 bins, orientation sampling using SOI (s = 72) for binary classification tasks (5-fold cross validated).*

SDR	**Accuracy**	**AUC**	**SAM**	**Accuracy**	**AUC**
NAD/NADP	0.77	0.82	Biopolymer/ small molecule	0.85	0.92
Phenols	0.77	0.73	Phenols	0.83	0.88
Sterols	0.78	0.76	N-Heterocycles	0.78	0.79
Acyl-CoA	0.86	0.80			

Discussion

As our results show, enzymatic function can be predicted from feature representations based on the three-dimensional structures of SDRs and SAM-MTAses with good accuracy (> 80% binary classification accuracy, > 0.8 AUC). Moreover, the predictive power is not strongly dependent on the parameters of the representation, for example, type of the SFC encoding. As we show in the SI, other encoding parameters

such as the amino acid encoding and the number of bins do not affect the accuracy of the predictions. The consistence of the findings across unrelated protein families indicates that the performance of SFC-based encodings is not tied to similarities between related proteins. As a result, the methodology described here should generalize well to other protein families and other functional prediction tasks. In addition, the reversibility of the SFC encoding allows the importance of the original three-dimensional structural features to be evaluated, aiding interpretability. The predictions in this work were made with the XGBoost classification model, which has been shown to generalize well and to be relatively insensitive to model parameters. In the SI, we show the results of hyperparameter optimization of the XGBoost model for the NAD/NADP cofactor specificity of SDRs. Interestingly, the results obtained using optimized hyperparameters do not differ strongly from the parameter set used through this work.

We have also investigated the accuracy of the predictions with different binary classification methods. We find that logistic regression underperforms the XGBoost results somewhat, while the SVMC and LightGBM models give similar accuracy to XGBoost. For the NAD/NADP cofactor specificity, logistic regression yield an accuracy of 0.64 (AUC = 0.67), SVM classification gives 0.71 accuracy (AUC = 0.78), and LightGBM gives 0.77 accuracy (AUC = 0.83), compared to 0.77 accuracy (AUC = 0.84) with XGBoost classification. The model parameters and complete results are shown in the SI.

The results obtained without orientational sampling are slightly worse, which suggests that structure alignment is important to the prediction. However, the differences between the results obtained with and without orientational sampling are not very large. Additionally, global structural alignment may not be optimal for proteins with very different compositions, while methods for aligning active site regions may produce further improvements.

The fidelity of the 3D to 1D mapping can be analyzed in terms of the correlation between the inter-residue distances in three-dimensional space and along the SFC ordering. The average Pearson correlation coefficients for the Euclidean inter-residue distances and their encoding using the Hilbert SFC are $R^2 = 0.45$ for SDRs and 0.47 for SAM-MTases. For the encoding using the Morton curve, the correlation coefficients are similar, giving $R^2 = 0.43$ for SDRs and 0.50 for SAM-MTases. The correlation is affected by its

dependence on the ordering of the x, y, z Cartesian components, in that the influence of the change in the Cartesian coordinates decreases with $x > y > z$. Orientation-invariant encodings might be helpful in addressing this issue.

The greatest surprise in our results was a significant amount of redundancy contained in the feature representations. The bag-of-amino acids representation, which consists of the counts of different amino acid types, yields predictions that are quite similar to the full feature representation containing both amino acid types and their spatial locations. On the other hand, the structure-only representation, which elides the amino acid types and only shows the presence of some amino acid at a location in space, also produces predictions with comparable accuracy. Two possible explanations may be responsible for this behavior. First, the sequence and the corresponding three-dimensional structure may be more strongly correlated with respect to the protein function than thus far appreciated. The correlations seem to be present within each SDR and SAM-MTase protein families. This can be understood since the three-dimensional structures are obtained by applying the AlphaFold2 model to the protein sequence. While this work is focused on naturally occurring protein structures, it might be useful to consider synthetic data sets with a lesser degree of correlation between the structure and sequence in order to explore this phenomenon. Second, the uniform discretization of the three-dimensional structure might be insufficiently sensitive to the active site structure, which would require finer resolution. In order to better explore this hypothesis, multiscale features representation should be explored, which are not based on an encoding of a regular grid in three- (or four-dimensional) space but instead on an adaptive-resolution grid.

In order to better understand the performance of SFC-based feature representations, it is instructive to compare their predictions with the current state-of-the art methods. In a parallel study, we have generated enzyme sequence embeddings of the SDRs and SAM-MTases of our data set using the ESM protein language model [93] with pre-trained weights (ESM-1b) and 1280 length [94]. These sequence embeddings have been used to train a logistic regression model for the same binary classification tasks as in this work. The ESM sequence embeddings demonstrate excellent performance with AUC = 0.98 obtained for the NAD/NADP classification task, compared to AUC = 0.67 using SFC-based feature representations with

logistic regression and AUC = 0.84 with XGBoost binary classification. In the small molecule/biopolymer classification task for SAM-MTases, AUC = 0.99 with ESM sequence embeddings, compared to AUC = 0.78 for SFC-based encodings with logistic regression and AUC = 0.92 with XGBoost. The reader is referred to Ref. [94] for details of the sequence encoding and additional results. Within the logistic regression model, the ESM sequence embeddings outperform SFC-based structure encodings, however, the results of SFC-based feature representations still provide very good predictions in combination with XGBoost binary classification. The comparison of ESM and SFC feature representations is particularly interesting because these models are completely different in their design and implementation. Whereas ESM sequence embeddings utilize protein language models and were generated using unsupervised learning on a very large protein data set [93], the SFC-based feature representations are not data-driven and rely only on a geometric construction. An advantage of SFC-based encodings is that they are reversible and deterministic and thus allow for a bidirectional transformation between one-dimensional feature vector and the full three-dimensional coordinates and residue types. In this aspect, they complement the existing toolbox of protein feature representations.

Supporting Information

Protein sequences, AlphaFold2-generated three-dimensional structures, substrate structures and enzyme clusters, and labels for classification tasks. Accuracy of classification tasks using different sets of encoding parameters and classification algorithms. Training and test sizes for the classification tasks. Hyperparameter optimization of the XGBoost algorithm.

Data and Software Availability

The compiled benchmark data are available from Zenodo online repository using Creative Commons 4.0 license [99]. Substructure definitions, labels for classification tasks, and substrate cluster definitions are in the SI. The structure encoding algorithm is implemented in the molz Python library and is released under the

Apache 2 open-source license [133]. The implementation of the classification tasks and the evaluation of their accuracy using the open-source scikit-learn package [132] are reported in the SI.

Acknowledgements

DR was supported by the National Science Foundation under Grant No. CHE-2227112. AJ would like to thank the Howard Hughes Medical Institute's (HHMI) Hanna Gray Fellowship for support. The authors are grateful to Dr. Benjamin Sanchez-Lengeling for his assistance with structural clustering of enzyme substrates.

References

(1) UniProt Consortium. UniProt: The Universal Protein Knowledgebase in 2021. *Nucleic Acids Res.* **2021**, *49* (D1), D480–D489. https://doi.org/10.1093/nar/gkaa1100.

(2) Rhee, K. Y.; Jansen, R. S.; Grundner, C. Activity-Based Annotation: The Emergence of Systems Biochemistry. *Trends Biochem. Sci.* **2022**. https://doi.org/10.1016/j.tibs.2022.03.017.

(3) Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G. R.; Wang, J.; Cong, Q.; Kinch, L. N.; Schaeffer, R. D.; Millán, C.; Park, H.; Adams, C.; Glassman, C. R.; DeGiovanni, A.; Pereira, J. H.; Rodrigues, A. V.; van Dijk, A. A.; Ebrecht, A. C.; Opperman, D. J.; Sagmeister, T.; Buhlheller, C.; Pavkov-Keller, T.; Rathinaswamy, M. K.; Dalwadi, U.; Yip, C. K.; Burke, J. E.; Garcia, K. C.; Grishin, N. V.; Adams, P. D.; Read, R. J.; Baker, D. Accurate Prediction of Protein Structures and Interactions Using a Three-Track Neural Network. *Science* **2021**, *373* (6557), 871–876. https://doi.org/10.1126/science.abj8754.

(4) 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. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2.

(5) Torng, W.; Altman, R. B. 3D Deep Convolutional Neural Networks for Amino Acid Environment Similarity Analysis. *BMC Bioinformatics* **2017**, *18* (1), 302. https://doi.org/10.1186/s12859-017-1702-0.

(6) Torng, W.; Altman, R. B. High Precision Protein Functional Site Detection Using 3D Convolutional Neural Networks. *Bioinformatics* **2019**, *35* (9), 1503–1512. https://doi.org/10.1093/bioinformatics/bty813.

(7) Amidi, A.; Amidi, S.; Vlachakis, D.; Megalooikonomou, V.; Paragios, N.; Zacharaki, E. I. EnzyNet: Enzyme Classification Using 3D Convolutional Neural Networks on Spatial Representation. *PeerJ* **2018**, *6*, e4750. https://doi.org/10.7717/peerj.4750.

(8) Bagley, S. C.; Altman, R. B. Characterizing the Microenvironment Surrounding Protein Sites. *Protein Sci.* **1995**, *4* (4), 622–635. https://doi.org/10.1002/pro.5560040404.

(9) 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. Structure-Based Protein Function Prediction Using Graph Convolutional Networks. *Nat. Commun.* **2021**, *12* (1), 3168. https://doi.org/10.1038/s41467-021-23303-9.

(10) Mills, C. L.; Garg, R.; Lee, J. S.; Tian, L.; Suciu, A.; Cooperman, G. D.; Beuning, P. J.; Ondrechen, M. J. Functional Classification of Protein Structures by Local Structure Matching in Graph Representation. *Protein Sci.* **2018**, *27* (6), 1125–1135. https://doi.org/10.1002/pro.3416.

(11) Zacharaki, E. I. Prediction of Protein Function Using a Deep Convolutional Neural Network Ensemble. *PeerJ Comput. Sci.* **2017**, *3*, e124. https://doi.org/10.7717/peerj-cs.124.

(12) Göbel, U.; Sander, C.; Schneider, R.; Valencia, A. Correlated Mutations and Residue Contacts in Proteins. *Proteins* **1994**, *18* (4), 309–317. https://doi.org/10.1002/prot.340180402.

(13) Sobolev, V.; Sorokine, A.; Prilusky, J.; Abola, E. E.; Edelman, M. Automated Analysis of Interatomic Contacts in Proteins. *Bioinformatics* **1999**, *15* (4), 327–332. https://doi.org/10.1093/bioinformatics/15.4.327.

(14) Morcos, F.; Pagnani, A.; Lunt, B.; Bertolino, A.; Marks, D. S.; Sander, C.; Zecchina, R.; Onuchic, J. N.; Hwa, T.; Weigt, M. Direct-Coupling Analysis of Residue Coevolution Captures Native Contacts across Many Protein Families. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108* (49), E1293-301. https://doi.org/10.1073/pnas.1111471108.

(15) Shindyalov, I. N.; Kolchanov, N. A.; Sander, C. Can Three-Dimensional Contacts in Protein

Structures Be Predicted by Analysis of Correlated Mutations? *Protein Eng.* **1994**, *7* (3), 349–358. https://doi.org/10.1093/protein/7.3.349.
(16) Lockless, S. W.; Ranganathan, R. Evolutionarily Conserved Pathways of Energetic Connectivity in Protein Families. *Science* **1999**, *286* (5438), 295–299. https://doi.org/10.1126/science.286.5438.295.
(17) Burger, L.; van Nimwegen, E. Disentangling Direct from Indirect Co-Evolution of Residues in Protein Alignments. *PLoS Comput. Biol.* **2010**, *6* (1), e1000633. https://doi.org/10.1371/journal.pcbi.1000633.
(18) Goodfellow, I.; Bengio, Y.; Courville, A. Convolutional Networks. In *Deepe Learning*; MIT Press: Cambridge MA, 2016; pp 321–361.
(19) O'Shea, K.; Nash, R. An Introduction to Convolutional Neural Networks. *arXiv [cs.NE]*, 2015. http://arxiv.org/abs/1511.08458.
(20) Friedman, J. H. Greedy Function Approximation: A Gradient Boosting Machine. *aos* **2001**, *29* (5), 1189–1232. https://doi.org/10.1214/aos/1013203451.
(21) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; KDD '16; Association for Computing Machinery: New York, 2016; pp 785–794. https://doi.org/10.1145/2939672.2939785.
(22) Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems*; von Luxburg, U., Guyon, I., Bengio, S., Wallach, H., Fergus, R., Eds.; 2017; Vol. 30, pp 3149–3157.
(23) Schölkopf, B.; Smola, A. J. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*; MIT Press: Cambridge MA, 2002.
(24) Smola, A. J.; Schölkopf, B. A Tutorial on Support Vector Regression. *Stat. Comput.* **2004**, *14* (3), 199–222. https://doi.org/10.1023/B:STCO.0000035301.49549.88.
(25) Vapnik, V. *The Nature of Statistical Learning Theory*; Springer: New York, 2010. https://doi.org/10.1007/978-1-4757-3264-1.
(26) Indyk, P.; Motwani, R. Approximate Nearest Neighbors: Towards Removing the Curse of Dimensionality. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing*; STOC '98; Association for Computing Machinery: New York, NY, USA, 1998; pp 604–613. https://doi.org/10.1145/276698.276876.
(27) Indyk, P. Nearest Neighbors in High-Dimensional Spaces. In *Handbook of Discrete and Computational Geometry, Second Edition*; Chapman and Hall/CRC, 2004. https://doi.org/10.1201/9781420035315.ch39.
(28) Indyk, P.; Matoušek, J.; Sidiropoulos, A. Low-distortion embeddings of finite metric spaces. In *Handbook of Discrete and Computational Geometry*; Goodman, J. E., O'Rourke, J., Tóth, C. D., Eds.; Chapman and Hall/CRC: Boca Raton FL, 11 2017; pp 211–230.
(29) Falconer, K. J. *Fractal Geometry: Mathematical Foundations and Applications*; Wiley, 2003. https://doi.org/10.1002/0470013850.
(30) Sagan, H. *Space-Filling Curves*; Springer: New York, 1994. https://doi.org/10.1007/978-1-4612-0871-6.
(31) Bader, M. *Space-Filling Curves: An Introduction With Applications in Scientific Computing*; Springer: Berlin Heidelberg, 2012. https://doi.org/10.1007/978-3-642-31046-1.
(32) Peano, G. Sur Une Courbe, Qui Remplit Toute Une Aire Plane. *Math. Ann.* **1890**, *36* (1), 157–160. https://doi.org/10.1007/BF01199438.
(33) Hilbert, D. Ueber die stetige Abbildung einer Line auf ein Flächenstück. *Math. Ann.* **1891**, *38* (3), 459–460. https://doi.org/10.1007/bf01199431.
(34) Morton, G. M. *A Computer Oriented Geodetic Data Base and a New Technique in File Sequencing*; IBM Canada, 1966. https://dominoweb.draco.res.ibm.com/0dabf9473b9c86d48525779800566a39.html.
(35) Butz, A. R. Space Filling Curves and Mathematical Programming. *Inf. Control* **1968**, *12* (4), 314–330. https://doi.org/10.1016/S0019-9958(68)90367-7.

(36) Butz, A. R. Convergence with Hilbert's space filling curve. *J. Comput. System Sci.* **1969**, *3* (2), 128–146. https://doi.org/10.1016/S0022-0000(69)80010-3.
(37) Bartholdi, J. J.; Platzman, L. K. Heuristics Based on Spacefilling Curves for Combinatorial Problems in Euclidean Space. *Manage. Sci.* **1988**, *34* (3), 291–305.
(38) Platzman, L. K.; Bartholdi, J. J. Spacefilling Curves and the Planar Travelling Salesman Problem. *J. ACM* **1989**, *36* (4), 719–737. https://doi.org/10.1145/76359.76361.
(39) Gaede, V.; Günther, O. Multidimensional Access Methods. *ACM Comput. Surv.* **1998**, *30* (2), 170–231. https://doi.org/10.1145/280277.280279.
(40) Böhm, C.; Berchtold, S.; Keim, D. A. Searching in High-Dimensional Spaces: Index Structures for Improving the Performance of Multimedia Databases. *ACM Comput. Surv.* **2001**, *33* (3), 322–373. https://doi.org/10.1145/502807.502809.
(41) Griebel, M.; Knapek, S.; Zumbusch, G. *Numerical Simulation in Molecular Dynamics: Numerics, Algorithms, Parallelization, Applications*; Springer: Berlin Heidelberg, 2007. https://doi.org/10.1007/978-3-540-68095-6.
(42) Behrens, J. *Adaptive Atmospheric Modeling: Key Techniques in Grid Generation, Data Structures, and Numerical Operations with Applications*; Springer: Berlin Heidelberg, 2006. https://doi.org/10.1007/3-540-33383-5.
(43) Chatterjee, S.; Lebeck, A. R.; Patnala, P. K.; Thottethodi, M. Recursive Array Layouts and Fast Matrix Multiplication. *IEEE Trans. Parallel Distrib. Syst.* **2002**, *13* (11), 1105–1123. https://doi.org/10.1109/TPDS.2002.1058095.
(44) Bader, M.; Zenger, C. Cache Oblivious Matrix Multiplication Using an Element Ordering Based on a Peano Curve. *Linear Algebra Appl.* **2006**, *417* (2), 301–313. https://doi.org/10.1016/j.laa.2006.03.018.
(45) Gotsman, C.; Lindenbaum, M. On the metric properties of discrete space-filling curves. *IEEE Trans. Image Process.* **1996**, *5* (5), 794–797. https://doi.org/10.1109/83.499920.
(46) Haverkort, H.; van Walderveen, F. Locality and bounding-box quality of two-dimensional space-filling curves. *Comput. Geom.* **2010**, *43* (2), 131–147. https://doi.org/10.1016/j.comgeo.2009.06.002.
(47) Radivojac, P.; Clark, W. T.; Oron, T. R.; Schnoes, A. M.; Wittkop, T.; Sokolov, A.; Graim, K.; Funk, C.; Verspoor, K.; Ben-Hur, A.; Pandey, G.; Yunes, J. M.; Talwalkar, A. S.; Repo, S.; Souza, M. L.; Piovesan, D.; Casadio, R.; Wang, Z.; Cheng, J.; Fang, H.; Gough, J.; Koskinen, P.; Törönen, P.; Nokso-Koivisto, J.; Holm, L.; Cozzetto, D.; Buchan, D. W. A.; Bryson, K.; Jones, D. T.; Limaye, B.; Inamdar, H.; Datta, A.; Manjari, S. K.; Joshi, R.; Chitale, M.; Kihara, D.; Lisewski, A. M.; Erdin, S.; Venner, E.; Lichtarge, O.; Rentzsch, R.; Yang, H.; Romero, A. E.; Bhat, P.; Paccanaro, A.; Hamp, T.; Kaßner, R.; Seemayer, S.; Vicedo, E.; Schaefer, C.; Achten, D.; Auer, F.; Boehm, A.; Braun, T.; Hecht, M.; Heron, M.; Hönigschmid, P.; Hopf, T. A.; Kaufmann, S.; Kiening, M.; Krompass, D.; Landerer, C.; Mahlich, Y.; Roos, M.; Björne, J.; Salakoski, T.; Wong, A.; Shatkay, H.; Gatzmann, F.; Sommer, I.; Wass, M. N.; Sternberg, M. J. E.; Škunca, N.; Supek, F.; Bošnjak, M.; Panov, P.; Džeroski, S.; Šmuc, T.; Kourmpetis, Y. A. I.; van Dijk, A. D. J.; ter Braak, C. J. F.; Zhou, Y.; Gong, Q.; Dong, X.; Tian, W.; Falda, M.; Fontana, P.; Lavezzo, E.; Di Camillo, B.; Toppo, S.; Lan, L.; Djuric, N.; Guo, Y.; Vucetic, S.; Bairoch, A.; Linial, M.; Babbitt, P. C.; Brenner, S. E.; Orengo, C.; Rost, B.; Mooney, S. D.; Friedberg, I. A Large-Scale Evaluation of Computational Protein Function Prediction. *Nat. Methods* **2013**, *10* (3), 221–227. https://doi.org/10.1038/nmeth.2340.
(48) Kim, G. B.; Kim, W. J.; Kim, H. U.; Lee, S. Y. Machine Learning Applications in Systems Metabolic Engineering. *Curr. Opin. Biotechnol.* **2020**, *64*, 1–9. https://doi.org/10.1016/j.copbio.2019.08.010.
(49) Faulon, J.-L.; Faure, L. In Silico, in Vitro, and in Vivo Machine Learning in Synthetic Biology and Metabolic Engineering. *Curr. Opin. Chem. Biol.* **2021**, *65*, 85–92. https://doi.org/10.1016/j.cbpa.2021.06.002.
(50) Carbonell, P.; Radivojevic, T.; García Martín, H. Opportunities at the Intersection of Synthetic Biology, Machine Learning, and Automation. *ACS Synth. Biol.* **2019**, *8* (7), 1474–1477.

https://doi.org/10.1021/acssynbio.8b00540.
(51) Patra, P.; B R, D.; Kundu, P.; Das, M.; Ghosh, A. Recent Advances in Machine Learning Applications in Metabolic Engineering. *Biotechnol. Adv.* **2023**, *62*, 108069. https://doi.org/10.1016/j.biotechadv.2022.108069.
(52) Jang, W. D.; Kim, G. B.; Kim, Y.; Lee, S. Y. Applications of Artificial Intelligence to Enzyme and Pathway Design for Metabolic Engineering. *Curr. Opin. Biotechnol.* **2022**, *73*, 101–107. https://doi.org/10.1016/j.copbio.2021.07.024.
(53) Roy, A.; Yang, J.; Zhang, Y. COFACTOR: An Accurate Comparative Algorithm for Structure-Based Protein Function Annotation. *Nucleic Acids Res.* **2012**, *40* (Web Server issue), W471-7. https://doi.org/10.1093/nar/gks372.
(54) Dobson, P. D.; Doig, A. J. Predicting Enzyme Class from Protein Structure without Alignments. *J. Mol. Biol.* **2005**, *345* (1), 187–199. https://doi.org/10.1016/j.jmb.2004.10.024.
(55) Borro, L. C.; Oliveira, S. R. M.; Yamagishi, M. E. B.; Mancini, A. L.; Jardine, J. G.; Mazoni, I.; Santos, E. H. dos; Higa, R. H.; Kuser, P. R.; Neshich, G. Predicting Enzyme Class from Protein Structure Using Bayesian Classification. *Genet. Mol. Res.* **2006**, *5* (1), 193–202.
(56) Dalkiran, A.; Rifaioglu, A. S.; Martin, M. J.; Cetin-Atalay, R.; Atalay, V.; Doğan, T. ECPred: A Tool for the Prediction of the Enzymatic Functions of Protein Sequences Based on the EC Nomenclature. *BMC Bioinformatics*. 2018. https://doi.org/10.1186/s12859-018-2368-y.
(57) Nagao, C.; Nagano, N.; Mizuguchi, K. Prediction of Detailed Enzyme Functions and Identification of Specificity Determining Residues by Random Forests. *PLoS One* **2014**, *9* (1), e84623. https://doi.org/10.1371/journal.pone.0084623.
(58) Jensen, L. J.; Gupta, R.; Blom, N.; Devos, D.; Tamames, J.; Kesmir, C.; Nielsen, H.; Staerfeldt, H. H.; Rapacki, K.; Workman, C.; Andersen, C. A. F.; Knudsen, S.; Krogh, A.; Valencia, A.; Brunak, S. Prediction of Human Protein Function from Post-Translational Modifications and Localization Features. *J. Mol. Biol.* **2002**, *319* (5), 1257–1265. https://doi.org/10.1016/S0022-2836(02)00379-0.
(59) Qiu, J.-D.; Luo, S.-H.; Huang, J.-H.; Liang, R.-P. Using Support Vector Machines to Distinguish Enzymes: Approached by Incorporating Wavelet Transform. *J. Theor. Biol.* **2009**, *256* (4), 625–631. https://doi.org/10.1016/j.jtbi.2008.10.026.
(60) Davidson, N. J.; Wang, X. Non-Alignment Features Based Enzyme/Non-Enzyme Classification Using an Ensemble Method. *Proc. Int. Conf. Mach. Learn. Appl.* **2010**, 546–551. https://doi.org/10.1109/ICMLA.2010.167.
(61) Wang, Y.-C.; Wang, X.-B.; Yang, Z.-X.; Deng, N.-Y. Prediction of Enzyme Subfamily Class via Pseudo Amino Acid Composition by Incorporating the Conjoint Triad Feature. *Protein Pept. Lett.* **2010**, *17* (11), 1441–1449. https://doi.org/10.2174/0929866511009011441.
(62) Wang, Y.-C.; Wang, Y.; Yang, Z.-X.; Deng, N.-Y. Support Vector Machine Prediction of Enzyme Function with Conjoint Triad Feature and Hierarchical Context. *BMC Syst. Biol.* **2011**, *5 Suppl 1*, S6. https://doi.org/10.1186/1752-0509-5-S1-S6.
(63) Kumar, C.; Choudhary, A. A Top-down Approach to Classify Enzyme Functional Classes and Sub-Classes Using Random Forest. *EURASIP J. Bioinform. Syst. Biol.* **2012**, *2012* (1), 1. https://doi.org/10.1186/1687-4153-2012-1.
(64) Kumar, N.; Skolnick, J. EFICAz2.5: Application of a High-Precision Enzyme Function Predictor to 396 Proteomes. *Bioinformatics* **2012**, *28* (20), 2687–2688. https://doi.org/10.1093/bioinformatics/bts510.
(65) De Ferrari, L.; Aitken, S.; van Hemert, J.; Goryanin, I. EnzML: Multi-Label Prediction of Enzyme Classes Using InterPro Signatures. *BMC Bioinformatics* **2012**, *13*, 61. https://doi.org/10.1186/1471-2105-13-61.
(66) Volpato, V.; Adelfio, A.; Pollastri, G. Accurate Prediction of Protein Enzymatic Class by N-to-1 Neural Networks. *BMC Bioinformatics* **2013**, *14 Suppl 1*, S11. https://doi.org/10.1186/1471-2105-14-S1-S11.
(67) Matsuta, Y.; Ito, M.; Tohsato, Y. ECOH: An Enzyme Commission Number Predictor Using Mutual Information and a Support Vector Machine. *Bioinformatics*. 2013, pp 365–372.

https://doi.org/10.1093/bioinformatics/bts700.
(68) Che, Y.; Ju, Y.; Xuan, P.; Long, R.; Xing, F. Identification of Multi-Functional Enzyme with Multi-Label Classifier. *PLoS One* **2016**, *11* (4), e0153503. https://doi.org/10.1371/journal.pone.0153503.
(69) Li, Y.; Wang, S.; Umarov, R.; Xie, B.; Fan, M.; Li, L.; Gao, X. DEEPre: Sequence-Based Enzyme EC Number Prediction by Deep Learning. *Bioinformatics* **2018**, *34* (5), 760–769. https://doi.org/10.1093/bioinformatics/btx680.
(70) Zhou, X.-B.; Chen, C.; Li, Z.-C.; Zou, X.-Y. Using Chou's Amphiphilic Pseudo-Amino Acid Composition and Support Vector Machine for Prediction of Enzyme Subfamily Classes. *Journal of Theoretical Biology*. 2007, pp 546–551. https://doi.org/10.1016/j.jtbi.2007.06.001.
(71) Shen, H.-B.; Chou, K.-C. EzyPred: A Top-down Approach for Predicting Enzyme Functional Classes and Subclasses. *Biochem. Biophys. Res. Commun.* **2007**, *364* (1), 53–59. https://doi.org/10.1016/j.bbrc.2007.09.098.
(72) Lu, L.; Qian, Z.; Cai, Y.-D.; Li, Y. ECS: An Automatic Enzyme Classifier Based on Functional Domain Composition. *Comput. Biol. Chem.* **2007**, *31* (3), 226–232. https://doi.org/10.1016/j.compbiolchem.2007.03.008.
(73) Huang, W.-L.; Chen, H.-M.; Hwang, S.-F.; Ho, S.-Y. Accurate Prediction of Enzyme Subfamily Class Using an Adaptive Fuzzy K-Nearest Neighbor Method. *Biosystems.* **2007**, *90* (2), 405–413. https://doi.org/10.1016/j.biosystems.2006.10.004.
(74) Nasibov, E.; Kandemir-Cavas, C. Efficiency Analysis of KNN and Minimum Distance-Based Classifiers in Enzyme Family Prediction. *Comput. Biol. Chem.* **2009**, *33* (6), 461–464. https://doi.org/10.1016/j.compbiolchem.2009.09.002.
(75) Zhang, T.; Tian, Y.; Yuan, L.; Chen, F.; Ren, A.; Hu, Q.-N. Bio2Rxn: Sequence-Based Enzymatic Reaction Predictions by a Consensus Strategy. *Bioinformatics* **2020**, *36* (11), 3600–3601. https://doi.org/10.1093/bioinformatics/btaa135.
(76) Watanabe, N.; Murata, M.; Ogawa, T.; Vavricka, C. J.; Kondo, A.; Ogino, C.; Araki, M. Exploration and Evaluation of Machine Learning-Based Models for Predicting Enzymatic Reactions. *J. Chem. Inf. Model.* **2020**, *60* (3), 1833–1843. https://doi.org/10.1021/acs.jcim.9b00877.
(77) Pathak, A.; Jayaram, B. Seq2Enz: An Application of Mask BLAST Methodology with a New Chemical Logic of Amino Acids for Improved Enzyme Function Prediction. *Biochim. Biophys. Acta: Proteins Proteomics* **2022**, *1870* (1), 140721. https://doi.org/10.1016/j.bbapap.2021.140721.
(78) Khan, K. A.; Memon, S. A.; Naveed, H. A Hierarchical Deep Learning Based Approach for Multi-functional Enzyme Classification. *Protein Science*. 2021, pp 1935–1945. https://doi.org/10.1002/pro.4146.
(79) Sequeira, A. M.; Rocha, M. Recurrent Deep Neural Networks for Enzyme Functional Annotation. *Practical Applications of Computational Biology & Bioinformatics, 15th International Conference (PACBB 2021)*. 2022, pp 62–73. https://doi.org/10.1007/978-3-030-86258-9_7.
(80) Ryu, J. Y.; Kim, H. U.; Lee, S. Y. Deep Learning Enables High-Quality and High-Throughput Prediction of Enzyme Commission Numbers. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116* (28), 13996–14001. https://doi.org/10.1073/pnas.1821905116.
(81) Concu, R.; Cordeiro, M. N. D. S. Alignment-Free Method to Predict Enzyme Classes and Subclasses. *Int. J. Mol. Sci.* **2019**, *20* (21). https://doi.org/10.3390/ijms20215389.
(82) Pradhan, D.; Sahoo, B.; Misra, B. B.; Padhy, S. A Multiclass SVM Classifier with Teaching Learning Based Feature Subset Selection for Enzyme Subclass Classification. *Applied Soft Computing*. 2020, p 106664. https://doi.org/10.1016/j.asoc.2020.106664.
(83) Memon, S. A.; Khan, K. A.; Naveed, H. HECNet: A Hierarchical Approach to Enzyme Function Classification Using a Siamese Triplet Network. *Bioinformatics* **2020**, *36* (17), 4583–4589. https://doi.org/10.1093/bioinformatics/btaa536.
(84) Visani, G. M.; Hughes, M. C.; Hassoun, S. Enzyme Promiscuity Prediction Using Hierarchy-Informed Multi-Label Classification. *Bioinformatics* **2021**. https://doi.org/10.1093/bioinformatics/btab054.
(85) Semwal, R.; Aier, I.; Tyagi, P.; Varadwaj, P. K. DeEPn: A Deep Neural Network Based Tool for

Enzyme Functional Annotation. *J. Biomol. Struct. Dyn.* **2021**, *39* (8), 2733–2743. https://doi.org/10.1080/07391102.2020.1754292.
(86) Sarker, B.; Ritchie, D. W.; Aridhi, S. GrAPFI: Predicting Enzymatic Function of Proteins from Domain Similarity Graphs. *BMC Bioinformatics* **2020**, *21* (1), 168. https://doi.org/10.1186/s12859-020-3460-7.
(87) Shahraki, M. F.; Atanaki, F. F.; Ariaeenejad, S.; Ghaffari, M. R.; Norouzi-Beirami, M. H.; Maleki, M.; Salekdeh, G. H.; Kavousi, K. A Computational Learning Paradigm to Targeted Discovery of Biocatalysts from Metagenomic Data: A Case Study of Lipase Identification. *Biotechnology and Bioengineering*. 2022, pp 1115–1128. https://doi.org/10.1002/bit.28037.
(88) Duhan, N.; Norton, J. M.; Kaundal, R. DeepNEC: A Novel Alignment-Free Tool for the Identification and Classification of Nitrogen Biochemical Network-Related Enzymes Using Deep Learning. *Brief. Bioinform.* **2022**, *23* (3). https://doi.org/10.1093/bib/bbac071.
(89) Nallapareddy, M. V.; Dwivedula, R. ABLE: Attention Based Learning for Enzyme Classification. *Comput. Biol. Chem.* **2021**, *94*, 107558. https://doi.org/10.1016/j.compbiolchem.2021.107558.
(90) Baldazzi, D.; Savojardo, C.; Martelli, P. L.; Casadio, R. BENZ WS: The Bologna ENZyme Web Server for Four-Level EC Number Annotation. *Nucleic Acids Res.* **2021**, *49* (W1), W60–W66. https://doi.org/10.1093/nar/gkab328.
(91) Strodthoff, N.; Wagner, P.; Wenzel, M.; Samek, W. UDSMProt: Universal Deep Sequence Models for Protein Classification. *Bioinformatics* **2020**, *36* (8), 2401–2409. https://doi.org/10.1093/bioinformatics/btaa003.
(92) Blum, M.; Chang, H.-Y.; Chuguransky, S.; Grego, T.; Kandasaamy, S.; Mitchell, A.; Nuka, G.; Paysan-Lafosse, T.; Qureshi, M.; Raj, S.; Richardson, L.; Salazar, G. A.; Williams, L.; Bork, P.; Bridge, A.; Gough, J.; Haft, D. H.; Letunic, I.; Marchler-Bauer, A.; Mi, H.; Natale, D. A.; Necci, M.; Orengo, C. A.; Pandurangan, A. P.; Rivoire, C.; Sigrist, C. J. A.; Sillitoe, I.; Thanki, N.; Thomas, P. D.; Tosatto, S. C. E.; Wu, C. H.; Bateman, A.; Finn, R. D. The InterPro Protein Families and Domains Database: 20 Years On. *Nucleic Acids Res.* **2021**, *49* (D1), D344–D354. https://doi.org/10.1093/nar/gkaa977.
(93) Rives, A.; Meier, J.; Sercu, T.; Goyal, S.; Lin, Z.; Liu, J.; Guo, D.; Ott, M.; Zitnick, C. L.; Ma, J.; Fergus, R. Biological Structure and Function Emerge from Scaling Unsupervised Learning to 250 Million Protein Sequences. *Proc. Natl. Acad. Sci. U. S. A.* **2021**, *118* (15), e2016239118. https://doi.org/10.1073/pnas.2016239118.
(94) Jinich, A.; Nazia, S. Z.; Tellez, A. V.; Rappoport, D.; AlQuraishi, M.; Rhee, K. Y. Predicting Enzyme Substrate Chemical Structure with Protein Language Models. *bioRxiv*, 2022, 2022.09.28.509940. https://doi.org/10.1101/2022.09.28.509940.
(95) UniProt Consortium. UniProt: A Hub for Protein Information. *Nucleic Acids Res.* **2015**, *43* (Database issue), D204-12. https://doi.org/10.1093/nar/gku989.
(96) Bansal, P.; Morgat, A.; Axelsen, K. B.; Muthukrishnan, V.; Coudert, E.; Aimo, L.; Hyka-Nouspikel, N.; Gasteiger, E.; Kerhornou, A.; Neto, T. B.; Pozzato, M.; Blatter, M.-C.; Ignatchenko, A.; Redaschi, N.; Bridge, A. Rhea, the Reaction Knowledgebase in 2022. *Nucleic Acids Res.* **2022**, *50* (D1), D693–D700. https://doi.org/10.1093/nar/gkab1016.
(97) Hastings, J.; Owen, G.; Dekker, A.; Ennis, M.; Kale, N.; Muthukrishnan, V.; Turner, S.; Swainston, N.; Mendes, P.; Steinbeck, C. ChEBI in 2016: Improved Services and an Expanding Collection of Metabolites. *Nucleic Acids Res.* **2016**, *44* (D1), D1214-9. https://doi.org/10.1093/nar/gkv1031.
(98) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28* (1), 31–36. https://doi.org/10.1021/ci00057a005.
(99) Jinich, A.; Rappoport, D. *Enzyme Substrate Classification Dataset for SDRs and SAM-MTases*; 2022. https://doi.org/10.5281/zenodo.7141435.
(100) Landrum, G.; co-authors. *RDKit: Open-Source Cheminformatics Software*; 2022. https://doi.org/10.5281/zenodo.6388425.
(101) Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures-A

Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5* (2), 107–113. https://doi.org/10.1021/c160017a018.
(102) McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv [stat.ML]*, 2018. http://arxiv.org/abs/1802.03426.
(103) MacQueen, J. Some Methods for Classification and Analysis of Multivariate Observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*; University of California Press: Berkeley, CA, 1967; Vol. 5.1, pp 281–298.
(104) Lloyd, S. Least Squares Quantization in PCM. *IEEE Trans. Inf. Theory* **1982**, *28* (2), 129–137. https://doi.org/10.1109/TIT.1982.1056489.
(105) Kaufman, L.; Rousseeuw, P. J. *Finding Groups in Data: An Introduction to Cluster Analysis*; Wiley: New York, 1990. https://doi.org/10.1002/9780470316801.
(106) Daylight Chemical Information Inc. *SMARTS - A Language for Describing Molecular Patterns*. https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html (accessed 2022-07-08).
(107) Mirdita, M.; Schütze, K.; Moriwaki, Y.; Heo, L.; Ovchinnikov, S.; Steinegger, M. ColabFold - Making Protein Folding Accessible to All. *bioRxiv*, 2022, 2021.08.15.456425. https://doi.org/10.1101/2021.08.15.456425.
(108) Steinegger, M.; Söding, J. MMseqs2 Enables Sensitive Protein Sequence Searching for the Analysis of Massive Data Sets. *Nat. Biotechnol.* **2017**, *35* (11), 1026–1028. https://doi.org/10.1038/nbt.3988.
(109) Mirdita, M.; Steinegger, M.; Söding, J. MMseqs2 Desktop and Local Web Server App for Fast, Interactive Sequence Searches. *Bioinformatics* **2019**, *35* (16), 2856–2858. https://doi.org/10.1093/bioinformatics/bty1057.
(110) Koehl, P. Protein Structure Classification. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Cundari, T. R., Gillet, V. J., Boyd, D. B., Eds.; Wiley: Hoboken NJ, 2006; Vol. 22, pp 1–55. https://doi.org/10.1002/0471780367.ch1.
(111) Kmiecik, S.; Gront, D.; Kolinski, M.; Wieteska, L.; Dawid, A. E.; Kolinski, A. Coarse-Grained Protein Models and Their Applications. *Chem. Rev.* **2016**, *116* (14), 7898–7936. https://doi.org/10.1021/acs.chemrev.6b00163.
(112) M. O. Dayhoff, R. M. Schwartz, B. C. Orcutt. A Model of Evolutionary Change in Proteins. In *Atlas of Protein Sequence and Structure*; Dayhoff, M. O., Ed.; National Biomedical Research Foundation: Washington DC, 1978; Vol. 5, Suppl. 3, pp 345–352.
(113) Schwartz, R. M.; Dayhoff, M. O. Matrices for Detecting Distant Relationships. In *Atlas of Protein Sequence and Structure*; Dayhoff, M. O., Ed.; National Biomedical Research Foundation: Washington DC, 1978; Vol. 5, Suppl. 3, pp 353–358.
(114) Swanson, R. A Vector Representation for Amino Acid Sequences. *Bull. Math. Biol.* **1984**, *46* (4), 623–639. https://doi.org/10.1007/BF02459507.
(115) Henikoff, S.; Henikoff, J. G. Amino Acid Substitution Matrices from Protein Blocks. *Proc. Natl. Acad. Sci. U. S. A.* **1992**, *89* (22), 10915–10919. https://doi.org/10.1073/pnas.89.22.10915.
(116) Maetschke, S.; Towsey, M.; Bodén, M. Blomap: An Encoding of Amino Acids Which Improves Signal Peptide Cleavage Site Prediction. In *Proceedings of the 3rd Asia-Pacific Bioinformatics Conference*; World Scientific: Singapore, 2005. https://doi.org/10.1142/9781860947322_0014.
(117) Li, J.; Koehl, P. 3D Representations of Amino Acids-Applications to Protein Sequence Comparison and Classification. *Comput. Struct. Biotechnol. J.* **2014**, *11* (18), 47–58. https://doi.org/10.1016/j.csbj.2014.09.001.
(118) Koehl, P.; Orland, H.; Delarue, M. Numerical Encodings of Amino Acids in Multivariate Gaussian Modeling of Protein Multiple Sequence Alignments. *Molecules* **2018**, *24* (1). https://doi.org/10.3390/molecules24010104.
(119) French, S.; Robson, B. What Is a Conservative Substitution? *J. Mol. Evol.* **1983**, *19* (2), 171–175. https://doi.org/10.1007/BF02300754.
(120) Kidera, A.; Konishi, Y.; Oka, M.; Ooi, T.; Scheraga, H. A. Statistical Analysis of the Physical Properties of the 20 Naturally Occurring Amino Acids. *J. Protein Chem.* **1985**, *4* (1), 23–55.

https://doi.org/10.1007/BF01025492.
(121) He, Y.; Rackovsky, S.; Yin, Y.; Scheraga, H. A. Alternative Approach to Protein Structure Prediction Based on Sequential Similarity of Physical Properties. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112* (16), 5029–5032. https://doi.org/10.1073/pnas.1504806112.
(122) He, Y.; Maisuradze, G. G.; Yin, Y.; Kachlishvili, K.; Rackovsky, S.; Scheraga, H. A. Sequence-, Structure-, and Dynamics-Based Comparisons of Structurally Homologous CheY-like Proteins. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114* (7), 1578–1583. https://doi.org/10.1073/pnas.1621344114.
(123) Zvelebil, M. J.; Barton, G. J.; Taylor, W. R.; Sternberg, M. J. Prediction of Protein Secondary Structure and Active Sites Using the Alignment of Homologous Sequences. *J. Mol. Biol.* **1987**, *195* (4), 957–961. https://doi.org/10.1016/0022-2836(87)90501-8.
(124) Hellberg, S.; Sjöström, M.; Skagerberg, B.; Wold, S. Peptide Quantitative Structure-Activity Relationships, a Multivariate Approach. *J. Med. Chem.* **1987**, *30* (7), 1126–1135. https://doi.org/10.1021/jm00390a003.
(125) Sandberg, M.; Eriksson, L.; Jonsson, J.; Sjöström, M.; Wold, S. New Chemical Descriptors Relevant for the Design of Biologically Active Peptides. A Multivariate Characterization of 87 Amino Acids. *J. Med. Chem.* **1998**, *41* (14), 2481–2491. https://doi.org/10.1021/jm9700575.
(126) Kaushik, R.; Singh, A.; Jayaram, B. Where Informatics Lags Chemistry Leads. *Biochemistry* **2018**, *57* (5), 503–506. https://doi.org/10.1021/acs.biochem.7b01073.
(127) Atchley, W. R.; Zhao, J.; Fernandes, A. D.; Drüke, T. Solving the Protein Sequence Metric Problem. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102* (18), 6395–6400. https://doi.org/10.1073/pnas.0408677102.
(128) Altschul, S. F. Substitution Matrices. *eLS*; Wiley: Chichester, UK, 2013. https://doi.org/10.1002/9780470015902.a0005265.pub3.
(129) Kabsch, W. A Discussion of the Solution for the Best Rotation to Relate Two Sets of Vectors. *Acta Crystallogr. A* **1978**, *34* (5), 827–828. https://doi.org/10.1107/S0567739478001680.
(130) Theobald, D. L. Rapid Calculation of RMSDs Using a Quaternion-Based Characteristic Polynomial. *Acta Crystallogr. A* **2005**, *61* (Pt 4), 478–480. https://doi.org/10.1107/S0108767305015266.
(131) Yershova, A.; Jain, S.; Lavalle, S. M.; Mitchell, J. C. Generating Uniform Incremental Grids on SO(3) Using the Hopf Fibration. *Int. J. Rob. Res.* **2010**, *29* (7), 801–812. https://doi.org/10.1177/0278364909352700.
(132) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12* (85), 2825–2830.
(133) Rappoport, D. *Z-Curve Representations with Molecular 3D Coordinates https://bitbucket.org/rappoport/molz*; 2022.

Figure Captions

Figure 1. *Recursive construction of the two-dimensional Hilbert curve (top) and Morton curve (bottom). Dashed lines indicate subsquare boundaries.*

Figure 2. *Structural clustering of substrates and products of annotated short-chain dehydrogenase / reductase (SDR) enzymes. Each circle represents a unique substrate or product. These compounds were clustered according to their chemical structure by (1) obtaining Morgan fingerprints from their SMILES representation; (2) projecting the Morgan fingerprints to 2-dimensions using UMAP; and (3) grouping compounds using k-means clustering.*

Figure 3. *ROC curves for NAD/NADP classification of SDRs using 8-bit Hilbert SFC with modified LK encoding, 4096 bins, orientation sampling using SOI (s = 72, left), same after random label shuffle (right).*

Figures

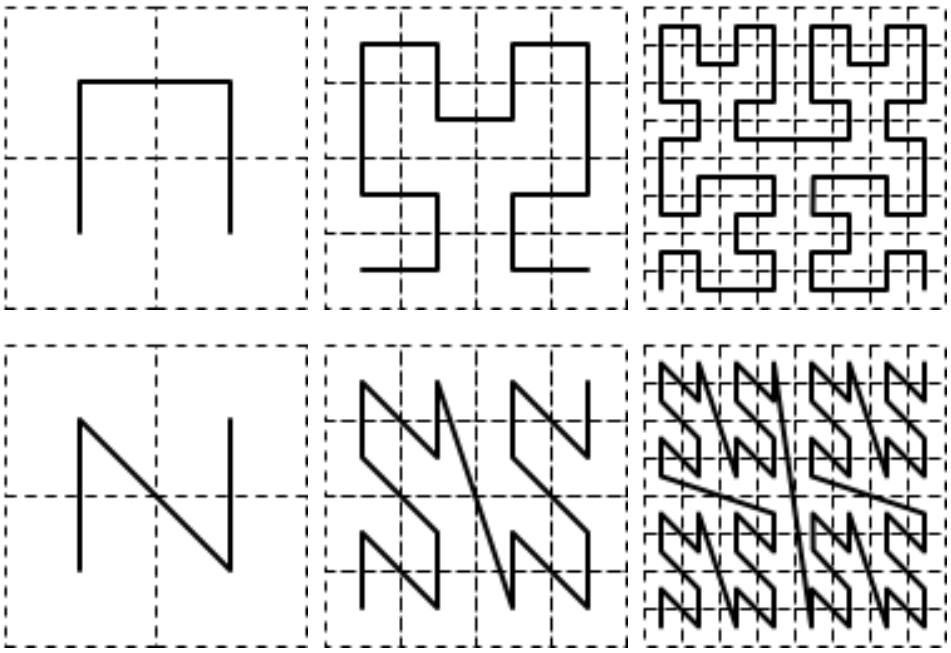


Figure 1

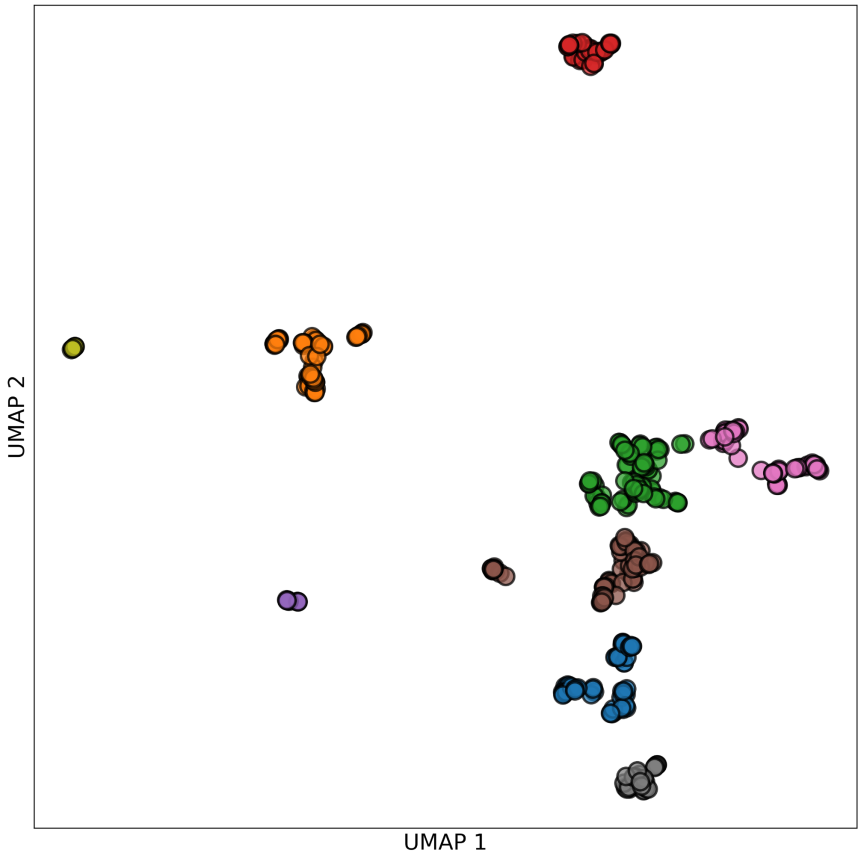


Figure 2

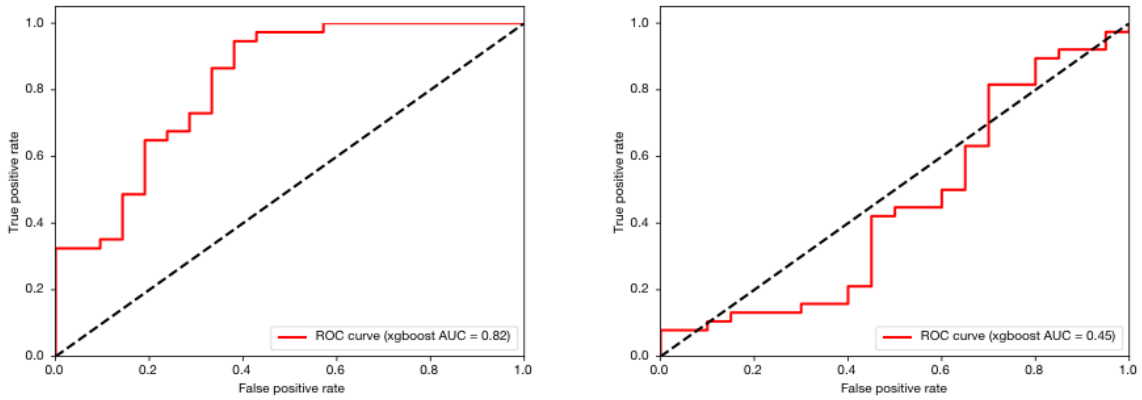
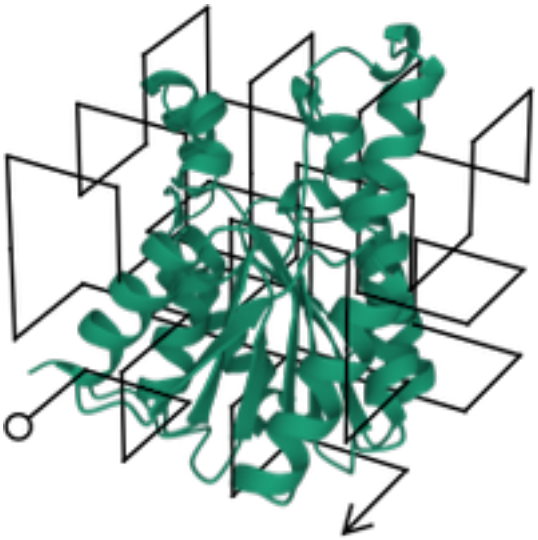


Figure 3

TOC Graphic



This work introduces a novel type of molecular feature representations based on space-filling curves (SFC) that provide discrete, reversible, and interpretable mappings from (3+1)-D protein structure and amino acid type to 1D encodings. Using a new database of enzymatic functional data in two large enzyme families (short-chain dehydrogenase/reductases (SDRs) and S-adenosylmethionine-dependent methyltransferases (SAM-MTases), SFC-based classification models generate predictions that go beyond enzyme categories (EC classes) and yield specific structural and experimentally testable hypotheses.