SCREEN: A Graph-based Contrastive Learning Tool to Infer Catalytic Residues and Assess Enzyme Mutations

Tong Pan[1,2], Yue Bi[1,2], Xiaoyu Wang[1,2], Ying Zhang[1,3], Geoffrey I. Webb[4], Robin B. Gasser[5,*], Lukasz Kurgan[6,*], Jiangning Song[1,2,7,*]

[1]*Monash Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Monash University, Clayton, VIC 3800, Australia*

[2]*Monash Biomedicine Discovery Institute-Wenzhou Medical University Alliance in Clinical and Experimental Biomedicine, Monash University, Clayton, VIC 3800, Australia*

[3]*School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China*

[4]*Department of Data Science and Artificial Intelligence, Monash University, Clayton, VIC 3800, Australia*

[5]*Department of Veterinary Biosciences, Melbourne Veterinary School, The University of Melbourne, Parkville, VIC 3010, Australia*

[6]*Department of Computer Science, Virginia Commonwealth University, Richmond 23284, USA*

[7]*Key Laboratory of Clinical Laboratory Diagnosis and Translational Research of Zhejiang Province, Department of Clinical Laboratory, The First Affiliated Hospital of Wenzhou Medical University, Wenzhou 325015, China*

* Corresponding authors.

Email: robinbg@unimelb.edu.au (Gasser RB), lkurgan@vcu.edu (Kurgan L), jiangning.song@monash.edu (Song J).

Running Title: *Pan T et al / SCREEN for Catalytic Residue Prediction*

The counts of words (8540 words), references (64), tables (1) and figures (6), supplementary figures (15), supplementary tables (3), The counts of letters in article title (90 characters), running title (44 characters), count of keywords (5 words) and words in Abstract (181 words).

Abstract

The accurate identification of catalytic residues contributes to our understanding of enzyme functions in biological processes and pathways. The increasing number of protein sequences necessitates computational tools for the automated prediction of catalytic residues in enzymes.

Here, we introduce SCREEN, a graph neural network for the high-throughput prediction of catalytic residues via the integration of enzyme functional and structural information. SCREEN constructs residue representations based on spatial arrangements and incorporates enzyme function priors into such representations through contrastive learning. We demonstrate that SCREEN (i) consistently outperforms currently-available predictors; (ii) provides accurate results when applied to inferred enzyme structures; and (iii) generalizes well to enzymes dissimilar from those in the training set. We also show that the putative catalytic residues predicted by SCREEN mimic key structural and biophysical characteristics of native catalytic residues. Moreover, using experimental data sets, we show that SCREEN's predictions can be used to distinguish residues with a high mutation tolerance from those likely to cause functional loss when mutated, indicating that this tool might be used to infer disease-associated mutations. SCREEN is publicly available at https://github.com/BioColLab/SCREEN and https://ngdc.cncb.ac.cn/biocode/tool/7580.

Keywords: Catalytic residue; Enzyme structure; Evolutionary conservation; Graph neural network; Contrastive learning

Introduction

Enzymes are critical for a wide range of diverse biochemical, molecular and physiological processes and pathways which sustain life [1]. The extraordinary catalytic proficiency of an enzyme is often intricately orchestrated by a selected set of amino acids within its active site(s), referred to as the catalytic residues [2]. These spatially proximate catalytic residues can be engaged in critical interactions with substrate molecules, catalyzing chemical reactions and ensuring the catalytic efficiency and specificity of enzymes [3]. Catalytic residues often exhibit conservation across species, particularly those within the same taxonomic groups [4], such that mutations in catalytic sites can affect enzyme function(s), potentially inducing the onset of diseases, such as cancers and metabolic disorders [5]. For instance, mutations of catalytic residues in the tumor suppressor phosphatase and tensin homolog (PTEN) have been shown to culminate in various forms of cancers, such as glioblastoma multiforme, melanoma and breast cancer [6]. Mutations in the catalytic sites of enzymes, such as CYP2C9, which are responsible for the biotransformation of small molecule drugs, can impact individual drug responses and

potentially increase the risk of metabolic disorders [7].

The number for which enzymes with detailed catalytic residue annotations available in the Mechanism and Catalytic Site Atlas (M-CSA) database [8] is substantially lower than the vast number of enzyme sequences in the UniProt database [9] and enzyme structures in repositories such as the Protein Data Bank (PDB) [10]. This gap relates primarily to the absence of high-throughput methods for identifying catalytic residues. Traditionally, the active sites in enzymes have been established using site-directed mutagenesis and biochemical assays, providing insight into the corresponding kinetic and thermodynamic parameters [11]. However, these laboratory methods are relatively low throughput, time-consuming and labor-intensive, thereby restricting analyses to small numbers of residues and constraining "down-stream" applications such as the design of novel enzymes and inhibitors [12]. Furthermore, although M-CSA offers information on enzyme functions, its coverage of the function space is not comprehensive, particularly for oxidoreductases and translocases (with only 45.6% and 30% coverage, respectively), potentially attributable to curation backlog and/or limited functional data/information [13]. There is significant demand for an *in silico* approach for the reliable and reproducible identification of catalytic residues in enzymes from sequence and/or structure data to enable the exploration of enzyme functions and accelerate biomolecular design.

A significant effort has been directed towards developing computational methods for the identification of catalytic residues in enzymes [14]. These methods include homology-based approaches, machine learning-based, and deep learning-based approaches. Homology-based methods annotate catalytic residues by comparing the query enzyme's sequence or structural similarity to that of the target enzymes with experimentally validated residues. Wallace et al. firstly introduced TESS [15], a program that uses a geometric hashing algorithm to identify enzyme active sites by aligning the query enzyme to a structural template database. Mistry et al. later proposed a sequence-based method that transfers previously verified catalytic residues to other chains within the same Pfam family using a set of strict rules [16]. However, these Homology-based methods are constrained by the availability of reliable templates. Compared with homology-based methods, machine learning-based, and deep learning-based methods utilize the "ground truth" annotations of catalytic residues from curated public databases to train predictive models, which, in turn, can be applied to infer catalytic residues in most unknown enzymes [17]. For one of the earliest tools, Gutteridge et al. [18] trained a neural network model

to identify catalytic residues using enzyme structure- and sequence-derived features as inputs. Subsequently developed predictors often employed support vector machine (SVM) models or random forest models to a collection of manually curated features [19–23]. Interestingly, Chea et al. [24] did not use a machine learning model but, instead, predicted catalytic residues using statistical scores calculated from a network representation of protein structure and solvent accessibility.

Structural information of the enzymes has proven highly effective in addressing diverse challenges, ranging from predicting enzyme function [25] to guiding enzyme engineering [26]. However, most existing methods for catalytic residue prediction, which rely primarily on enzyme sequences or manually curated structural features, often struggle to capture the spatial arrangements of amino acid residues, as catalytic residues tend to form spatial clusters. Moreover, the intricate chemical reactions catalyzed by enzymes are typically driven by a small subset of catalytic residues. Despite their critical role in enzyme function, current methods often fail to incorporate comprehensive enzyme functional data, thereby limiting the ability to fully explore the connections between catalytic residues, enzyme structure, and function.

We anticipate that recognizing patterns in the spatial arrangement of residues within enzyme structures can substantially enhance the performance of catalytic residue prediction tools and deepen our understanding of enzyme function through the use of modern deep neural networks. To this end, we propose here a deep learning-based solution, called SCREEN, for the accurate prediction of catalytic residues in enzymes. SCREEN employs a graph neural network that models the spatial arrangement of active sites in enzyme structures and combines data derived from enzyme structure, sequence embedding, and evolutionary information obtained by using two complementary methods – BLAST (Basic Local Alignment Search Tool) [27] and HMMER (sequence analysis tool using profile hidden Markov models) [28]. Moreover, we apply the contrastive learning framework to further enhance the predictive performance of SCREEN by incorporating enzyme functional information.

Method

Training and test datasets

We curated a dataset comprising 1055 enzymes with annotated catalytic residues, which we used to train and optimize our predictive model (**Figure 1**A). First, we collected data from M-CSA

database, which contains catalytic residue annotations and details about enzymic reaction mechanisms [29]. We combined the EF family dataset that includes catalytic residue annotations for enzyems from different SCOP families [22]. We filtered the combined dataset by clustering the proteins with the CD-HIT software at 40% sequence identity [30], then randomly selected one protein from each cluster. This prevented overfitting the model into larger clusters of similar enzymes. Next, we collected enzyme three-dimensional structures from the Protein Data Bank (PDB) [9]. We obtained the Enzyme Commission (EC) numbers [31] using the SIFTS database [32]. We used the first-level EC numbers due to the relatively sparse/incomplete nature of these data at the lower levels [33]. We shuffled the dataset randomly and then divided it into training (90%) and validation (10%) subsets (Figure S1).

We acquired five widely-used test datasets to conduct a comparative evaluation of our model against existing tools. These test sets included the EF-fold dataset, EF-superfamily dataset, and HA superfamily dataset, representing enzymes from every SCOP fold and superfamily respectively. We also collected the PC test dataset [20], which was originally obtained from Catalytic Residue Dataset (CATRES) and represents proteins of the PIRSF protein groups [34]. Finally, we obtained the NN test dataset [18], which comprises enzymes of six main classes. Importantly, we excluded enzymes from these five test datasets from our training/validation dataset. Table S1 summarizes the training and the five test datasets.

Overview of the SCREEN model

SCREEN is a supervised deep learner that integrates information derived from atomic structures, sequences, and evolutionary profiles. Specifically, SCREEN presents the input (enzyme structures) as graphs at the residue level, leveraging evolutionary information, sequence embeddings – generated by a modern language model – and relevant structural characteristics, such as B-factors and solvent accessibility (Figure 1B) [35]. We correspondingly employed a graph convolutional neural network to generate propensities for catalytic residues from these inputs. The training process employed enzyme function information (Enzyme Commission numbers) [36] via a contrastive learning framework, utilizing the Triplet Margin Loss function to enable clustering enzymes of the same classes and separating enzymes from different classes in a latent feature space (Figure 1C). This allowed our model to develop class-specific latent feature spaces, leading to improvements in predictive performance/capacity. Moreover, we

employed a dynamic training strategy, in which we initially trained the network using the contrastive learning and then applied our model to accurately identify catalytic residues (Figures S2 and S3).

Graph-based representation of protein structure

We represented the input enzyme structure composed of n residues $\boldsymbol{Enz}$= $\{\mathbf{r_1}, \mathbf{r_1}, \ldots, \mathbf{r_n}\}$ as an attributed graph encoded with evolutionary context, sequence embeddings and relevant structural characteristics, such as solvent accessibility and B-factors. Specifically, the graph representation $\boldsymbol{G} = (\boldsymbol{V}, \boldsymbol{A}, \boldsymbol{X})$ consists of three-dimensional enzyme structure by taking the residue set $\boldsymbol{V} \subseteq \boldsymbol{Enz}$ as graph nodes, the adjacency matrix $\boldsymbol{A}$ with $\boldsymbol{n} \times \boldsymbol{n}$ size that quantifies connectivity of nodes/residues, and the feature matrix $\boldsymbol{X} \in \boldsymbol{\mathcal{R}^{n \times \theta}}$.

The feature matrix $\boldsymbol{X}$= $(\boldsymbol{X_L}, \boldsymbol{X_G}, \boldsymbol{X_A})$ covers the evolutionary, sequence and structural features. The $\boldsymbol{X_L}$ descriptor quantifies evolutionary conservation by utilizing two complementary tools: PSI-BLAST, which is a heuristic algorithm that relies on the dynamic programming [27], and HMMER that is based on the hidden Markov model (HMM) [28]. We run PSI-BLAST on the NCBI's non-redundant (nr) database, with three iterations and the E-value threshold of $< \mathbf{10^{-3}}$. We normalize the output position-specific scoring matrix (PSSM) of size $\boldsymbol{n} \times \mathbf{20}$ with the sigmoid function: $\overline{\boldsymbol{x}} = \frac{\mathbf{1}}{\mathbf{1}+\boldsymbol{e}^{-\boldsymbol{x}}}$. We use HMMER with the uniclust30 database [37] and default parameters to generate the $\boldsymbol{n} \times \mathbf{30}$ HMM matrix that we normalize to the [0, 1] range [28]. The $\boldsymbol{X_G}$ descriptor captures sequence information computed by ProtT5 model, a deep learning language model that was pre-trained on 390 billion amino acids [38]. The enzyme sequence is encoded into residue-level feature embeddings denoted as $\boldsymbol{X_G} \in \boldsymbol{\mathcal{R}^{n \times h_1}}$, where $\boldsymbol{h_1}$ defaults to 1024. These vectors encapsulate information about individual residues that are adjacent in the sequence, and broader protein-level information. Lastly, the $\boldsymbol{X_A}$ descriptor encompasses several key properties that are derived from the atomic-level data: atom types and atomic mass when excluding hydrogen atoms, $\boldsymbol{B}$-factor, residue side-chain presence, the count of bonded hydrogen atoms, ring membership, van der Waals radius, and solvent accessibility. Given that residues might have different numbers of atoms, we compute the average values across all atoms, resulting in atomic descriptor $\boldsymbol{X_A} \in \boldsymbol{\mathcal{R}^{n \times h_2}}$, with $\boldsymbol{h_2}$ = 14.

Predictive model

We designed the graph convolutional neural network (GCN) with three convolutional layers to facilitate the propagation of feature embeddings for residues that share spatial proximity.

For a given graph defined by the adjacency matrix $\boldsymbol{A} \in \{\mathbf{0}, \mathbf{1}\}^{\boldsymbol{n}\times\mathbf{n}}$ and the feature matrix $\boldsymbol{X}$= ($\boldsymbol{X_L}, \boldsymbol{X_G}, \boldsymbol{X_A}$), our model produces residue-level representations $\boldsymbol{H}^{(\mathbf{i})} \in \boldsymbol{\mathcal{R}}^{\boldsymbol{n}\times\boldsymbol{d}_\mathbf{i}}$ where $\boldsymbol{d}_\mathbf{i}$ represents the embedding dimension for the $\boldsymbol{i}$th convolutional layer.

$$H^{(\mathrm{i})} = \mathrm{GCN}(A, [X_L, X_A]) \tag{1}$$

We refine residue representations through the process of neighbor aggregations as follows:

$$H^{(\mathrm{i})} = ReLU\left(\widetilde{D}^{-\frac{1}{2}}(A + I_n)\widetilde{D}^{-\frac{1}{2}}H^{(\mathrm{i}-1)}W^{(\mathrm{i})}\right) \tag{2}$$

$$H^{(0)} = [X_L, X_A] \tag{3}$$

where $\boldsymbol{I_n} \in \boldsymbol{\mathcal{R}}^{\boldsymbol{n}\times\mathbf{n}}$ is the identity matrix, $\widetilde{\boldsymbol{D}} \in \boldsymbol{\mathcal{R}}^{\boldsymbol{n}\times\mathbf{n}}$ is the diagonal degree matrix with entries $\boldsymbol{D_{ii}}$=$\sum_{\boldsymbol{j}}(\boldsymbol{A} + \boldsymbol{I_n})_{\boldsymbol{ij}}$, $\boldsymbol{W}^{(\mathbf{1})} \in \boldsymbol{\mathcal{R}}^{\boldsymbol{\theta}\times\boldsymbol{d}_\mathbf{i}}$ is the trainable weight matrix for the $\boldsymbol{i}$th convolutional layer, ReLU denotes the Rectified Linear Unit activation function, and [] denotes the concatenation operation. The above architecture generates graph representation $\boldsymbol{X_E} \in \boldsymbol{\mathcal{R}}^{\boldsymbol{n}\times\mathbf{d}}$, where d =512 (Figure S4), which we combine using multilayer perception (MLP) network as follows:

$$X_E = ReLU\left(MLP\left(\left[ReLU\left(MLP([H^1, H^2, H^3])\right), ReLU\left(MLP(X_G)\right)\right]\right)\right) \tag{4}$$

We employ three fully connected layers in the MLP network to reduce the feature space to the final output vector Y $\in$ $\mathrm{R}^{n\times 2}$ that gives numeric propensities for putative catalytic residues.

Contrastive learning

We used contrastive learning with Triplet Margin Loss to craft enzyme representations that improve the catalytic residue predictions. Using the graph representation $\boldsymbol{X_E} \in \boldsymbol{\mathcal{R}}^{\boldsymbol{n}\times\mathbf{d}}$, we employed average aggregation across residues, generating a fixed-sized sequence representation vector Z $\in$ $\mathbf{R}^{\mathbf{1}\times\boldsymbol{h_3}}$, with $\boldsymbol{h_3}$ set to 1024. During training in each epoch, we iteratively refined every sequence representation vector and computed enzyme class cluster centres. When training with a query enzyme $\boldsymbol{z_a}$, we selected the enzyme cluster centre embedding from the same

enzyme class as the positive sample $\boldsymbol{z_p}$, and randomly sampled another cluster centre from a different enzyme class as the negative sample $\boldsymbol{z_n}$, which resulted in the following Triplet Margin Loss function:

$$\mathcal{L}^{TM} = \left\| z_a - z_p \right\|_2 - \| z_a - z_n \|_2 + \alpha \quad (5)$$

where we set the margin α to the default value of 1. This loss function minimizes the Euclidean distance between enzyme representations belonging to the same main enzyme class while maximizing the distance between those form different main enzyme classes. We implemented a dynamic training strategy, where we performed contrastive learning for enzyme classification during early training epochs, and gradually shifted towards the default training that converges to produce accurate propensities for catalytic residues.

The multiplexed assays of variant effects data analyses

We gathered the multiplexed assays of variant effects (MAVE) measurements for four enzymes, which provided insight into the impact of a broad collection of substitutions on both enzyme function [6,39]. We categorized the missense variants of PTEN into two main groups: functional or inactive, regardless of their effect on abundance. The classification thresholds for the scores generated by each MAVE were guided by an established methodology [40]. Specifically, we used a minimal number of Gaussians (three) to ensure a reliable fit to the variant score distributions, and the intersection point between the first and last Gaussian served as the classification cut-off. Adopting this binary classification approach allowed us to categorize variants into two classes: (1) Wild Type-Like (WTL), variants characterized by high activity; (2) Functional Loss (FL), variants assigned that exhibit low activity.

Results

SCREEN accurately predicts catalytic residues

We collected five commonly-used test datasets to comparatively assess SCREEN against eight current solutions. These test datasets included the EF superfamily and EF fold datasets [22], the HA superfamily dataset [24], the NN dataset [18], and the PC dataset [20]. We compared the results from SCREEN with those of a conventional sequence-based method, CRpred [19], and six tools that employed different predictive models based on enzyme structures. These tools

included a neural network-based approach [18], three SVM-based methods [20], a random forests-based PREvaIL [23], and a statistical approach [24]. We compared the precision, recall, and F1 score to evaluate the predictions of catalytic residues, referencing the reported performance from the original paper (**Table 1**, Figure S5). We also evaluated the recently proposed graph-based method, AEGAN [41], by retraining and testing it on the same training and test sets, with the hyperparameter for negative sample size set to 20. We showed that SCREEN consistently outperformed the eight tools for the five test datasets. Compared to these methods, SCREEN achieved a higher F1 score across all five test datasets. The high F1 scores achieved by SCREEN were coupled with balanced and high values of precision and recall that ranged between 61.0 and 69.3 and between 61.2 and 82.0, respectively. We also quantified and compared two other popular metrics, the area under the receiver operating curve (AUC) and the area under the precision-recall curve (AUPR). **Figure 2**A reveals that SCREEN achieved substantially higher AUC and AUPR scores when compared with the latest structure-based (PREvaIL) and the sequence-based (CRpred) tools, except for the EF superfamily and EF fold dataset, where AUC and AUPRC values were comparable.

Using SCREEN, we also measured Best-F1 score and AUPR values for specific enzyme types (Figure S6). Particularly, for hydrolases, which represent the largest portion of the training dataset (313 of 1055; 29.7%), SCREEN obtained notable consistency across the five test datasets, with Best-F1 scores ranging from 0.63 to 0.78, AUPR values ranging from 0.56 to 0.68. For the isomerase data, the predictive performance was particularly high, with the Best-F1 score exceeding 0.74 and AUPR surpassing 0.80, even though these enzymes represent only a small portion of the training set (87 of 1055; 8.2%).

The use of enzyme structure information in SCREEN markedly improves the catalytic residue prediction

The catalytic residues typically tend to form cohesive clusters within the three-dimensional enzyme structures. Thus, we systematically investigated the spatial distribution of residues in enzyme structures by measuring the Euclidian distances to the nearest catalytic residue for both catalytic and non-catalytic residues in individual protein sequences. Figure 2B shows there was a clear difference in the distribution of Euclidian distances, with the catalytic residues peaking at ~ 6Å, and most non-catalytic residues exceeding 15Å, supporting that catalytic residues form

cohesive active enzymatic sites. We investigated whether SCREEN could reconstruct the same spatial distributions for enzyme catalytic residues. A performance assessment of SCREEN using five test datasets (Figure 2C, Figure S7) showed that the majority of catalytic residues predicted grouped together in the structure and that the distance values were consistent among the datasets, with median values being ~ 6Å. These results agree with findings presented in Figure 2B, implying SCREEN accurately captures the spatial distribution of predicted catalytic residues.

These findings suggest the use of structure information in SCREEN likely results in predictive performance improvements. We further investigated whether SCREEN could improve results compared with a "baseline model" that excludes structure-based inputs and replaces the graph convolutional neural network (GCN) with a sequence-based convolutional neural network (CNN). The comparison of SCREEN with the baseline using five test datasets employing the Best-F1 score, AUPR, and AUC metrics (Figure 2D) showed that the use of graph network led to a marked improvement in predictive performance.

To assess predictions, we selected enzymes from the EF superfamily dataset ranked around the 10%, 50%, and 90% percentiles based on the Best-F1 score. We plotted catalytic residues predicted by the structure-based SCREEN and the sequence-based baseline model against ground-truth catalytic residues within enzyme structures. In the top 10%, for carboxylic ester hydrolase (PDB ID:1LE6), SCREEN accurately predicted all catalytic residues, whereas the sequence-based baseline introduced three false positives (Figure 2E). In the median ranking, for casein kinase-1 (PDB ID:1CSN), SCREEN successfully identified Asp-131, Lys-133, and Thr-181 as key residues, but the baseline model failed to identify Thr-181 (Figure 2F). In the 90% ranking, for dihydropteroate synthase (PDB ID:1AD1), SCREEN identified Arg-239 and gave one false positive, whereas the baseline misidentified two non-essential residues (Figure 2G). These findings indicate that SCREEN has a superior performance compared with the sequence-based baseline model. This supports our design and, in particular, the use of the graph network and enzyme structure as a key input. We also investigated various graph convolution types, including the extensively employed Graph Convolutional Layer (GCN), Graph Attention (GAT), and Graph Isomorphism Network (GIN), but none of the models outperformed another using the same test sets and metrics (Figure S8).

SCREEN accurately predicts catalytic residues using structure models

Although we showed SCREEN's predictive performance benefits from the use of enzyme structure data, this information is often missing for many proteins/enzymes. Recent advances in protein structure prediction, like the AlphaFold algorithm [42-44], make it possible to accurately predict protein structure from sequences and to use such structure models as the input to SCREEN. We evaluated whether the use of predicted structures rather than experimentally determined structures would alter SCREEN's predictive performance (Table S2). We generated C_α-C_α contact maps for enzymes based on experimental protein structures sourced from PDB, as well as putative structures from AlphaFold. We tested SCREEN's performance by using each of these two sets of contact maps, comparing it against a sequence-based baseline model devoid of structural information. **Figure 3**A showed that SCREEN consistently benefits from utilizing putative enzyme structures (with Best-F1 = 0.702, AUPR = 0.644, and AUC = 0.985) compared with sequence data alone (with Best-F1 = 0.691, AUPR = 0.624, and AUC = 0.973).

To further assess SCREEN's denoising power on predicted structure error, we evaluated the model performance employing AlphaFold-predicted structures with varying quality. Specifically, we quantified the quality of predicted structures employing root-mean-square deviation (RMSD) metric as compared with experimental solved structures (RMSD = 0). Figure 3B revealed that SCREEN's performance was better employing AlphaFold-derived structures than using sequence data alone. SCREEN also out-performed currently-employed tools CRHunter and PREVAIL across the entire RMSD range, achieving the Best-F1 score of > 0.6 using AlphaFold-predicted structures, contrasting average Best-F1 scores of 0.45 and 0.26 for CRHunter and PREvaIL, respectively.

SCREEN performed relatively well using predicted structures (Figure 3A and B), irrespective of the quality of predictions via AlphaFold. Here, we used two examples to illustrate SCREEN's ability to accurately identify catalytic residues even in relatively low-quality predicted structures (Figure 3C). For the human calcineurin heterodimer (PDB ID: 1AUI, Chain A) [45], where the AlphaFold-predicted structure had an RMSD score of up to 6.76 Å, SCREEN successfully identified all 10 catalytic residues. Similarly, SCREEN accurately identified catalytic residues (Ser-53, Pro-54, and Asp-96) in the PVUII DNA methyltransferase (PDB ID: 1BOO, Chain A) [46], for which the structure predicted had an RMSD score of 4.75 Å. Taken together, these results suggested that SCREEN can accurately predict catalytic residues from AlphaFold-predicted enzyme structures, which might be attributed to the robustness of the input features

and how they are represented in the graph network model.

SCREEN predicts catalytic residues in "previously-unseen" enzymes

We investigated SCREEN's ability to generate accurate predictions for "previously unseen" enzymes. To this end, we categorized enzymes in the five test sets into three distinct groups based on their sequence identities, namely ≤ 30% (low), 30 to 70% (moderate) and > 70% (high), to enzymes in the training dataset. Figure 3D showed SCREEN's Best-F1 scores across three sequence identity ranges, *i.e.*, ≤ 30% (low), 30 to 70% (moderate), and > 70% (high), for each of the five test datasets. We showed that SCREEN consistently outperformed the sequence-based baseline model for each of the five datasets and the three identity ranges. Importantly, SCREEN's predictions were accurate also for enzymes with limited sequence identities to those in the training datasets, achieving the Best-F1 scores of 0.725, 0.738, 0.652, 0.718, and 0.794. These predictions were significantly better than those achieved using the existing tools, such as CRHunter (Best-F1 values of 0.328, 0.315, 0.450, 0.365, and 0.140, respectively) and PREVAIL (Best-F1 of 0.264, 0.261, 0.263, 0.264, and 0.263, respectively). We also used CATH [47] assignments to evaluate the model's robustness on enzymes sharing no homologous superfamilies with those in the training dataset, as shown in Table S3 and Figure S9.

Model training for enzyme classes improves the prediction of catalytic residues

We investigated whether the training of the graph network model with distinct enzyme function implications would improve predictive performance, considering that the small subset of catalytic residues contributes to the intricate functions of enzymes. We used enzyme class information (first-level EC numbers) to refine enzyme representation through a contrastive learning framework during the training process. This led to a separation of latent feature spaces in SCREEN's deep network model for different types of enzymes. We displayed these latent feature spaces among different enzyme classes employing t-distributed stochastic neighbor embedding (t-SNE) [48]. We showed that predictive performance (metrics: F1, AUPR, and AUC) of SCREEN was enhanced compared to when enzyme function was not incorporated for each of the data sets (EF fold, HA superfamily, EF superfamily, NN, and PC) (**Figure 4**A, Figure S10) and that SCREEN was able to group enzymes with similar functions together and separating enzymes with distinct functions (Figure 4C). Taken together, these findings indicate that the

enzyme function incorporation through contrastive learning during the training process improves predictive performance as SCREEN can differentiate catalytic from non-catalytic residues for different distinct types of enzymes (Figures S11 and S12).

SCREEN can capture selected features of catalytic residues

We analyzed catalytic residues predicted by SCREEN, in order to investigate whether they possess structural and biophysical characteristics expected for enzymes. To better understand the relevance of the features learned by SCREEN, we initially displayed the general chemical properties (including hydrophobicity, charge, and hydrogen bonds), along with low-dimensional projections of residue-level representations (Figure 4B). We observed that charged amino acids dominated in the catalytic residues predicted (Figure 4D), consistent with previous findings showing that electrostatic filtering has a marked effect on enzyme substrate selection [49]. Moreover, catalytic residues were inferred to be more rigid (structurally) than non-catalytic residues (based on low *vs.* high B-factor values; see Figure 4E), which accords with a previous study of native catalytic residues [50]. Fewer hydrophobic residues were associated with catalytic residues (Figure 4D), which is consistent with limited solvent accessibility (Figure 4F) and suggests substrate avoidance in substrate–enzyme interactions [51]. Collectively, these results show that SCREEN can capture key features that typify native catalytic residues in distinct classes of enzymes.

Linking catalytic residues to enzyme function and structure

Based on these catalytic residues, we further analyzed the sequence-structure-function relationship of enzymes to gain deeper insights into enzymes' catalytic mechanisms. We categorized enzymes according to their catalytic functions defined by third-level EC numbers and then by (complete) fourth-level EC numbers (which link to substrates). For enzyme clusters sharing the same catalytic function/mechanism, we assessed structural similarity by their TM-scores among cluster members [52], selected enzymes from individual clusters and mapped the catalytic residue predictions to respective three-dimensional structures (**Figure 5**).

The results indicated that same catalytic motif may represent distinct enzymes that serve different functions linked to diverse ligands or substrates. Figure 5A shows both 4-hydroxyproline betaine 2-epimerase (PDB ID: 4h2h, Chain A) and L-Ala-D/L-Glu epimerase

(PDB ID: 1TKK, Chain A) belong to the same superfamily and share a common catalytic motif (KDEDK) [53], but they differ in their substrate specificity: 4-hydroxyproline betaine 2-epimerase facilitates the 2-epimerization of trans-4-hydroxy-L-proline betaine (tHyp-B) to cis-4-hydroxy-D-proline betaine (cHyp-B), whereas L-Ala-D/L-Glu epimerase catalyzes the reversible epimerization of L-Ala-D-Glu to L-Ala-L-Glu.

Enzymes in a particular family can have similar structures and functions, despite undergoing sequence divergence through evolution. A compelling illustration emerges when we studied two related enzymes, the glutamate racemase (PDB ID: 1b73, Chain A) [54] and aspartate racemase (PDB ID: 1jfl, Chain A) [55] (Figure 5A). Despite enabling similar reactions via the same mechanism, their catalytic residues are significantly different but have analogous tertiary structures. Another example relates to protein-tyrosine-phosphatase non-receptor class (PDB ID: 1ytw, Chain A) [56] and protein-tyrosine-phosphatase non-receptor type 1 (PDB ID: 1bzc, Chain A) [57] (Figure 5B). Here, although both enzymes are tyrosine phosphatases and catalyze the same reaction to remove phosphoryl groups from tyrosine residues in proteins, their respective catalytic residues are distinctly different.

The shared structural arrangements of catalytic residues can be associated with functional similarity. Figure 5B shows phosphatase 5 (PDB ID: 1S95, Chain A) and phosphatase 2B (PDB ID: 1aui, Chain A) exhibite significant structural similarity and share catalytic residues (motif: DHDDRNHHRH) pertaining to serine/threonine phosphatase function(s), characterized by executing a “nucleophilic assault” on the phosphorus atom within a phosphorylated serine or threonine residue [45].

Although enzymes catalyzing the same reactions often exhibit marked sequence and/or structural similarity, exceptions exist where structurally dissimilar enzymes enable similar reactions via the same mechanism. This is expected since enzymes facilitate numerous reactions using a finite set of building blocks in their residues, resulting in multiple enzymes inevitably sharing components of their catalytic mechanisms. Here, we showed that non-homologous proteins, protein phosphatase 5 (PDB ID: 1S95, Chain A) and dual-specificity phosphatase (PDB ID: 1d5r, Chain A) employ distinct structural motifs to execute the same reaction that dephosphorylates a phosphoprotein substrate (Figure 5B, Figure S13).

Associating mutations with the SCREEN-predicted catalytic pockets

Here, we studied the patterns of mutations in the context of their proximity to the catalytic pockets predicted by SCREEN. We collected the multiplexed assays of variant effects data, probing mutation effect(s) on the functions of four different enzymes, namely PTEN tumor suppressor (PDB ID: 1d5r, Chain A) [58], human cytochrome P450 CYP2C9 (PDB ID: 1og5, Chain A) [59], NUDT15 (PDB ID: 5lpg, Chain A) [60], and *Escherichia coli* TEM1 beta-lactamase (PDB ID: 1btl, Chain A) [61], encompassing a total of 15,665 variants across 1,343 residues.

To function effectively, enzymes must be present at sufficiently high levels and have suitable catalytic residues in the active sites [62]; however, mutations can affect both of these aspects, potentially resulting in impaired enzymatic function. We used the MAVE data [63] for two residue classes: (1) wild type-like (WTL) residues that exhibit high functional tolerance to mutations, whose most missense mutations do not adversely impact enzyme function; and (2) functional loss (FL) residues that are prone to mutations that either decrease abundance (*e.g.*, unstable structures) and/or impair function, leading to diminished enzyme activity (Figure S14).

We systematically analyzed key characteristics of mutations in the context of the predicted catalytic residues. Specifically, we applied an additional tree-structured model to SCREEN (**Figure 6**A). The Euclidean distance values to the closest catalytic residue predicted by SCREEN combined with solvent accessibility, which, as expected, were inferred to vary according to residue type (WTL or FL), allowing to differentiate among different mutation groups. Figure 6B illustrates the catalytic residues predicted by SCREEN along with residue mutation type predictions. We performed five-fold cross-validation on all MAVE data; our results revealed an average accuracy of 0.70 and 0.84, and precision of 0.58 and 0.88 on the validation data and the entire dataset, respectively (Figure 6C). We found distinct spatial distribution patterns for the WTL and FL residues based on their Euclidean distances from the putative catalytic residues (Figure 6D). Our result aligns well with experimental data showing that FL residues are relatively close to the catalytic site, while WTL residues are distributed throughout the structure (Figure S15). This result indicates that SCREEN can be useful to establish the impact of mutations on enzyme structure and function and provides a tool to guide the identification of disease-associated mutations in enzymes.

Discussion

Enzymes can catalyze a broad set of chemical reactions using a limited set of catalytic residues [64]. Identifying these residues allows us to understand how existing enzymes function at the molecular level and to design new ones. In this work, we hypothesize that the structural organization of catalytic residues in spatial space, along with their generally high evolutionary conservation, collectively contributes to catalytic residue identification. To this end, we conceptualized, designed, and assessed SCREEN, a structure-based graph network that uses functional priors through contrastive learning and combines structure-, sequence-, and evolutionary profile-based representations to accurately predict catalytic residues in enzymes.

Comparative empirical assessments using five commonly-utilized test datasets and seven currently-available (published) predictors revealed that SCREEN (1) accurately predicts catalytic residues in known and computationally-modeled enzymes; (2) outperforms current tools; and (3) generalizes well to enzymes that have limited similarity to enzymes used to train the model, suggesting that SCREEN is applicable to currently-unknown enzymes. Incorporating enzyme function as a prior could improve the prediction of catalytic residues by enhancing the consistency of enzyme's latent representations based on their functions. However, we did not explore enzyme function prediction in depth, *e.g.*, addressing questions such as can we also utilize this model to solve enzyme function classification tasks? A more comprehensive understanding of the spatial distribution across diverse enzyme functions and the sequence-structure-function relationship could be achieved by analyzing a larger sample of enzymes that thoroughly covers the EC space. Additionally, our analysis is limited by treating enzymes as independent units. Enzymatic reactions involve multiple residues, substrates, and cofactors interacting across various chemical steps [13]; as such, an integrated analysis would be necessary for a more comprehensive understanding of the catalytic chemical activity. Further, future research should focus on determining the level of confidence that can be assigned to model predictions of catalytic residues, as well as exploring the techniques that can effectively assess this confidence.

We demonstrate that SCREEN could infer key structural and biophysical features, including amino acid charge, solvent accessibility and structural rigidity, of predicted and known catalytic residues. Further, We undertook sequence-structure-function analyses to link catalytic residues to enzyme structure and function. Our analyses revealed that while enzymes that catalyze identical reactions often display significant sequence and/or structural similarity, exceptions

arise wherein dissimilar sequences and/or structures can catalyze reactions via the same mechanism. In addition, using multiplexed enzyme mutation data, we showed that SCREEN could infer the tolerance of individual catalytic residues to mutations and, thus, predict which mutations in catalytic residues likely lead to the functional loss of an enzyme. Taken together, SCREEN should provide a useful tool for the reliable prediction of catalytic residues to support studies of known and unknown enzyme groups/classes as well as enable *in silico* investigations of diseases linked to mutations.

Conclusion

SCREEN is an efficient and robust method for high-throughput prediction of catalytic residues by integrating enzyme functional and structural information. We demonstrate SCREEN's effectiveness and robustness across various widely used datasets, illustrating that the predicted putative catalytic residues closely align with the key structural and biophysical characteristics of native catalytic residues. Furthermore, we performed sequence-structure-function analyses to establish connections between catalytic residues and enzyme structure and function. This highlights SCREEN's potential for reliably predicting catalytic residues in both known and unknown enzyme groups/classes, thereby supporting studies of the molecular mechanisms underlying enzyme functions.

Code availability

All source data needed to evaluate the conclusions in the paper can be found at https://huggingface.co/datasets/Biocollab/SCREEN/tree/main. All source code is available at https://github.com/BioColLab/SCREEN and https://ngdc.cncb.ac.cn/biocode/tool/7580.

CRediT author statement

Tong Pan: Conceptualization, Methodology, Investigation, Writing – original draft. **Yue Bi:** Visualization. **Xiaoyu Wang:** Visualization. **Ying Zhang:** Visualization. **Geoffrey I. Webb:** Methodology, Supervision. **Robin B. Gasser:** Investigation, Supervision, Writing – review & editing. **Lukasz Kurgan:** Methodology, Investigation, Supervision, Writing – review & editing. **Jiangning Song:** Conceptualization, Methodology, Investigation, Supervision, Writing – review

& editing. All authors have read and approved the final manuscript.

Competing interests

All authors declare they have no competing interests.

Supplementary material

Supplementary material is available at *Genomics, Proteomics & Bioinformatics* online (https://doi.org/10.1093/gpbjnl/qzaxxxx).

Acknowledgments

Financial support to RBG and JS from the Australian Research Council (Grant No. ARC LP220200614) is gratefully acknowledged.

ORCID

0009-0003-6676-5727 (Tong Pan)

0009-0001-6220-7475 (Yue Bi)

0000-0003-4444-6197 (Xiaoyu Wang)

0000-0003-1792-0121 (Ying Zhang)

0000-0001-9963-5169 (Geoffrey I. Webb)

0000-0002-4423-1690 (Robin B. Gasser)

0000-0002-7749-0314 (Lukasz Kurgan)

0000-0001-8031-9086 (Jiangning Song)

Reference

[1] Benkovic SJ, Hammes-Schiffer S. A perspective on enzyme catalysis. Science 2003;301:1196–202.

[2] O'maille PE, Malone A, Dellas N, Andes Hess Jr B, Smentek L, Sheehan I, et al. Quantitative exploration of the catalytic landscape separating divergent plant sesquiterpene synthases. Nat Chem Biol 2008;4:617–23.

[3] Zhou J, Yan W, Hu G, Shen B. Amino acid network for prediction of catalytic residues in enzymes: a comparison survey. Curr Protein Pept Sci 2016;17:41–51.

[4] Makarova KS, Wolf YI, Iranzo J, Shmakov SA, Alkhnbashi OS, Brouns SJ, et al. Evolutionary classification of CRISPR-Cas systems: a burst of class 2 and derived variants. Nat Rev Microbiol 2020;18:67–83.
[5] Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods 2010;7:248–9.
[6] Mighell TL, Evans-Dutson S, O'Roak BJ. A saturation mutagenesis approach to understanding PTEN lipid phosphatase activity and genotype-phenotype relationships. Am J Hum Genet 2018;102:943–55.
[7] Markin CJ, Mokhtari DA, Sunden F, Appel MJ, Akiva E, Longwell SA, et al. Revealing enzyme functional architecture via high-throughput microfluidic enzyme kinetics. Science 2021;373:eabf8761.
[8] Porter CT, Bartlett GJ, Thornton JM. The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. Nucleic Acids Res 2004;32:D129–D33.
[9] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The protein data bank. Nucleic Acids Res 2000;28:235–42.
[10] Consortium U. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res 2019;47:D506–D15.
[11] Storici F, Lewis LK, Resnick MA. In vivo site-directed mutagenesis using oligonucleotides. Nat Biotechnol 2001;19:773–6.
[12] Süel GM, Lockless SW, Wall MA, Ranganathan R. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. Nat Struct Biol 2003;10:59–69.
[13] Ribeiro AJ, Tyzack JD, Borkakoti N, Holliday GL, Thornton JM. A global analysis of function and conservation of catalytic residues in enzymes. J Biol Chem 2020;295:314–24.
[14] Radivojac P, Clark WT, Oron TR, Schnoes AM, Wittkop T, Sokolov A, et al. A large-scale evaluation of computational protein function prediction. Nat Methods 2013;10:221–7.
[15] Wallace AC, Borkakoti N, Thornton JM. TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. Protein Sci 1997;6:2308–23.
[16] Mistry J, Bateman A, Finn RD. Predicting active site residue annotations in the Pfam database. BMC Bioinformatics 2007;8:1–14.
[17] Zhang T, Zhang H, Chen K, Shen S, Ruan J, Kurgan L. Accurate sequence-based prediction of catalytic residues. Bioinformatics 2008;24:2329–38.
[18] Gutteridge A, Bartlett GJ, Thornton JM. Using a neural network and spatial clustering to predict the location of active sites in enzymes. J Mol Biol 2003;330:719–34.
[19] Zhang T, Zhang H, Chen K, Shen S, Ruan J, Kurgan L. Accurate sequence-based prediction of catalytic residues. Bioinformatics 2008;24:2329–38.
[20] Petrova NV, Wu CH. Prediction of catalytic residues using Support Vector Machine with selected protein sequence and structural properties. BMC Bioinformatics 2006;7:1–12.
[21] Sun J, Wang J, Xiong D, Hu J, Liu R. CRHunter: integrating multifaceted information to predict catalytic residues in enzymes. Sci Rep 2016;6:34044.
[22] Youn E, Peters B, Radivojac P, Mooney SD. Evaluation of features for catalytic residue prediction in novel folds. Protein Sci 2007;16:216–26.
[23] Song J, Li F, Takemoto K, Haffari G, Akutsu T, Chou K-C, et al. PREvaIL, an integrative approach for inferring catalytic residues using sequence, structural, and network features in a machine-learning framework. J Theor Biol 2018;443:125–37.

[24] Chea E, Livesay DR. How accurate and statistically robust are catalytic site predictions based on closeness centrality? BMC Bioinformatics 2007;8:1–14.
[25] Pan T, Li C, Bi Y, Wang Z, Gasser RB, Purcell AW, et al. PFresGO: an attention mechanism-based deep-learning approach for protein annotation by integrating gene ontology inter-relationships. Bioinformatics 2023;39:btad094.
[26] Chen K, Arnold FH. Engineering new catalytic activities in enzymes. Nat Catal 2020;3:203–13.
[27] Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–402.
[28] Potter SC, Luciani A, Eddy SR, Park Y, Lopez R, Finn RD. HMMER web server: 2018 update. Nucleic Acids Res 2018;46:W200–W4.
[29] Ribeiro AJM, Holliday GL, Furnham N, Tyzack JD, Ferris K, Thornton JM. Mechanism and Catalytic Site Atlas (M-CSA): a database of enzyme reaction mechanisms and active sites. Nucleic Acids Res 2018;46:D618–D23.
[30] Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 2006;22:1658–9.
[31] Bairoch A. The ENZYME database in 2000. Nucleic Acids Res 2000;28:304–5.
[32] Velankar S, Dana JM, Jacobsen J, Van Ginkel G, Gane PJ, Luo J, et al. SIFTS: structure integration with function, taxonomy and sequences resource. Nucleic Acids Res 2012;41:D483–D9.
[33] Schomburg I, Chang A, Ebeling C, Gremse M, Heldt C, Huhn G, et al. BRENDA, the enzyme database: updates and major new developments. Nucleic Acids Res 2004;32:D431–D3.
[34] Bartlett GJ, Porter CT, Borkakoti N, Thornton JM. Analysis of catalytic residues in enzyme active sites. J Mol Biol 2002;324:105–21.
[35] Yip KM, Fischer N, Paknia E, Chari A, Stark H. Atomic-resolution protein structure determination by cryo-EM. Nature 2020;587:157–61.
[36] Yu T, Cui H, Li JC, Luo Y, Jiang G, Zhao H. Enzyme function prediction using contrastive learning. Science 2023;379:1358–63.
[37] Mirdita M, Von Den Driesch L, Galiez C, Martin MJ, Söding J, Steinegger M. Uniclust databases of clustered and deeply annotated protein sequences and alignments. Nucleic Acids Res 2017;45:D170–D6.
[38] Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L, et al. Prottrans: Toward understanding the language of life through self-supervised learning. IEEE Trans Pattern Anal Mach Intell 2021;44:7112–27.
[39] Matreyek KA, Starita LM, Stephany JJ, Martin B, Chiasson MA, Gray VE, et al. Multiplex assessment of protein variant abundance by massively parallel sequencing. Nat Genet 2018;50:874–82.
[40] Cagiada M, Johansson KE, Valanciute A, Nielsen SV, Hartmann-Petersen R, Yang JJ, et al. Understanding the origins of loss of protein function by analyzing the effects of thousands of variants on activity and abundance. Mol Biol Evol 2021;38:3235–46.
[41] Shen X, Zhang S, Long J, Chen C, Wang M, Cui Z, et al. A Highly Sensitive Model Based on Graph Neural Networks for Enzyme Key Catalytic Residue Prediction. J Chem Inf Model 2023;63:4277–90.
[42] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Applying and improving AlphaFold at CASP14. Proteins 2021;89:1711–21.

[43] Elofsson A. Progress at protein structure prediction, as seen in CASP15. Curr Opin Struct Biol 2023;80:102594.
[44] Simpkin AJ, Mesdaghi S, Sanchez Rodriguez F, Elliott L, Murphy DL, Kryshtafovych A, et al. Tertiary structure assessment at CASP15. Proteins 2023;91:1616–35.
[45] Kissinger CR, Parge HE, Knighton DR, Lewis CT, Pelletier LA, Tempczyk A, et al. Crystal structures of human calcineurin and the human FKBP12-FK506-calcineurin complex. Nature 1995;378:641–4.
[46] Gong W, O'Gara M, Blumenthal RM, Cheng X. Structure of pvu II DNA-(cytosine N4) methyltransferase, an example of domain permutation and protein fold assignment. Nucleic Acids Res 1997;25:2702–15.
[47] Pearl FM, Bennett C, Bray JE, Harrison AP, Martin N, Shepherd A, et al. The CATH database: an extended protein family resource for structural and functional genomics. Nucleic Acids Res 2003;31:452–5.
[48] Van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res 2008;9:2579–2605.
[49] Shah NH, Wang Q, Yan Q, Karandur D, Kadlecek TA, Fallahee IR, et al. An electrostatic selection mechanism controls sequential kinase signaling downstream of the T cell receptor. Elife 2016;5:e20105.
[50] Bartlett GJ, Porter CT, Borkakoti N, Thornton JM. Analysis of catalytic residues in enzyme active sites. J Mol Biol 2002;324:105–21.
[51] Johnson JL, Yaron TM, Huntsman EM, Kerelsky A, Song J, Regev A, et al. An atlas of substrate specificities for the human serine/threonine kinome. Nature 2023;613:759–66.
[52] Koehler Leman J, Szczerbiak P, Renfrew PD, Gligorijevic V, Berenberg D, Vatanen T, et al. Sequence-structure-function relationships in the microbial protein universe. Nat Commun 2023;14:2351.
[53] Zhao S, Kumar R, Sakai A, Vetting MW, Wood BM, Brown S, et al. Discovery of new enzymes and metabolic pathways by using structure and genome context. Nature 2013;502:698–702.
[54] Hwang KY, Cho C-S, Kim SS, Sung H-C, Yu YG, Cho Y. Structure and mechanism of glutamate racemase from Aquifex pyrophilus. Nat Struct Biol 1999;6:422–6.
[55] Liu L, Iwata K, Kita A, Kawarabayasi Y, Yohda M, Miki K. Crystal structure of aspartate racemase from Pyrococcus horikoshii OT3 and its implications for molecular mechanism of PLP-independent racemization. J Mol Biol 2002;319:479–89.
[56] Fauman EB, Yuvaniyama C, Schubert HL, Stuckey JA, Saper MA. The X-ray crystal structures of Yersinia tyrosine phosphatase with bound tungstate and nitrate: mechanistic implications. J Biol Chem 1996;271:18780–8.
[57] Groves MR, Yao Z-J, Roller PP, Burke TR, Barford D. Structural basis for inhibition of the protein tyrosine phosphatase 1B by phosphotyrosine peptide mimetics. Biochemistry 1998;37:17773–83.
[58] Furnari FB, Huang HS, Cavenee WK. The phosphoinositol phosphatase activity of PTEN mediates a serum-sensitive G1 growth arrest in glioma cells. Cancer Res 1998;58:5002–8.
[59] Amorosi CJ, Chiasson MA, McDonald MG, Wong LH, Sitko KA, Boyle G, et al. Massively parallel characterization of CYP2C9 variant enzyme activity and abundance. Am J Hum Genet 2021;108:1735–51.

[60] Moriyama T, Nishii R, Perez-Andreu V, Yang W, Klussmann FA, Zhao X, et al. NUDT15 polymorphisms alter thiopurine metabolism and hematopoietic toxicity. Nat Genet 2016;48:367–73.
[61] Sirot D, Recule C, Chaibi E, Bret L, Croize J, Chanal-Claris C, et al. A complex mutant of TEM-1 beta-lactamase with mutations encountered in both IRT-4 and extended-spectrum TEM-15, produced by an Escherichia coli clinical isolate. Antimicrob Agents Chemother 1997;41:1322–5.
[62] Jack BR, Meyer AG, Echave J, Wilke CO. Functional sites induce long-range evolutionary constraints in enzymes. PLoS Biol 2016;14:e1002452.
[63] Cagiada M, Bottaro S, Lindemose S, Schenstrøm SM, Stein A, Hartmann-Petersen R, et al. Discovering functionally important sites in proteins. Nat Commun 2023;14:4175.
[64] Kraut DA, Carroll KS, Herschlag D. Challenges in enzyme mechanism and energetics. Annu Rev Biochem 2003;72:517–71.

Figure legends

Figure 1 SCREEN – a predictor of catalytic residues in enzymes

A. Data collection. We curate catalytic residue annotations from the M-CSA database, extract the corresponding enzyme structures from the RCSB PDB database, and retrieve corresponding enzyme function information using the SIFTS program [32] . **B.** Generation of inputs that include evolutionary profiles based on multiple sequence alignments (MSA), sequence embeddings that leverage a large-scale protein language model, and structural characteristics derived from the atomic structures. **C.** The architecture of SCREEN's predictive model.

Figure 2 Predictive performance of SCREEN

A. Comparison of the sequence-based CRpred and structure-based PREvaIL on the five test datasets. **B.** The distribution of the Euclidean distances to the nearest catalytic residues for the residues in enzymes from the M-CSA database [13]. **C.** The distribution of the Euclidean distances to the nearest catalytic residue for the catalytic residues predicted by SCREEN. For boxplots, the center line represents the median, top and bottom edges are the first and third quartiles, respectively. **D.** Comparison of the structure-based SCREEN model (GCN encoder) with a sequence-based baseline model (CNN encoder) on the five test datasets. **E.–G.** Three examples, ranked approximately at 10% (E), 50% (F), and 90% (G) based on the Best-F1 scores for catalytic residues predicted by SCREEN. We compare SCREEN with a sequence-based baseline model (CNN encoder). The native catalytic residues are in green in a zoomed figure.

The correctly identified catalytic residues by SCREEN and baseline model (CNN encoder) are marked in red, while misidentified residues are in gray.

Figure 3 Analysis of predictive performance when using putative enzyme structure and low similarity test proteins

A. Comparison of results produced by the SCREEN using the native structure, structure predicted from sequence with AlphaFold, and sequence-based baseline predictor (CNN encoder); the error bars represent standard deviation of the mean based on 10 independent runs. **B.** Comparison of results produced by the SCREEN using the native structure, AlphaFold-predicted structure, and sequence-based baseline predictor in the context of the quality of the AlphaFold-predicted structure. The dashed horizontal lines represent the Best-F1 scores generated by CRHunter (blue line) and PREvaIL (yellow line). **C.** Examples of contact maps by ground-truth of native enzyme structures (PDB) and AlphaFold-predicted enzyme structures, and corresponding catalytic residues identified by SCREEN for enzymes 1aui-A and 1boo-A. **D.** Evaluation for enzymes that share varying levels of sequence identify with the training proteins on the five test datasets.

Figure 4 Analysis of putative catalytic residues generated by SCREEN

A. Comparison of the SCREEN models with its variant SCREEN_NoEC that does not incorporate enzyme function through contrastive learning on the five test datasets. This figure also shows results from CRpred and PREvaIL. **B.** The t-SNE based visualization of latent feature spaces in the SCREEN model for residue characteristics, such as hydrophobicity (left), hydrogen bond types (medium) and charges (right). The "H-bond acceptor" denotes residues exclusively serving as hydrogen bond acceptors without containing H-bond donor atoms. **C.** The t-SNE based visualization of latent feature spaces in the SCREEN model for different color-coded enzyme classes. **D.–F.** Analysis of structural and biophysical characteristics, which include biophysical properties of amino acids (D), B-factor (E) and solvent accessibility (F) for the putative catalytic residues generated by SCREEN.

Figure 5 Diversity of enzyme structure and catalytic residues with the same catalytic mechanism

We examine enzymes that have the same catalytic mechanism: Isomerases acting on amino acids and derivatives with EC number 5.1.1 (**A**) and phosphoric monoester hydrolases with EC number 3.1.3 (**B**). We plot the TM-score as a measure of structural similarity as a heatmap, with larger numbers (more yellow) representing more similar structures. We also map the catalytic residues prediction by SCREEN onto the structures as well as enzyme reactions on the right.

Figure 6 Assessing the tolerance of catalytic residues to mutations

A. Architecture of the tree-structured model for characterizing mutated residues based on catalytic residue predictions by SCREEN. **B.** SCREEN identified catalytic residues of four different enzymes: PTEN tumor suppressor (PDB ID: 1d5r, Chain A), human cytochrome P450 CYP2C9 (PDB ID: 1og5, Chain A), NUDT15 (PDB ID: 5lpg, Chain A), and *Escherichia coli* TEM1 beta-lactamase (PDB ID: 1btl, Chain A) (top). Residues within enzyme structures are colored according to their predicted mutant class: blue corresponds to the Wild Type-like (WTL) residues, while gray to the Functional Loss (FL) residues (bottom). **C**. Quality of distance-based classification of residues with different mutation classes, measured by accuracy and precision on both the validation and entire datasets. **D.** Distribution of the Euclidean distances for residues of different mutation classes.

Table 1 Comparison with current predictors of catalytic residues

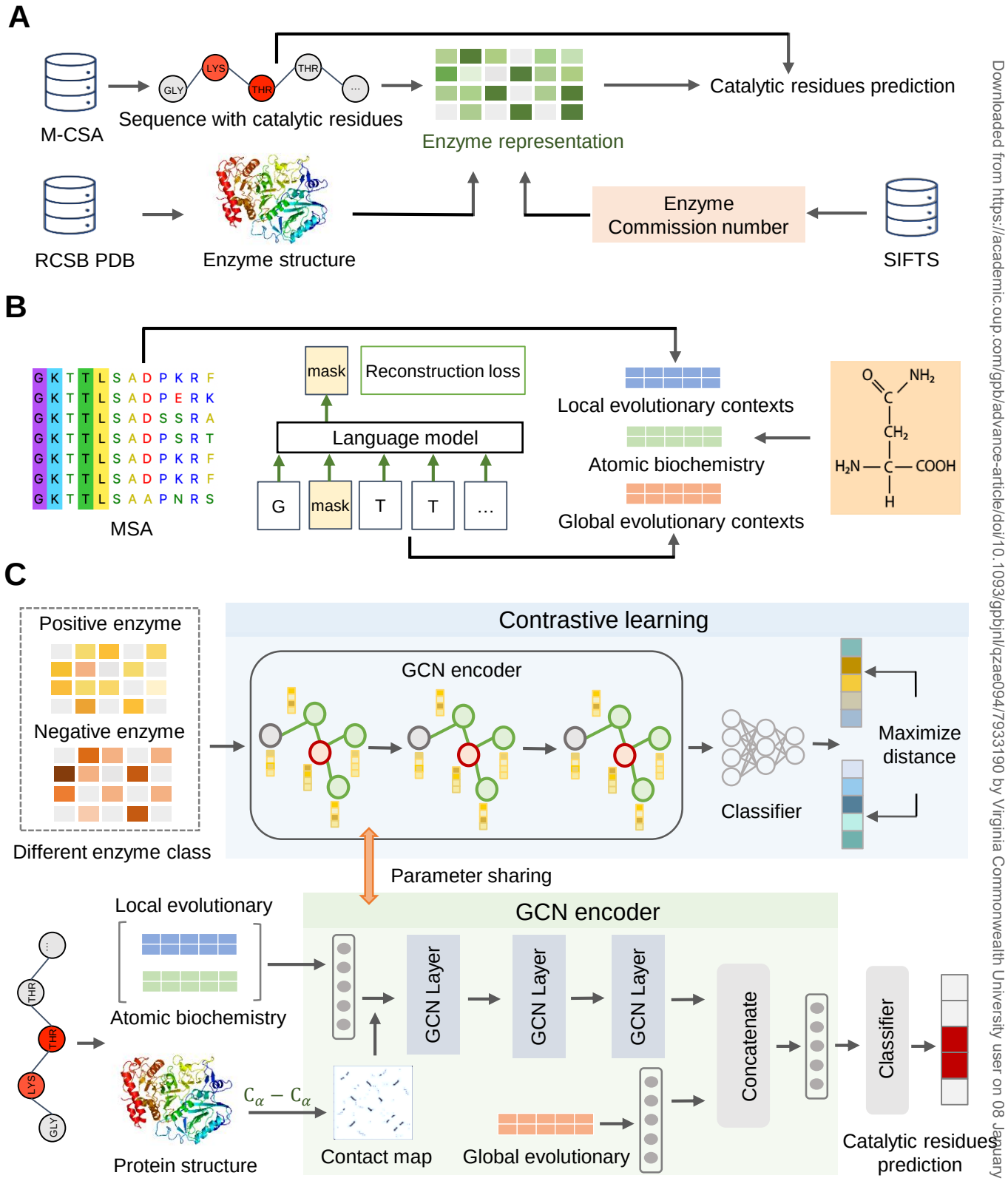
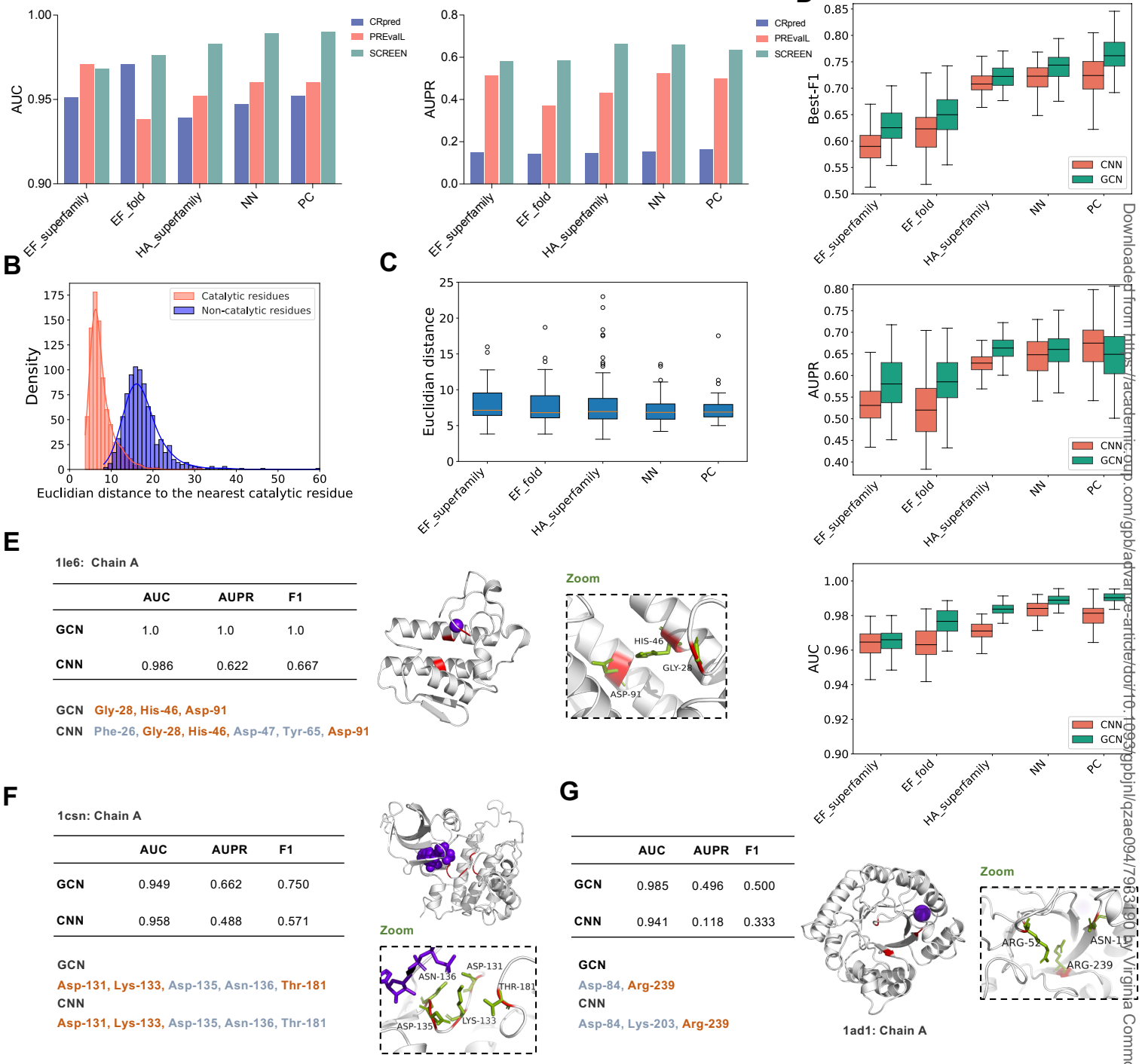


Figure 2



E

1le6: Chain A

	AUC	AUPR	F1
GCN	1.0	1.0	1.0
CNN	0.986	0.622	0.667

GCN **Gly-28, His-46, Asp-91**

CNN Phe-26, **Gly-28, His-46,** Asp-47, Tyr-65, **Asp-91**

F

1csn: Chain A

	AUC	AUPR	F1
GCN	0.949	0.662	0.750
CNN	0.958	0.488	0.571

GCN

Asp-131, Lys-133, Asp-135, Asn-136, **Thr-181**

CNN

Asp-131, Lys-133, Asp-135, Asn-136, Thr-181

G

	AUC	AUPR	F1
GCN	0.985	0.496	0.500
CNN	0.941	0.118	0.333

GCN

Asp-84, **Arg-239**

CNN

Asp-84, Lys-203, **Arg-239**

1ad1: Chain A

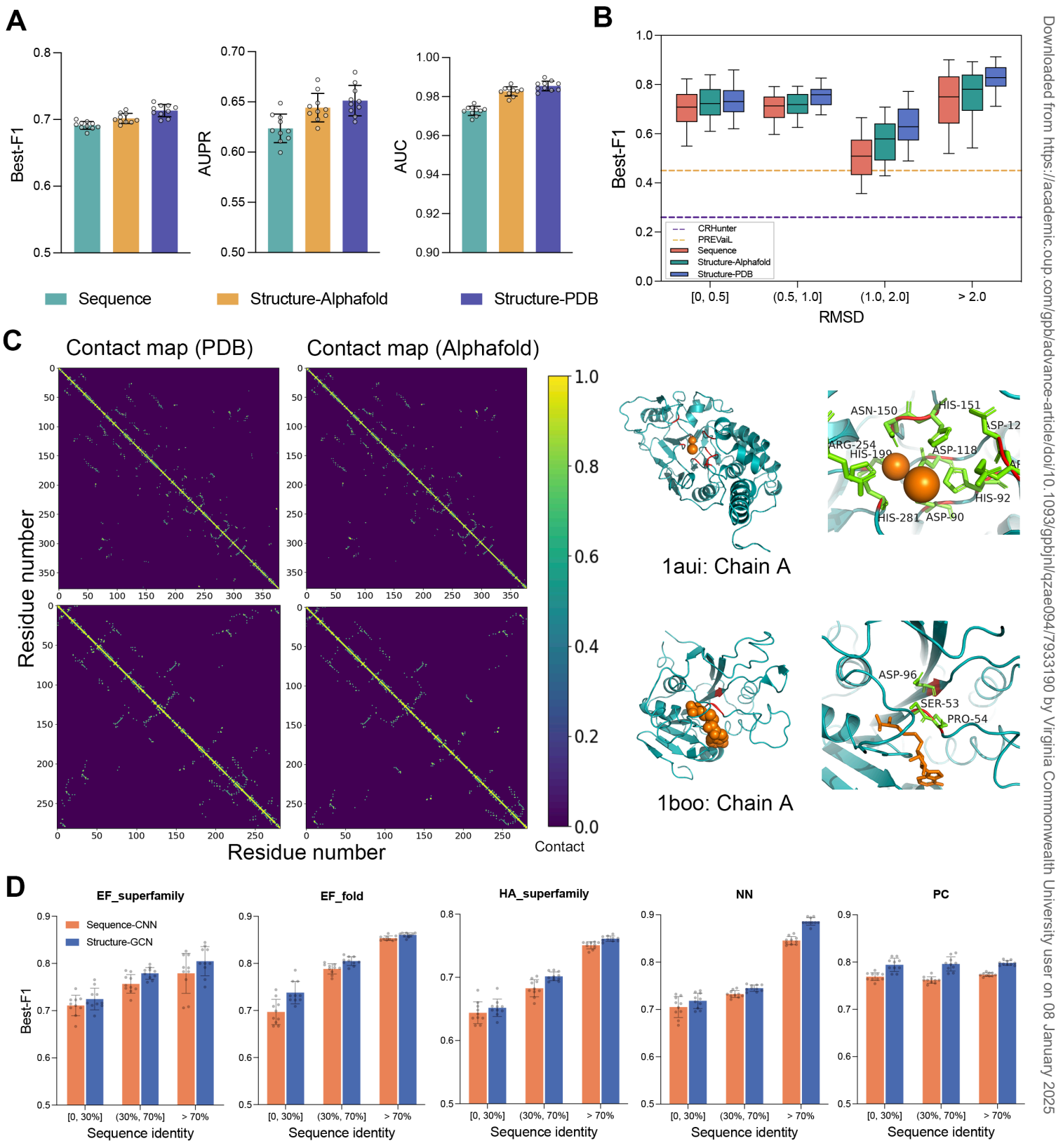
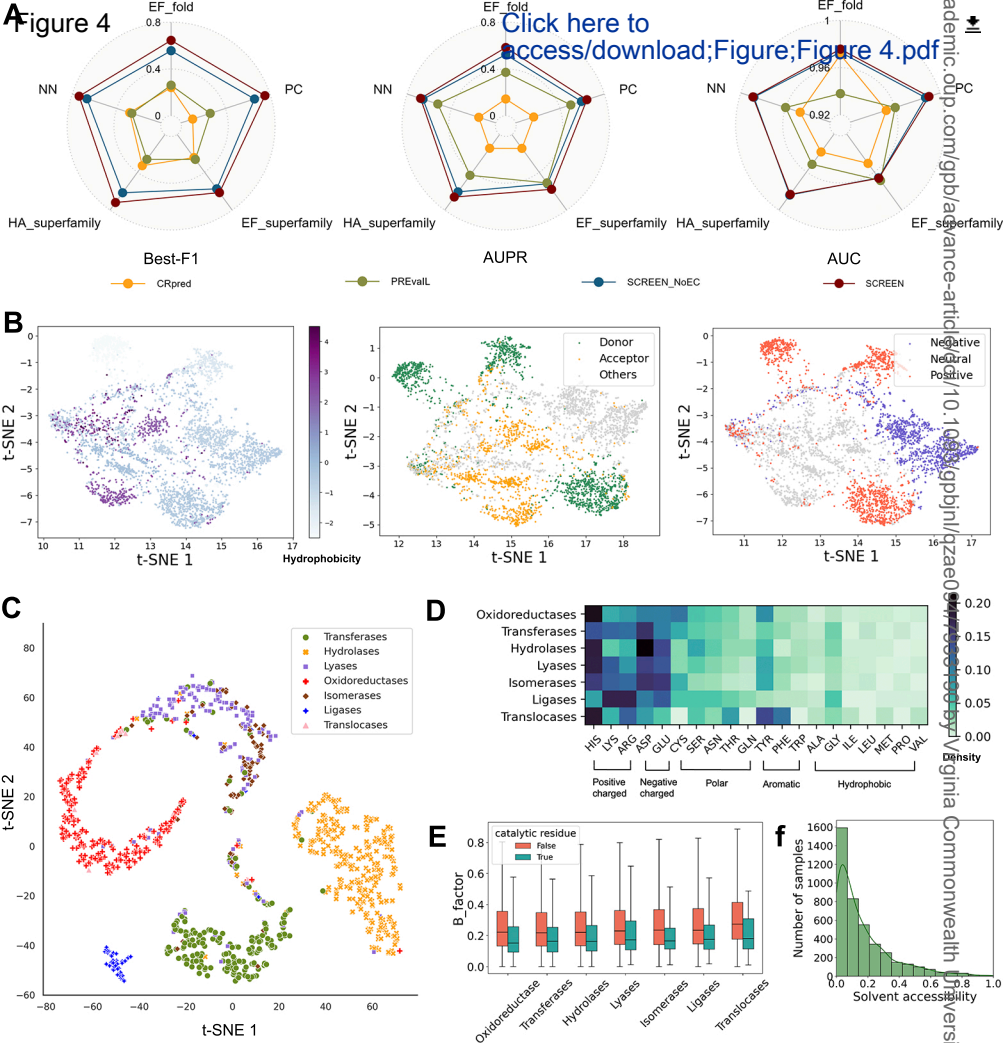


Figure 4



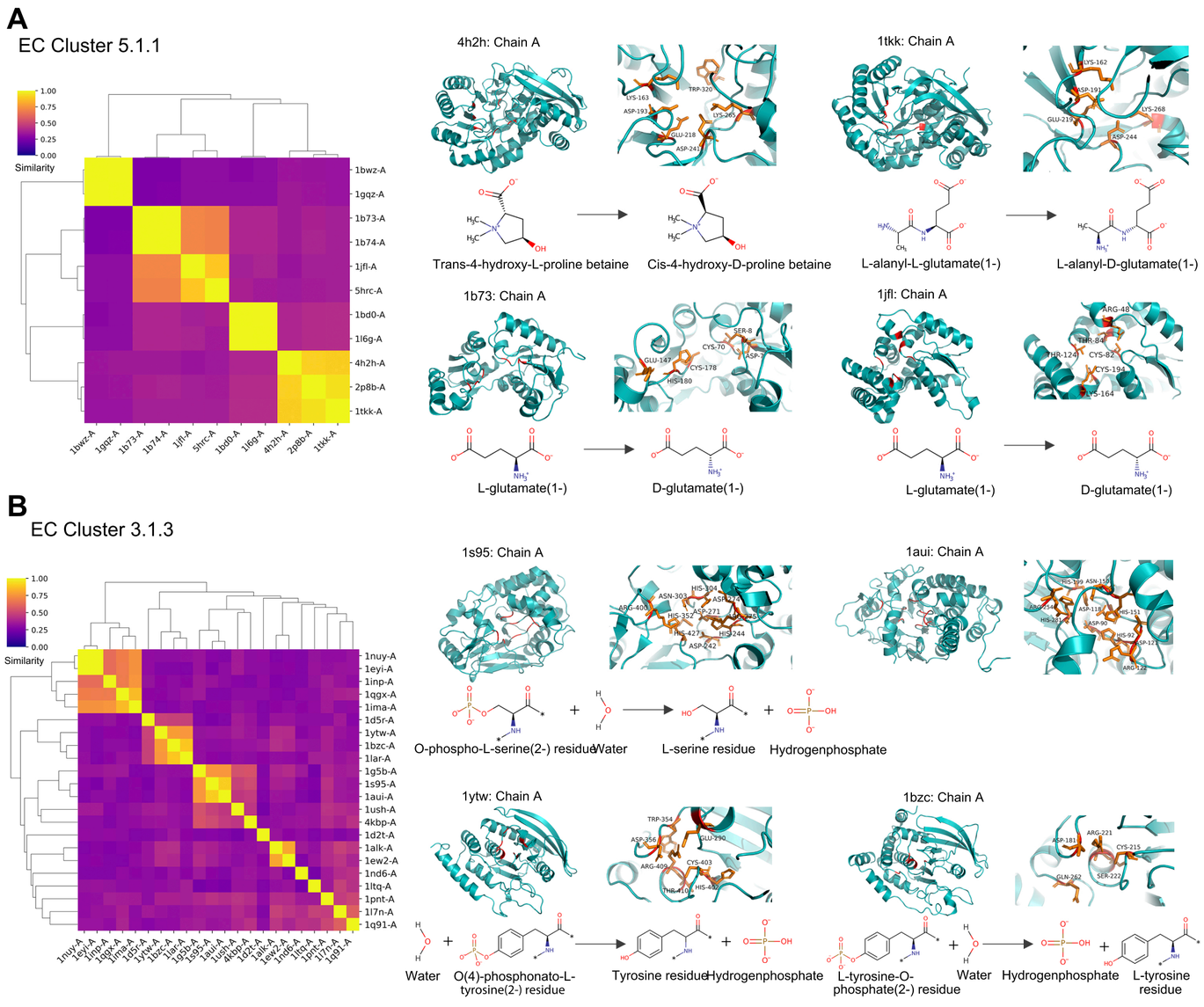


Figure 6

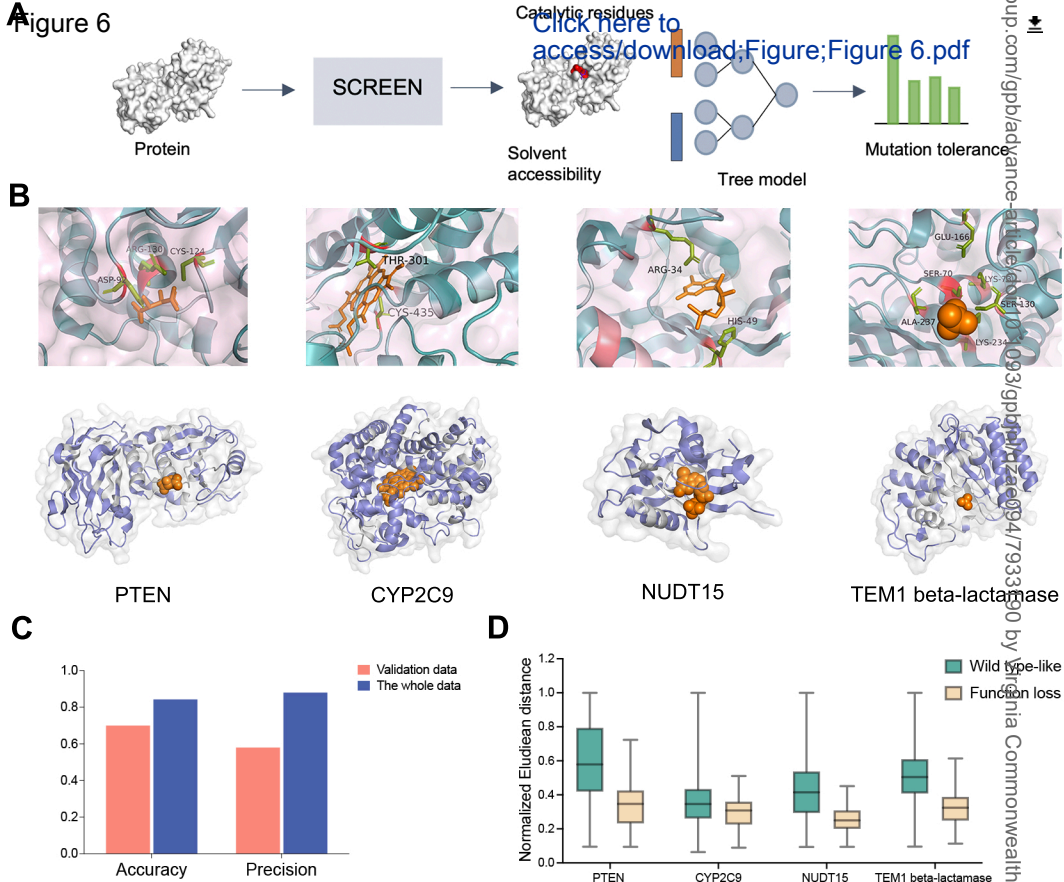


Table 1 Comparison with current predictors of catalytic residues

Methods	Measurement (%)	EF Superfamily dataset	EF fold dataset	HA superfamily dataset	NN dataset	PC dataset
Methods from Youn et al.[22]; Chea et al.[24]; Gutteridge et al.[18]; Petrova and Wu [20]	Precision (Recall)	16.9 (53.9)[a]	17.1 (51.1)[a]	16.5 (29.3)[b]	56.0 (14.0)[c]	7.0 (90.0)[d]
	F1	25.7	25.6	21.1	22.4	13.0
CRpred	Precision (Recall)	15.9 (52.1)	16.1 (48.0)	24.7 (49.7)	65.9 (18.0)	5.6 (84.5)
	F1	24.4	24.1	33	28.3	10.5
CRHunter	Precision (Recall)	21.5 (68.7)	21.0 (62.7)	33.2 (69.6)	76.4 (24.0)	7.6 (92.1)
	F1	32.8	31.5	45.0	36.5	14.0
PREvaIL	Precision (Recall)	17.0 (59.4)	17.0 (56.5)	17.0 (57.9)	58.9 (17.0)	17.0 (58.1)
	F1	26.4	26.1	26.3	26.4	26.3
AEGAN	Precision (Recall)	31.7 (85.7)	31.0 (83.7)	29.8 (85.7)	29.6 (83.9)	28.6 (86.6)
	F1	45.9	44.6	43.6	43.3	41.3
SCREEN (This study)	Precision (Recall)	61.9 (61.2)	61.0 (68.4)	69.3 (74.8)	68.5 (79.9)	67.6 (82.0)
	F1	61.5	64.5	72.0	73.8	74.1

Note: [a] Model performance on EF superfamily and EF fold datasets by Youn et al. [22]; [b] Model performance on HA family dataset by Chea et al. [24]; [c] Molde performance on NN dataset using the structure-based method without spatial clustering by Gutteridge et al. [18]; [d] Model Performance on PC dataset by Petrova and Wu [20].