PREvaIL, an integrative approach for inferring catalytic residues using sequence, structural, and network features in a machine-learning framework

著者	Song Jiangning, Li Fuyi, Takemoto Kazuhiro, Haffari Gholamreza, Akutsu Tatsuya, Chou Kuo-Chen, Webb Geoffrey I.
journal or publication title	Journal of Theoretical Biology
volume	443
page range	125-137
year	2018-02-01
その他のタイトル	PREvaIL, an integrative approach for inferring catalytic residues using sequence, structural and network features in a machine learning framework
URL	http://hdl.handle.net/10228/00008223

doi: https://doi.org/10.1016/j.jtbi.2018.01.023

PREvaIL, an integrative approach for inferring catalytic residues using sequence, structural and network features in a machine learning framework

Jiangning Song[1,2,*], Fuyi Li[2], Kazuhiro Takemoto[3], Gholamreza Haffari[1], Tatsuya Akutsu[4], Kuo-Chen Chou[5,6,7,*], and Geoffrey I. Webb[1,*]

[1]Monash Centre of Data Science, Faculty of Information Technology, Melbourne, VIC 3800, Australia

[2]Department of Biochemistry and Molecular Biology and Biomedicine Discovery Institute, Monash University, Melbourne, VIC 3800, Australia

[3]Department of Bioscience and Bioinformatics, Kyushu Institute of Technology, Iizuka, Fukuoka 820-8502, Japan

[4]Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011, Japan

[5]Gordon Life Science Institute, Boston, MA 02478, USA

[6]Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu, 610054, China

[7]Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah 21589, Saudi Arabia

To be submitted to ***Journal of Theoretical Biology*** (A* journal) as a regular research paper

*To whom correspondence should be addressed: Jiangning.Song@monash.edu; kcchou@gordonlifescience.org; Geoff.Webb@monash.edu

Abstract:

Determining the catalytic residues in an enzyme is critical to our understanding the relationship between protein sequence, structure, function, and enhancing our ability to design novel enzymes and their inhibitors. Although many enzymes have been sequenced, and their primary and tertiary structures determined, experimental methods for enzyme functional characterization lag behind. Because experimental methods used for identifying catalytic residues are resource- and labor-intensive, computational approaches have considerable value and are highly desirable for their ability to complement experimental studies in identifying catalytic residues and helping to bridge the sequence-structure-function gap. In this study, we describe a new computational method called PREvaIL for predicting enzyme catalytic residues. This method was developed by leveraging a comprehensive set of informative features extracted from multiple levels, including sequence, structure, and residue-contact network, in a random forest machine-learning framework. Extensive benchmarking experiments on eight different datasets based on 10-fold cross-validation and independent tests, as well as side-by-side performance comparisons with seven modern sequence- and structure-based methods, showed that PREvaIL achieved competitive predictive performance, with an area under the receiver operating characteristic curve and area under the precision-recall curve ranging from 0.896–0.973 and from 0.294–0.523, respectively. We demonstrated that this method was able to capture useful signals arising from different levels, leveraging such differential but useful types of features and allowing us to significantly improve the performance of catalytic residue prediction. We believe that this new method can be utilized as a valuable tool for both understanding the complex sequence-structure-function relationships of proteins and facilitating the characterization of novel enzymes lacking functional annotations.

Keywords:
enzyme catalytic residues; sequence-structure-function relationship; functional annotation; bioinformatics; pattern recognition; machine learning; sequence analysis

1 INTRODUCTION

As powerful biological catalysts, enzymes can effectively catalyze biochemical reactions at extremely high rates and are thus indispensable for many biological processes and pathways (Khosla and Harbury, 2001). Many important findings acquired from enzyme fast reaction systems (Chou and Zhou, 1982; Kuo-chen and Shou-ping, 1974; Zhou and Zhong, 1982) significantly impact both basic research (Gardner, et al., 2015) and drive changes in medicinal chemistry (Chou, 2017). However, the residues comprising an enzyme differ greatly in functional significance, with only a small number directly involved in catalytic activity (Furnham, et al., 2014). Accordingly, understanding which of these are catalytic residues is critical for our determining relationships between protein sequence, structure, function, and enhancing our ability to design novel inhibitors and enzymes. This has important implications in the post-genomic era, with its challenge of bridging the widening protein sequence-structure gap. Although sequence information for many enzymes is known, relatively few enzymes have been functionally characterized. Therefore, detailed information regarding catalytic residues and enzyme active sites explicitly involved in catalysis remains lacking. Because experimental methods for identifying catalytic residues are resource- and labor-intensive, high-throughput *in silico* approaches have considerable value and are highly desirable for complementing experimental efforts in identifying catalytic residues and helping to bridge the sequence-structure-function gap.

In recent years, a variety of computational methods have been developed for predicting catalytic residues or functional residues involved in catalytic reactions (Chou and Cai, 2004). These methods differ in several ways, including in the machine-learning or statistical-scoring technique used, the types of sequence features used, whether or not structural features are used in addition to sequence features, and in the sources of training and testing datasets. According to the types of features used for constructing prediction models, existing methods can be generally categorized into four major groups.

The first group of methods was primarily developed based on protein sequence and typically relied upon extracting useful sequence features for inputs used to train the prediction models. Commonly used sequence features include evolutionary information in the form of position-specific scoring matrices (PSSMs) or sequence conservation inferred from multiple sequence alignments (Capra and Singh, 2007; Fischer, et al., 2008; La, et al., 2005; Pai, et al., 2015; Youn, et al., 2007; Zhang, et al., 2008) or other sequence-derived features, such as Jensen-Shannon divergence scores, relative entropies (Dou, et al., 2012; Dou, et al., 2010; Fischer, et al., 2008), and predicted structural information inferred from sequences, including secondary structure and solvent accessibility (Dou, et al., 2012; Kauffman and Karypis, 2009; Shen, et al., 2009).

In recent years, many research groups exploited the increasing quantity of structural data deposited in the Protein Data Bank (PDB) (Rose, et al., 2017), prompting the proliferation of the second group of methods, which leverage structural information to build the prediction models (Alterovitz, et al., 2009; Chea and Livesay, 2007; Cilia and Passerini, 2010; Gutteridge, et al., 2003; Han, et al., 2012; Kirshner, et al., 2013; Panchenko, et al., 2004; Petrova and Wu, 2006; Sun, et al., 2016; Xin, et al., 2010; Youn, et al., 2007). Xin et al. (2010) proposed a structure-based kernel algorithm for the prediction of catalytic residues by explicitly modelling the similarity between residue-centered neighborhoods in protein structures (Xin, et al., 2010). They showed that the geometry, physicochemical properties, and evolutionary conservation play an important role in determining catalytic residue activity. In a recent study, Sun et al. (2016) developed the CRHunter method which combined both sequence and structural information in an SVM framework that achieved stable performance when compared with other template-based predictors (Sun, et al., 2016). Chien and Huang proposed an approach EXIA based on residue side chain orientation and backbone flexibility of protein structure, which achieved a comparable performance to that of evolutionary sequence conservation (Chien and Huang, 2012). In another study, Kirshner et al. (2013) developed the Catsid (Catalytic site identification) search engine, which enables rapid searches for structural matches to a user-specified catalytic site among all PDB structures. Its capacity to rapidly search all known protein structures in the PDB is enabled by a logistic regression-based model that allows for systematic identification of true positives based on a set of feature descriptors (Kirshner, et al., 2013).

The third group of methods (Chea and Livesay, 2007; del Sol, et al., 2006; del Sol and O'Meara, 2005; Li, et al., 2011) involve graph-theoretical methods that essentially rely on representing protein three-dimensional (3D) structures as small world networks (Watts and Strogatz, 1998), where amino acid residues specify vertices within a graph while two residues in a proximal spatial neighborhood form edges. Zhou et al. provided a comprehensive review on recent progress in this area (Zhou, et al., 2016). Previous studies showed that representing protein structure as a topological residue-contact network can provide novel insights into protein folding mechanisms, stability, and function (del Sol, et al., 2006; del Sol and O'Meara, 2005; Jiao and Ranganathan, 2017; Song, et al., 2010; Tang, et al., 2008; Wang, et al., 2012; Zheng, et al., 2012). Chea and Livesay (2007) benchmarked the performance of one particular network measure called closeness centrality and showed that it provided statistically significant predictive power for catalytic residue predictions. They also demonstrated that solvent accessibility or residue identity could be used as an efficient filter by this network feature to further improve its predictive performance (Chea and Livesay, 2007).

The fourth group of methods uses heterogeneous features through the integration or fusion of sequence, structure, and other types of features (Li, et al., 2011; Sankararaman, et al., 2010; Tang, et al., 2008). Because the extracted features are heterogeneous, redundant, and noisy, a number of feature-selection and dimensionality reduction algorithms are often employed and used in combination with the learning algorithms to remove irrelevant features and improve model training in order to increase prediction accuracy. In terms of the algorithms used for training these prediction models, machine learning or statistical scoring approaches are often employed and used include neural networks (Gutteridge, et al., 2003), information-theoretic algorithms (Capra and Singh, 2007; Fischer, et al., 2008), genetic algorithms (Izidoro, et al., 2015), support vector machines (SVMs) (Chea and Livesay, 2007; Li, et al., 2011; Pai, et al., 2015; Petrova and Wu, 2006; Sun, et al., 2016; Youn, et al., 2007), kernel-based algorithms (Xin, et al., 2010), AdaBoost (Alterovitz, et al., 2009), and logistic regression (Dou, et al., 2012; Kirshner, et al., 2013; Sankararaman, et al., 2010). The consensus of these studies has been that evolutionary information, sequence conservation, and the structural neighborhood of catalytic residues are important predictive features, with machine learning-based approaches often providing competitive performance, making them particularly suitable for dealing with high-dimensional heterogeneous feature spaces.

Despite the development and increasing availability of such a wide range of methods, three main challenges need to be overcome to predict catalytic residues by machine leaning-based approaches: (1) Sequence and structural features are still not sufficient to predict the catalytic residues of certain proteins. Accordingly, it is necessary to find and exploit other novel and complementary groups or types of features that can be used to further improve prediction performance. (2) Methods for quantifying and characterizing the relative importance and contribution of each group of features according to model performance are needed. (3) It is necessary to determine which machine learning algorithm provides the overall highest and most reliable prediction performance.

To address these questions, in this study, we present a new machine learning-based approach called **PREvaIL** (PRotEin various Information-based cataLytic site predictor) for predicting catalytic residues based on a random forest (RF) algorithm. In terms of input features, this approach combines a variety of sequence and structural features, as well as residue-contact-network properties, and uses an efficient feature-selection technique to select a subset of more useful features for catalytic residue prediction. We performed extensive benchmark experiments using eight different test datasets to evaluate the performance of this approach and compared it with other competing methods. The results showed that this new approach performed favorably as compared with other methods, thereby illustrating its effectiveness.

2 MATERIALS and METHODS

According to the 5-step rules (Chou, 2011), the first important step in developing a new predictor involves construction or selection of an effective benchmark dataset. In this study, we addressed this problem as follows.

2.1 Design and generation of independent test datasets of catalytic residues

To comprehensively evaluate the predictive performance of our approach and compare it with other available methods, we employed the same nine datasets prepared by previous studies (Chea and Livesay, 2007; Gutteridge, et al., 2003; Petrova and Wu, 2006; Youn, et al., 2007; Zhang, et al., 2008). These nine datasets are briefly introduced below.

The first six of these datasets were carefully curated based on various levels of sequence homology and only contained one sequence per fold, family, and superfamily (Zhang, et al., 2008), thereby allowing rigorous and unbiased performance comparison between different methods. These curated datasets are the SCOP fold dataset (“EF_fold”), SCOP superfamily dataset (“EF_superfamily”), and SCOP family dataset (“EF_family”) originating from Youn et al. (2007), SCOP superfamily dataset (“HA_superfamily”) prepared by Chea and Livesay (2007), the dataset (“PC”) prepared by Petrova and Wu (2006), and a non-homologous dataset (“NN”) prepared by Gutteridge et al. (2003).

The ST-1109 dataset was originally prepared by the Kurgan group (Zhang, et al., 2008). In this dataset, all experimentally verified catalytic residues were extracted from the Catalytic Site Atlas (CSA) database (Furnham, et al., 2014; Porter, et al., 2004), which is a comprehensive resource for catalytic sites and residues identified in enzymes using structural data. The CD-HIT program (Fu, et al., 2012) was applied at a sequence identity cut-off of 40% to filter homologous sequences against those in the six aforementioned datasets in order to avoid bias introduced by homologous sequences used for independent tests. This final design dataset contained 1,109 PDB chains.

The T-124 and T-37 datasets were also originally prepared by the Kurgan group (Zhang, et al., 2008) and used as independent test datasets to evaluate the performance of different methods. The two datasets contain 124 and 37 PDB chains, respectively, and used the CSA database (version 2.2.5) (Furnham, et al., 2014; Porter, et al., 2004) to annotate the catalytic residues in each chain. The two datasets have low pairwise sequence identity (<30%) with respect to the two training datasets EF_fold and ST-1109 design set, which would be used by our method to train the RF models.

As the most comprehensive dataset, the ST-1109 design dataset was used to train the prediction models used by our method, select optimal features, and calibrate model parameters. The EF_fold dataset was used to train the models of our method, which was then tested using the T-124 and T-37 independent test datasets to compare the performance between different methods. Additionally, 10-fold cross-validation tests on the six curated datasets were also performed to assess the performance of other existing methods.

2.2 Overview of the PREvaIL methodology

The flowchart of our PREvaIL methodology for predicting catalytic residues based on the integration of sequence, structural, and residue-contact-network features is shown in **Figure 1**. There exist several major stages of developing the PREvaIL methodology, including dataset curation, feature extraction at the sequence, structure, and network levels, feature selection, model construction, and performance evaluation. Except for dataset curation, which was described in section 2.1, we discuss each of these major stages in the following sections.

2.3 Feature engineering

Similar to previous studies, we formalized the catalytic residue prediction as a classification problem. For this purpose, each candidate catalytic residue was represented by a feature vector, **x,** with ***D***-dimensional feature components, $\{x_1, x_2,\ldots, x_D\}$. The aim of this classification problem is to predict the label y given the feature representations of a residue, i, i.e., to predict whether a residue is catalytic (y = 1) or noncatalytic (y = 0). **Table 1** provides a comprehensive list of all feature components, $\{x_1, x_2,\ldots, x_D\}$, categorized according to 11 different feature groups and arranged in the order of feature encoding. A series of 3,424-dimensional feature vectors were extracted and used in this study. In the following sections, we describe in detail each feature type used for extracting feature-vector components.

2.3.1. Sequence features

With the explosive growth of biological sequence data generated in the post-genomic age, one of the most important but also most difficult problems in computational biology is formulation of a biological sequence using a discrete model or vector while maintaining considerable sequence order or pattern information. This is because all existing operation engines, such as SVMs and RF algorithms, can only handle vectors, but not sequence samples (Chou, 2015). However, a vector defined by a discrete model might lose all of the sequence-pattern information. To avoid complete loss of sequence-pattern information for proteins, the pseudo amino acid composition (PseAAC) (Chou, 2001; Chou, 2005) was proposed and has subsequently been utilized in many biomedicine and drug development areas (Zhong and Zhou, 2014), as well as all areas of computational proteomics [e.g., ((Khan, et al., 2017; Meher, et al., 2017; Rahimi, et al., 2017; Tahir, et al., 2017; Zhou, et al., 2007) and a long list of references previously cited (Amanzadeh, et al., 2014; Behbahani, et al., 2016; Beigi, et al., 2011; Esmaeili, et al., 2010; Hajisharifi, et al., 2014; Khosravian, et al., 2013; Mohabatkar, 2010; Mohabatkar, et al., 2011; Mohabatkar, et al., 2013; Mousavizadegan and Mohabatkar, 2016; Poorinmohammad, et al., 2015). Because of its wide use, three powerful open access software tools called "PseAAC-Builder" (Du, et al., 2012), "propy" (Cao, et al., 2013), and "PseAAC-General" (Du, et al., 2014) were established. The former two are used to generate various modes of special PseAACs, whereas the latter the general PseAAC, including not only all special modes of feature vectors for proteins but also higher level feature vectors, such as "Functional Domain," "Gene Ontology," and "Sequential Evolution" or "PSSM" modes. In this study, we considered the following five different sequence-derived feature types.

1) PSSM. Evolutionary information in the form of a PSSM (Jones, 1999) is particularly useful for improving the predictive performance of machine learning-based models in our previous studies,

including prediction of cis/trans isomerization (Song, et al., 2006), disulfide connectivity (Song, et al., 2007), protease cleavage sites (Song, et al., 2012), and metal-binding sites (Chen, et al., 2013; Song, et al., 2017). We used a sliding window comprised of 13 amino acids to extract PSSM features, resulting in a 20 × 13 = 260-dimensional vector.

2) EntWOP. This feature is calculated based on the Shannon entropy of the weighted observed percentages (WOPs) generated by performing three iterations of PSI-BLAST search and was originally designed by incorporating the sequence conservation information into the PSI-BLAST profiles (Zhang, et al., 2008). This feature was included based on its emergence as among the most important features used for predicting catalytic residues. We followed the same procedures as described by Zhang et al. (2008) to extract this feature.

3) Dist_Key. This one-dimensional feature describes the relative sequence distance of a catalytic residue relative to the N-terminus of the protein sequence.

4) Physicochemical property. Catalytic residues can be classified as charged, hydrophobic, or polar residues according to their physicochemical properties. We used binary encoding to represent this feature type (i.e., using a 3-dimensional binary vector; if the given catalytic residue is charged, the first dimension is set to "1," whereas the other two dimensions are set to "0").

5) CRPair. An enzyme normally has three catalytic residues based on the requirement to form a catalytic triad at the center of the active site (Carter and Wells, 1988). The notation of CRPair was first introduced by Zhang et al. and is defined as a pair of catalytic residues (Zhang, et al., 2008). The sequence distances between any two adjacent catalytic residues are all calculated and collectively encoded for the central catalytic residue. A total of 76 CRPairs were extracted in this study and we used binary encoding to represent each CRPair.

2.3.2. Structure features

To complement sequence-derived features and improve the predictive performance of our models, we also extracted a variety of structure features, including the following: 1) eight different types of secondary structures calculated using the DSSP program (Kabsch and Sander, 1983) and including the 3_{10} helix (denoted as G in DSSP), α helix (H), π helix (I), beta bridge (B), beta bulge (E), turns (T), high curvature region (S), and loops (C); 2) solvent accessibilities of all-atoms, total-side, main-chain, non-polar, and all-polar residues, which were calculated using the NACCESS program (Hubbard and Thornton, 1993); 3) solvent exposure features to include half sphere exposure (HSE), contact number (CN), and residue depth (RD), with HSE a two-dimensional measurement of the solvent exposure of a residue (Hamelryck, 2005; Song, et al., 2008). HSE features include HSEAU, HSEAD, HSEBU, and HSEBD, with the former two calculated using the Cα coordinates, whereas the latter two were calculated using the Cβ coordinates. We used the Biopython package (Cock, et al., 2009) to calculate the six solvent exposure features; 5) B-factor (atomic displacement parameter) measures residue mobility and reflects fluctuations of an atom in a crystallographic structure (Yuan, et al., 2005) based on its indication of residue flexibility and dynamics. We extracted the original B-factor value for each catalytic residue from its corresponding PDB structure and then normalized it using a previously described method (Smith, et al., 2003) prior to its being encoded as a feature vector and used as the input.

2.3.3. Residue contact network features

A protein can be represented as a connected network of contacting residues in the 3D structure space (del Sol and O'Meara, 2005). By representing protein structures as small world networks

(Watts and Strogatz, 1998), a number of important characteristic features can be extracted from the topology of the protein 3D structures to facilitate the identification of functionally important residues involved in both enzyme and nonenzyme protein families (del Sol, et al., 2006). Here, we defined two residues in a protein structure as in contact if the distance between their center points was ≤6.5 Å. This allowed conversion of a PDB structure into a residue contact network (Wang, et al., 2012). We used the iGraph network analysis package (Csardi and Nepusz, 2006) to calculate different network properties describing the local environment of the catalytic residue in the residue contact network, including degree, closeness, status, hubscore, clustering coefficient, cyclic coefficient, constraint, betweenness, eigenvector, cocitation, coreness, and eccentrality.

2.3.4. Network neighboring properties

Considering that neighboring nodes might affect the spatial arrangement and organization of the central node due to their differential distance to the central node in the network, the neighboring property of node i, $\phi(i)$, can be defined as follows to take this effect into consideration:

$$\phi(i) = \frac{1}{N-1}\sum_{j\neq i}\frac{f(j)}{d(i,j)},$$

where *f(j)* is a given property of node j, *d(i,j)* is the shortest path length between nodes *i* and j, and *N* is the total number of residues in a protein structure. To the best of our knowledge, this represents the first encoding and use of this form of network neighboring properties.

We encoded the network features in two different ways. One involves encoding residue-contact-network metrics as input features, and the other involves encoding network neighboring properties for a central residue as an input feature after representing the entire protein structure as the residue contact small world network.

2.4. RF algorithm

RF is an ensemble tree-structured algorithm used for classification and regression analyses (Breiman, 2001) and has been implemented as the randomForest package in R (Liaw and Wiener, 2002) and widely used in computational biology (Jia, et al., 2015; Jia, et al., 2016; Liu, et al., 2016; Qiu, et al., 2016). A typical RF model consists of hundreds of decision trees and uses majority voting to determine the final prediction outcome for unseen data samples. Compared with other machine-learning algorithms, RF has several attractive advantages that make it suitable for dealing with the current prediction task. It usually performs favorably and stably with high-dimensional feature vectors, which is particularly the case in this work. The model training process is often faster than that of other machine-learning algorithms, such as SVMs and neural networks. Importantly, this also permits variable or feature selection, thereby providing the opportunity to characterize important features that contribute the most to model performance. Additionally, use of RF includes both model training and prediction stages, which is similar to many other machine-learning algorithms.

2.5. Feature selection based on the RF mean decrease Gini index (MDGI)

RF provides a feature-selection method based on the MDGI, which can be calculated by the randomForest R package (Liaw and Wiener, 2002). The MDGI score measures the importance of individual vector elements of a feature for correctly classifying a residue as catalytic or noncatalytic. The mean MDGI was calculated as the averaged MDGI over 100 trials of randomly classifying a set of positive (i.e., catalytic residues) and negative residues (i.e., noncatalytic residues) with a ratio of

1:1. The mean MDGI Z-Score of each vector element was then calculated as:

$$MDGI\ \ Z - score = \frac{x_i - \bar{x}}{\sigma}$$

where x_i is the mean MDGI of the i-th feature, and σ is the standard deviation. We divided the vector elements into four zones according to their MDGI Z-Scores (i.e., Z-score > 2, 1.5 < Z-score ≤ 2, 1< Z-score ≤ 1.5, and 0.5 < Z-score ≤ 1, respectively). Vector elements with MDGI Z-Score s> 2.0 were considered as optimal feature candidates and used as the input features to train the RF classifier.

2.6. Performance evaluation by cross-validation and independent testing

We used several standard performance measures, including precision (PRE), recall (REC), false positive (FP) rate, the area under the curve (AUC), and the area under the recall-precision Curve (AURPC), to comprehensively evaluate and compare the predictive performance between different methods. Among these measures, AUC represents the area under the receiver operating characteristic (ROC) curve, which is a plot of the true positive (TP) rate against the FP rate, whereas AURPC is the area under the Recall-Precision curve (RPC) and used as a good alternative to AUC if there is a large skew in the class distribution. Both AUC and AURPC were used as primary measures to assess the predictive performance of different methods using the eight aforementioned independent test datasets.

PRE is defined as:

$$PRE = \frac{TP}{TP + FP}$$

REC (also referred to as the TP rate) is defined as:

$$REC = \frac{TP}{TP + FN}$$

FP rate is defined as:

$$FP\ rate = \frac{FP}{FP + TN}$$

where *TP* is the number of TPs, *TN* is the number of true negatives, *FP* is the number of FPs, and *FN* is the number of false negatives.

3 RESULTS AND DISCUSSION

3.1. Feature ranking by the MDGI Z-Score

We calculated and ranked the MDGI Z-Scores of all initial 3,424 features (See Table 1 for a summary of these features) using the randomForest R package in order to assess the relative importance and contribution of each feature type. As a result, we identified a total of 127 feature-vector elements with MDGI Z-score > 1.0, of which 41 had an MDGI Z-score > 2.0. The relative importance and ranking of these feature vectors are plotted in **Figure 2**. A detailed list of these feature vectors according to their MDGI Z-Score zones are provided in **Supplementary Table S1**.

To better understand the interrelationships between the significant sequence, structure, and network-based features, we performed classical multi-dimensional scaling (Lobley, et al., 2007) and visualized the distributions of these features in the feature space (**Figure 3**). Feature descriptors that are closely correlated tend to be closely clustered together in the feature space. As shown, there existed four clearly defined feature groupings. The first group included two solvent exposure measures, HSEAD and HSEBD, and several other network features, such as degree, closeness_centrality, cocitation, coreness, hubscore, and eigen_centrality. The second group contained all polar, total_side, main_chain, non-polar solvent accessibility, and two network features, cyclic_coeff and cluster_coeff. The third group included CN, RD, HSEAU, HSEBU, and betweenness centrality. The remaining features formed the fourth group, which included B-factor, constraint, Dist_Key, eccentrality, and status.

As expected, the feature interrelationships revealed by the multi-dimensional scaling plot in **Figure 3** agreed well with their pairwise Pearson's correlation coefficients. For example, five network features that were closely related to the degree feature in the feature space were coreness, cocitation, closeness_centrality, eigen_centrality, and hubscore, with Pearson's correlation coefficient values of 0.896, 0.896, 0.407, 0.364, and 0.364, respectively (See the **Supplementary Excel file 1 Correlation matrix between significant features** for the calculated Pearson's correlation coefficients between every two features). The observed correlations between these features indicated they encoded similar information within this feature grouping, and that such information was not represented by other feature groupings. Similar observations were noted for the second, third, and fourth feature groupings, as shown in **Figure 3** and **Supplementary Excel file 1**.

3.2. Analysis of the importance and relevance of different feature types

We performed unpaired two-sample *t* tests to examine whether the mean values of a given feature between catalytic residues and randomly selected noncatalytic residues were statistically significant in order to assess the potential of the given feature for discriminating the two sample sets. The results are shown in **Figures 4** and **S1**, with the mean values, standard deviations, and *P*-values listed in **Supplementary Table S2**. For the majority of features, the mean values between catalytic residues and noncatalytic residues were statistically significant, with most having a *P*-value < 1.7E-05. The boxplots for some of the selected features are shown in **Figures 4** and **S1**.

To investigate and assess the contribution of a variety of sequence, structural, and residue-contact-network features to catalytic residue prediction, we examined their contributions to gain insight into the ability of PREvaIL to discriminate between catalytic and noncatalytic residues. As shown in **Figure 2** and **Supplementary Table 1**, among the 3,424 features initially extracted, 41 had MDGI Z-scores >2.0 and were used in the final RF models. These features were distributed in eight specific types, including network feature (closeness centrality), Dist_Key, a number of neighboring properties, PSSM, EntWOP, CN, HSEAD, HSEBD, solvent accessibility, and amino acid physicochemical properties (charge and hydrophobicity).

These top ranked features are frequently associated with features identified in previous studies as highly correlated with catalytic residues. These include closeness centrality, which is a network centrality feature describing the status of a residue located in the protein structure. Highly central residues tend to have higher closeness values due to their interaction with a relatively larger number of residues in the structure space. Closeness was highly correlated with catalytic residues (Amitai, et al., 2004; Chea and Livesay, 2007; del Sol, et al., 2006; del Sol and O'Meara, 2005; Li, et al., 2011), and our results confirmed this observation.

However, as an enriched feature source, sequence-derived features have been extensively used in model training and crucial for ensuring model performance in a number of previous bioinformatics studies focusing on prediction of protein structural and functional properties (Li, et al., 2015; Li, et al., 2014; Song, et al., 2009; Song, et al., 2017; Wang, et al., 2014) (Disfani, et al., 2012; Jones and Cozzetto, 2015; Lobley, et al., 2007; Meng and Kurgan, 2016; Ofran and Rost, 2007; Zhang, et al., 2008). In our investigation, we focused on sequence-derived features, such as evolutionary information in the form of PSSM profile, EntWOP, and Dist_Key (describing sequence-specific distance between two adjacent catalytic residues). Among these, EntWOP was the top-ranked feature, with the highest MDGI Z-score of 24.3 (**Figure 2**). This can be considered a condensed type of sequence-conservation feature derived from the multiple sequence alignments of homologous proteins gathered from the nonredundant protein databases. Zhang et al. originally introduced the concept of this feature and used it in their SVM-based models to improve the prediction of catalytic residues (Zhang, et al., 2008). Here we confirmed EntWOP as a powerful feature, with its use in conjunction with PSSM features greatly benefiting catalytic residue prediction.

This study also confirmed the critical importance of physicochemical properties of amino acids, including hydrophobicity (Zhang, et al., 2008) and charge (Sankararaman, et al., 2010), as well as structural features, including solvent accessibility (Gutteridge, et al., 2003; Petrova and Wu, 2006) and B-factor (Sankararaman, et al., 2010; Youn, et al., 2007). The B-factor had a relatively smaller contribution according to its lower MDGI Z-score (1 < Z-score < 1.5) (**Supplementary Table 1**). Additionally, we identified several structural features as important for predicting catalytic residues, including CN, HSEAD, and HSEBD (Wang, et al., 2012), with these ranked as top features (MDGI Z-scores > 2.0). To our knowledge, this was the first application of these features to build models for identifying catalytic residues and thus represent novel informative features for this prediction task.

In summary, investigating the impact of integrating these sequence-derived, structural, and network level features might provide complementary information to existing methods and shed light on the sequence-structure-function relationships of functional residues. In the following sections, we examine the effectiveness of combining these features to train our PREvaIL models and compared them with other existing methods.

3.3. Performance comparison between PREvaIL and other methods

In this section, we compared the performance of our PREvaIL method with two sequence-based methods and five structure-based method by performing 10-fold cross-validation tests on six datasets (EF_fold, EF_superfamily, EF_family, HA_family, NN, and PC). Additionally, we also compared the performance of our method against four methods by performing independent tests on the T-124 dataset. To facilitate performance comparison, we used the TP rate and PRE by adjusting the prediction cut-off value to achieve an equal or close-to-equal PRE with different methods. For 10-fold cross-validation tests, performance results between different methods are shown in **Table 2**. In terms of independent tests, the performance results are shown in **Table 3**. The ROC curves and Precision-Recall curves of the CRpred method and our method are shown in **Figures 5** and **S2**.

The five structure-based methods in the benchmark included a neural-network method (two versions with and without spatial clustering) (Gutteridge, et al., 2003), two SVM methods (Petrova and Wu, 2006; Youn, et al., 2007), and a graph-theoretic method (Chea and Livesay, 2007), whereas the two sequence-based methods included a neural-network method (Gutteridge, et al., 2003) and CRpred (Zhang, et al., 2008).

The SVM-based method proposed by Youn et al. (2007) extracted a number of features from sequence, sequence alignments, 3D structures, and structural-environment conservation and used the SVM algorithm to perform automated catalytic site prediction and annotation (Youn, et al., 2007). In their study, the authors found that structural features of residue environments, such as solvent accessibility, together with sequence conservation were particularly important for predicting catalytic residues. Specifically, this method achieved TP rates of 51.1%, 53.9%, and 57.0%, and PRE values of 17.1%, 16.9%, and 18.5% on the EF_fold, EF_superfamily, and EF_family datasets, respectively.

The method proposed by Chea and Livesay (2007) is a graph-theoretic method that uses closeness centrality as the primary feature based on a network representation of protein structure to predict enzyme catalytic residues (Chea and Livesay, 2007). This method achieved a TP rate of 29.3% and Precision of 16.5%, respectively, on the HA_superfamily dataset.

The method proposed by Petrova and Wu is an SVM-based method that selected seven of the 24 attributes as an optimal subset of features, including sequence conservation, catalytic propensities of amino acids, and relative position on the protein surface as the most important features (Petrova and Wu, 2006). This method achieved a TP rate of 90.0%, but a significantly lower PRE value of only 7.0%, on the PC dataset.

The method proposed by Gutteridge *et al.* (2003) is a neural network based on analysis of both sequence and structural features and using solvent accessibility, secondary structure type, residue depth, and the pocket in which the catalytic residues are located, as well as conservation score and residue type, as inputs for training the neural-network models. After predicting the catalytic residues using these models, the output and spatial clustering of the high scoring residues were then used to predict the location of the active site (Gutteridge, et al., 2003). For the structure-based version with spatial clustering, the method achieved a better performance in terms of TP rate (68.0%) and PRE value (16.0%) on the NN dataset; however, its sequence-based version performed worse than the structure-based version, with a TP rate of 50.0% and a PRE value of 13.0% on the NN dataset.

CRpred is also an SVM-based method that takes advantage of a wide range of sequence features, including residue type, PSSM profile generated by PSI-BLAST, EntWOP, hydrophobicity, and catalytic residue pairs (544 features) (Zhang, et al., 2008). To reduce the dimensionality of the input features, CRpred uses feature selection to eliminate redundant and less relevant features (Zhang, et al., 2008). Due to its competitive performance and the optimized parameterization of the SVM models and carefully designed feature sets, the CRpred method is considered as a state-of-the-art method for enzyme catalytic residue prediction. CRpred was extensively tested by 10-fold cross-validation on all six datasets, achieving TP rates of 54.0% and 53.7% and PRE values of 14.9% and 17.5% on the HA_superfamily and PC datasets, respectively (**Table 2**).

As a comparison, the performance of the RF models of our PREvaIL method achieved TP rates of 56.5%, 59.4%, and 60.2% at the fixed PRE value of 17.0% on the EF_fold, EF_superfamily, and EF_family datasets, respectively. This was a consistently better performance relative to all other methods, including CRpred. On the HA_superfamily dataset, PREvaIL achieved a TP rate of 57.9% at the fixed PRE value of 17.0%, representing the best performance on this dataset and also achieved the best performance on the NN and PC datasets (**Table 3**). The improved performance of PREvaIL over CRpred was more pronounced when they were evaluated in terms of the ROC curve and RPC, especially the latter (**Figures 5** and **S2**). Depending on the particular test dataset, PREvaIL consistently achieved a larger AURPC value than CRpred. These results indicated that the PREvaIL method provided the overall best performance as compared with several sequence- and structure-based methods.

3.4 Performance evaluation by independent testing using the T-124 and T-37 datasets

We performed additional independent testing using the T-124 and T-37 datasets, which exhibit the lowest sequence identities (<30%) with the two training datasets (ST-1109 and EF_fold). As previously suggested (Zhang, et al., 2008), we trained the PREvaIL model using the entire EF_fold dataset, tested the trained model, and compared its performance with other methods using the independent test datasets. The performance results between PREvaIL, CRpred, and the structure-based HA method proposed by Chea and Livesay (2007) on the T-124 dataset are shown in **Table 3**. Additionally, the performance results between PREvaIL and CRpred on the two independent test datasets are also shown in **Figure S2** in terms of ROC curve and RPC.

As compared with CRpred (based on all residues and residues with coordinates) and the HA method (based on residue-identity and combination filters), PREvaIL outperformed these four different methods when evaluated on both the T-124 and T-37 datasets. More specifically, PREvaIL achieved a TP rate of 62.2% and a PRE value of 14.9% on the T-124 dataset, whereas CRpred (based on residues with coordinates) achieved a TP rate of 50.1% at 14.7% REC. In contrast, the HA method (based on residue-identity filter) achieved a TP rate of 27.71% at 16.1% REC. However, we noticed that all tested methods performed relatively poorly on these two independent test datasets, as reflected by the lower AURPC values (<0.35) (**Supplementary Figure S2D and E**). As observed in previous studies (Fischer, et al., 2008; Kauffman and Karypis, 2009), none of these methods achieved >30% PRE at 50% REC on the two independent test datasets. For example, the best performing PREvaIL method only achieved 20% PRE at 50% REC on the T-124 dataset. These results indicated that there remains a strong need to improve the performance of the predictors, especially at the higher REC. Future studies should investigate incorporation of other relevant features that might prove useful for improving the predictive performance of catalytic residues. In this regards, a recent work that proposed new network-based features that describe side chain orientation and residue contact density (Chien and Huang, 2012) might provide additional information for further improving the model performance.

In summary, the extensive benchmarking results on the eight datasets indicated that the combination of the different types of features descriptors extracted from sequence, structural, and residue-contact networks provided a more representative power that could be leveraged by the RF algorithm to achieve better performance for accurately differentiating enzyme catalytic residues.

3.5 Case study

To demonstrate the effectiveness and predictive capability of PREvaIL, we further performed a case study of catalytic residue prediction by selecting two different proteins. Specifically, we applied two selection criteria for choosing case study proteins. A primary consideration is that they should contain multiple catalytic residues (three or more catalytic residues), such that we can evaluate and compare the predictive performance of our method for predicting these catalytic residues within the same enzyme. Another consideration is that the case study proteins should have important biological functions, as indicated by their functional roles or involvement in significant biological pathways and processes. **Figure 6** shows their predicted catalytic residues. The first case study protein, anabolic ornithine transcarbamylase from *Escherichia coli* (PDB: 1AKM, chain A), is an essential metabolic enzyme that catalyzes the production of L-citrulline and phosphate from L-ornithine and carbamyl phosphate (Jin, et al., 1997) and has seven catalytic residues. The RF models of PREvaIL correctly predicted six of these, including R57, R106, H133, Q136, C273, and R319 (**Figure 6A and B, red**) while incorrectly classifying the seventh catalytic residue (T58)

(**Figure 6A and B, yellow**). The second case study protein, mitochondrial creatine kinase (PDB: 1CRK, chain A) (Fritz-Wolf, et al., 1996), is an enzyme that catalyzes the reversible transfer of a phosphoryl group from phosphocreatine to adenosine diphosphate and is considered important for energy metabolism in cells with high and fluctuating energy requirements (Fritz-Wolf, et al., 1996). It has five catalytic residues annotated in the CSA database. The RF model successfully predicted three of these, including R127, E227, and R231 (**Figure 6C and D, red**) while failing to identify the fourth and fifth catalytic residues: R287 and R315 (**Figure 6C and D, yellow**). It is challenging to explain why certain catalytic residues were incorrectly predicted as noncatalytic by the model. However, we found that catalytic residues tended to have smaller EntWOP and Dist_Key values, and larger CN, HSEAD, HSEBD values compared with noncatalytic residues. This was consistent with the observations from **Figures 4 and S1**. From this viewpoint, those catalytic residues with relatively larger EntWOP and Dist_Key values, or smaller CN, HSEAD, HSEBD values represent difficult samples to predict, which is the case for the three incorrectly catalytic residues. A detailed list of all the predicted catalytic residues is given in **Supplementary Table 3**. These results suggested that PREvaIL is a useful tool for novel catalytic residue prediction.

4. CONCLULSIONS

In this study, we demonstrated that the combinatorial application of machine learning techniques on multi-level protein features involving sequence-derived, structural, and residue-contact-network features allowed the development of a powerful bioinformatics predictor, PREvaIL. Previous methods explored these different levels of features separately; however, we illustrated their effective integration into a machine-learning framework to provide complementary information to collectively help improve predictive performance associated with catalytic residue prediction. Through in-depth feature analysis, we identified a smaller subset of features arising from all of these levels that significantly contributed to the prediction. Using 10-fold cross-validation and independent test datasets, we showed that the performance of PREvaIL compared favorably with two sequence-based methods and five structure-based methods. The improved performance of PREvaIL might be attributed to three major factors: 1) inclusion of a comprehensive set of informative features at the sequence, structure, and residue-contact-network levels; 2) selection of an optimal set of contributing features, and 3) use of the RF algorithm for machine learning-based model training, which provided a robust and competitive performance. A local stand-alone version of PREvaIL can be downloaded at http://prevail.erc.monash.edu/. We believe that this approach could be utilized as a valuable tool for both understanding the complex sequence-structure-function relationships of proteins and facilitating the characterization of novel enzymes with unknown functional annotations.

Acknowledgments

We would like to thank Drs. Tuo Zhang and Lukasz Kurgan who graciously made publicly available the datasets and extracted features used in the benchmarking testing of CRpred.

References:

Alterovitz, R., *et al.* ResBoost: characterizing and predicting catalytic residues in enzymes. *BMC Bioinformatics* 2009;10:197.

Amanzadeh, E., Mohabatkar, H. and Biria, D. Classification of DNA minor and major grooves binding proteins according to the NLSS by data analysis methods. *Applied biochemistry and biotechnology* 2014;174(1):437-451.

Amitai, G., *et al.* Network analysis of protein structures identifies functional residues. *J Mol Biol* 2004;344(4):1135-1146.

Behbahani, M., Mohabatkar, H. and Nosrati, M. Analysis and comparison of lignin peroxidases between fungi and bacteria using three different modes of Chou's general pseudo amino acid composition. *Journal of theoretical biology* 2016;411:1-5.

Beigi, M.M., Behjati, M. and Mohabatkar, H. Prediction of metalloproteinase family based on the concept of Chou's pseudo amino acid composition using a machine learning approach. *Journal of Structural and Functional Genomics* 2011;12(4):191-197.

Breiman, L. Random forests. *Machine learning* 2001;45(1):5-32.

Cao, D.S., Xu, Q.S. and Liang, Y.Z. propy: a tool to generate various modes of Chou's PseAAC. *Bioinformatics* 2013;29(7):960-962.

Capra, J.A. and Singh, M. Predicting functionally important residues from sequence conservation. *Bioinformatics* 2007;23(15):1875-1882.

Carter, P. and Wells, J.A. Dissecting the catalytic triad of a serine protease. *Nature* 1988;332(6164):564-568.

Chea, E. and Livesay, D.R. How accurate and statistically robust are catalytic site predictions based on closeness centrality? *BMC Bioinformatics* 2007;8:153.

Chen, Z., *et al.* ZincExplorer: an accurate hybrid method to improve the prediction of zinc-binding sites from protein sequences. *Molecular BioSystems* 2013;9(9):2213-2222.

Chien, Y.-T. and Huang, S.-W. Accurate prediction of protein catalytic residues by side chain orientation and residue contact density. *PLoS One* 2012;7(10):e47951.

Chou, K.C. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins-Structure Function and Genetics* 2001;43(3):246-255.

Chou, K.C. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics* 2005;21(1):10-19.

Chou, K.C. Some remarks on protein attribute prediction and pseudo amino acid composition. *Journal of Theoretical Biology* 2011;273(1):236-247.

Chou, K.C. Impacts of Bioinformatics to Medicinal Chemistry. *Med Chem* 2015;11(3):218-234.

Chou, K.C. An Unprecedented Revolution in Medicinal Chemistry Driven by the Progress of Biological Science. *Curr Top Med Chem* 2017;17(21):2337-2358.

Chou, K.C. and Cai, Y.D. A novel approach to predict active sites of enzyme molecules. *Proteins-Structure Function and Bioinformatics* 2004;55(1):77-82.

Chou, K.C. and Zhou, G.P. Role of the Protein Outside Active-Site on the Diffusion-Controlled Reaction of Enzyme. *J Am Chem Soc* 1982;104(5):1409-1413.

Cilia, E. and Passerini, A. Automatic prediction of catalytic residues by modeling residue structural neighborhood. *BMC Bioinformatics* 2010;11:115.

Cock, P.J., *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 2009;25(11):1422-1423.

Csardi, G. and Nepusz, T. The igraph software package for complex network research. *InterJournal, Complex Systems* 2006;1695(5):1-9.

del Sol, A., *et al.* Residue centrality, functionally important residues, and active site shape: analysis of enzyme and non-enzyme families. *Protein Sci* 2006;15(9):2120-2128.

del Sol, A. and O'Meara, P. Small-world network approach to identify key residues in protein-protein interaction. *Proteins* 2005;58(3):672-682.

Disfani, F.M., *et al.* MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins. *Bioinformatics* 2012;28(12):i75-83.

Dou, Y., *et al.* L1pred: a sequence-based prediction tool for catalytic residues in enzymes with the L1-logreg classifier. *PLoS One* 2012;7(4):e35666.

Dou, Y., *et al.* Prediction of catalytic residues based on an overlapping amino acid classification. *Amino Acids* 2010;39(5):1353-1361.

Du, P.F., Gu, S.W. and Jiao, Y.S. PseAAC-General: Fast Building Various Modes of General Form of Chou's Pseudo-Amino Acid Composition for Large-Scale Protein Datasets. *Int J Mol Sci* 2014;15(3):3495-3506.

Du, P.F., *et al.* PseAAC-Builder: A cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions. *Anal Biochem* 2012;425(2):117-119.

Esmaeili, M., Mohabatkar, H. and Mohsenzadeh, S. Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses. *Journal of theoretical biology* 2010;263(2):203-209.

Fischer, J.D., Mayer, C.E. and Soding, J. Prediction of protein functional residues from sequence by probability density estimation. *Bioinformatics* 2008;24(5):613-620.

Fritz-Wolf, K., *et al.* Structure of mitochondrial creatine kinase. *Nature* 1996;381(6580):341-345.

Fu, L., *et al.* CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 2012;28(23):3150-3152.

Furnham, N., *et al.* The Catalytic Site Atlas 2.0: cataloging catalytic sites and residues identified in enzymes. *Nucleic Acids Res* 2014;42(Database issue):D485-489.

Gardner, P.R., Gardner, D.P. and Gardner, A.P. Globins Scavenge Sulfur Trioxide Anion Radical. *J Biol Chem* 2015;290(45):27204-27214.

Gutteridge, A., Bartlett, G.J. and Thornton, J.M. Using a neural network and spatial clustering to predict the location of active sites in enzymes. *J Mol Biol* 2003;330(4):719-734.

Hajisharifi, Z., *et al.* Predicting anticancer peptides with Chou′ s pseudo amino acid composition and investigating their mutagenicity via Ames test. *Journal of Theoretical Biology* 2014;341:34-40.

Hamelryck, T. An amino acid has two sides: a new 2D measure provides a different view of solvent exposure. *Proteins* 2005;59(1):38-48.

Han, L., *et al.* Identification of catalytic residues using a novel feature that integrates the microenvironment and geometrical location properties of residues. *PLoS One* 2012;7(7):e41370.

Hubbard, S.J. and Thornton, J.M. Naccess. *Computer Program, Department of Biochemistry and Molecular Biology, University College London* 1993;2(1).

Izidoro, S.C., de Melo-Minardi, R.C. and Pappa, G.L. GASS: identifying enzyme active sites with genetic algorithms. *Bioinformatics* 2015;31(6):864-870.

Jia, J.H., *et al.* iPPI-Esml: An ensemble classifier for identifying the interactions of proteins by incorporating their physicochemical properties and wavelet transforms into PseAAC. *Journal of Theoretical Biology* 2015;377:47-56.

Jia, J.H., *et al.* pSuc-Lys: Predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach. *Journal of Theoretical Biology* 2016;394:223-230.

Jiao, X. and Ranganathan, S. Prediction of interface residue based on the features of residue interaction network. *J Theor Biol* 2017;432:49-54.

Jin, L., Seaton, B.A. and Head, J.F. Crystal structure at 2.8 A resolution of anabolic ornithine transcarbamylase from Escherichia coli. *Nat Struct Biol* 1997;4(8):622-625.

Jones, D.T. Protein secondary structure prediction based on position-specific scoring matrices. *Journal of molecular biology* 1999;292(2):195-202.

Jones, D.T. and Cozzetto, D. DISOPRED3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics* 2015;31(6):857-863.

Kabsch, W. and Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 1983;22(12):2577-2637.

Kauffman, C. and Karypis, G. LIBRUS: combined machine learning and homology information for sequence-based ligand-binding residue prediction. *Bioinformatics* 2009;25(23):3099-3107.

Khan, M*., et al.* Unb-DPC: Identify mycobacterial membrane protein types by incorporating un-biased dipeptide composition into Chou's general PseAAC. *Journal of Theoretical Biology* 2017;415:13-19.

Khosla, C. and Harbury, P.B. Modular enzymes. *Nature* 2001;409(6817):247-252.

Khosravian, M*., et al.* Predicting antibacterial peptides by the concept of Chou's pseudo-amino acid composition and machine learning methods. *Protein and Peptide Letters* 2013;20(2):180-186.

Kirshner, D.A., Nilmeier, J.P. and Lightstone, F.C. Catalytic site identification--a web server to identify catalytic site structural matches throughout PDB. *Nucleic Acids Res* 2013;41(Web Server issue):W256-265.

Kuo-chen, C. and Shou-ping, J. Studies on the rate of diffusion-controlled reactions of enzymes. Spatial factor and force field factor. *Sci Sin* 1974;27(5):664-680.

La, D., Sutch, B. and Livesay, D.R. Predicting protein functional sites with phylogenetic motifs. *Proteins* 2005;58(2):309-320.

Li, F*., et al.* GlycoMine: a machine learning-based approach for predicting N-, C- and O-linked glycosylation in the human proteome. *Bioinformatics* 2015;31(9):1411-1419.

Li, Y*., et al.* Novel feature for catalytic protein residues reflecting interactions with other residues. *PLoS One* 2011;6(3):e16932.

Li, Y*., et al.* Accurate in silico identification of species-specific acetylation sites by integrating protein sequence-derived and functional features. *Sci Rep* 2014;4:5765.

Liaw, A. and Wiener, M. Classification and regression by randomForest. *R news* 2002;2(3):18-22.

Liu, B., Long, R. and Chou, K.C. iDHS-EL: identifying DNase I hypersensitive sites by fusing three different modes of pseudo nucleotide composition into an ensemble learning framework. *Bioinformatics* 2016;32(16):2411-2418.

Lobley, A*., et al.* Inferring function using patterns of native disorder in proteins. *PLoS computational biology* 2007;3(8):e162.

Meher, P.K*., et al.* Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into Chou's general PseAAC. *Sci Rep-Uk* 2017;7.

Meng, F. and Kurgan, L. DFLpred: High-throughput prediction of disordered flexible linker regions in protein sequences. *Bioinformatics* 2016;32(12):i341-i350.

Mohabatkar, H. Prediction of cyclin proteins using Chou's pseudo amino acid composition. *Protein and peptide letters* 2010;17(10):1207-1214.

Mohabatkar, H., Beigi, M.M. and Esmaeili, A. Prediction of GABA A receptor proteins using the concept of Chou's pseudo-amino acid composition and support vector machine. *Journal of Theoretical Biology* 2011;281(1):18-23.

Mohabatkar, H., *et al.* Prediction of allergenic proteins by means of the concept of Chou's pseudo amino acid composition and a machine learning approach. *Med Chem* 2013;9(1):133-137.

Mousavizadegan, M. and Mohabatkar, H. An evaluation on different machine learning algorithms for classification and prediction of antifungal peptides. *Med Chem* 2016;12(8):795-800.

Ofran, Y. and Rost, B. Protein–protein interaction hotspots carved into sequences. *PLoS computational biology* 2007;3(7):e119.

Pai, P.P., Ranjani, S.S. and Mondal, S. PINGU: PredIction of eNzyme catalytic residues usinG seqUence information. *PLoS One* 2015;10(8):e0135122.

Panchenko, A.R., Kondrashov, F. and Bryant, S. Prediction of functional sites by analysis of sequence and structure conservation. *Protein Sci* 2004;13(4):884-892.

Petrova, N.V. and Wu, C.H. Prediction of catalytic residues using Support Vector Machine with selected protein sequence and structural properties. *BMC Bioinformatics* 2006;7:312.

Poorinmohammad, N., *et al.* Computational prediction of anti HIV-1 peptides and in vitro evaluation of anti HIV-1 activity of HIV-1 P24-derived peptides. *Journal of Peptide Science* 2015;21(1):10-16.

Porter, C.T., Bartlett, G.J. and Thornton, J.M. The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res* 2004;32(Database issue):D129-133.

Qiu, W.R., *et al.* iPhos-PseEn: Identifying phosphorylation sites in proteins by fusing different pseudo components into an ensemble classifier. *Oncotarget* 2016;7(32):51270-51283.

Rahimi, M., Bakhtiarizadeh, M.R. and Mohammadi-Sangcheshmeh, A. OOgenesis_Pred: A sequence-based method for predicting oogenesis proteins by six different modes of Chou's pseudo amino acid composition. *Journal of Theoretical Biology* 2017;414:128-136.

Rose, P.W., *et al.* The RCSB protein data bank: integrative view of protein, gene and 3D structural information. *Nucleic Acids Res* 2017;45(D1):D271-D281.

Sankararaman, S., *et al.* Active site prediction using evolutionary and structural information. *Bioinformatics* 2010;26(5):617-624.

Shen, H.-B., Song, J.-N. and Chou, K.-C. Prediction of protein folding rates from primary sequence by fusing multiple sequential features. *Journal of Biomedical Science and Engineering (JBiSE)* 2009;2:136-143.

Smith, D.K., *et al.* Improved amino acid flexibility parameters. *Protein Science* 2003;12(5):1060-1072.

Song, J., *et al.* Prediction of cis/trans isomerization in proteins using PSI-BLAST profiles and secondary structure information. *BMC bioinformatics* 2006;7(1):124.

Song, J., *et al.* MetalExplorer, a Bioinformatics Tool for the Improved Prediction of Eight Types of Metal-Binding Sites Using a Random Forest Algorithm with Two-Step Feature Selection. *Current Bioinformatics* 2017;12(6):480-489.

Song, J., *et al.* Prodepth: predict residue depth by support vector regression approach from protein sequences only. *PLoS One* 2009;4(9):e7072.

Song, J., *et al.* PROSPER: an integrated feature-based tool for predicting protease substrate cleavage sites. *PloS one* 2012;7(11):e50300.

Song, J., *et al.* Cascleave: towards more accurate prediction of caspase substrate cleavage sites. *Bioinformatics* 2010;26(6):752-760.

Song, J., *et al.* HSEpred: predict half-sphere exposure from protein sequences. *Bioinformatics*

2008;24(13):1489-1497.

Song, J., *et al.* PhosphoPredict: A bioinformatics tool for prediction of human kinase-specific phosphorylation substrates and sites by integrating heterogeneous feature selection. *Sci Rep* 2017;7(1):6862.

Song, J., *et al.* Predicting disulfide connectivity from protein sequence using multiple sequence feature vectors and secondary structure. *Bioinformatics* 2007;23(23):3147-3154.

Sun, J., *et al.* CRHunter: integrating multifaceted information to predict catalytic residues in enzymes. *Sci Rep* 2016;6:34044.

Tahir, M., Hayat, M. and Kabir, M. Sequence based predictor for discrimination of enhancer and their types by applying general form of Chou's trinucleotide composition. *Comput Meth Prog Bio* 2017;146:69-75.

Tang, Y.R., *et al.* An improved prediction of catalytic residues in enzyme structures. *Protein Eng Des Sel* 2008;21(5):295-302.

Wang, M., *et al.* FunSAV: predicting the functional effect of single amino acid variants using a two-stage random forest model. *PLoS One* 2012;7(8):e43847.

Wang, M., *et al.* Cascleave 2.0, a new approach for predicting caspase and granzyme cleavage targets. *Bioinformatics* 2014;30(1):71-80.

Watts, D.J. and Strogatz, S.H. Collective dynamics of 'small-world' networks. *Nature* 1998;393(6684):440-442.

Xin, F., *et al.* Structure-based kernels for the prediction of catalytic residues and their involvement in human inherited disease. *Bioinformatics* 2010;26(16):1975-1982.

Youn, E., *et al.* Evaluation of features for catalytic residue prediction in novel folds. *Protein Sci* 2007;16(2):216-226.

Yuan, Z., Bailey, T.L. and Teasdale, R.D. Prediction of protein B-factor profiles. *Proteins* 2005;58(4):905-912.

Zhang, T., *et al.* Accurate sequence-based prediction of catalytic residues. *Bioinformatics* 2008;24(20):2329-2338.

Zheng, C., *et al.* An integrative computational framework based on a two-step random forest algorithm improves prediction of zinc-binding sites in proteins. *PLoS One* 2012;7(11):e49716.

Zhong, W.Z. and Zhou, S.F. Molecular Science for Drug Development and Biomedicine. *Int J Mol Sci* 2014;15(11):20072-20078.

Zhou, G.Q. and Zhong, W.Z. Diffusion-controlled reactions of enzymes. A comparison between Chou's model and Alberty-Hammes-Eigen's model. *Eur J Biochem* 1982;128(2-3):383-387.

Zhou, J., *et al.* Amino acid network for prediction of catalytic residues in enzymes: a comparison survey. *Current Protein and Peptide Science* 2016;17(1):41-51.

Zhou, X.B., *et al.* Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes. *Journal of Theoretical Biology* 2007;248(3):546-551.

Figure legends

Figure 1. Flowchart describing the PREvaIL methodology for predicting catalytic residues based on the integration of sequence, structural, and residue-contact-network features using the RF learning framework.

Figure 2. The MDGI Z-Scores for the selected feature groups. The bar represents the corresponding MDGI Z-Score of a feature group. The feature group is the same as the feature category described in **Supplementary Table 1**. There were 41 features with MDGI Z-scores >2.0.

Figure 3. Two-dimensional scaling plot of the representative sequence, structure, and network-based features in the feature space. Feature descriptors that are closely correlated are closely clustered together in the feature space. The scale units of the plot are relative to the smallest correlation between feature pairs as measured by Pearson's correlation coefficient (Lobley *et al.*, 2007).

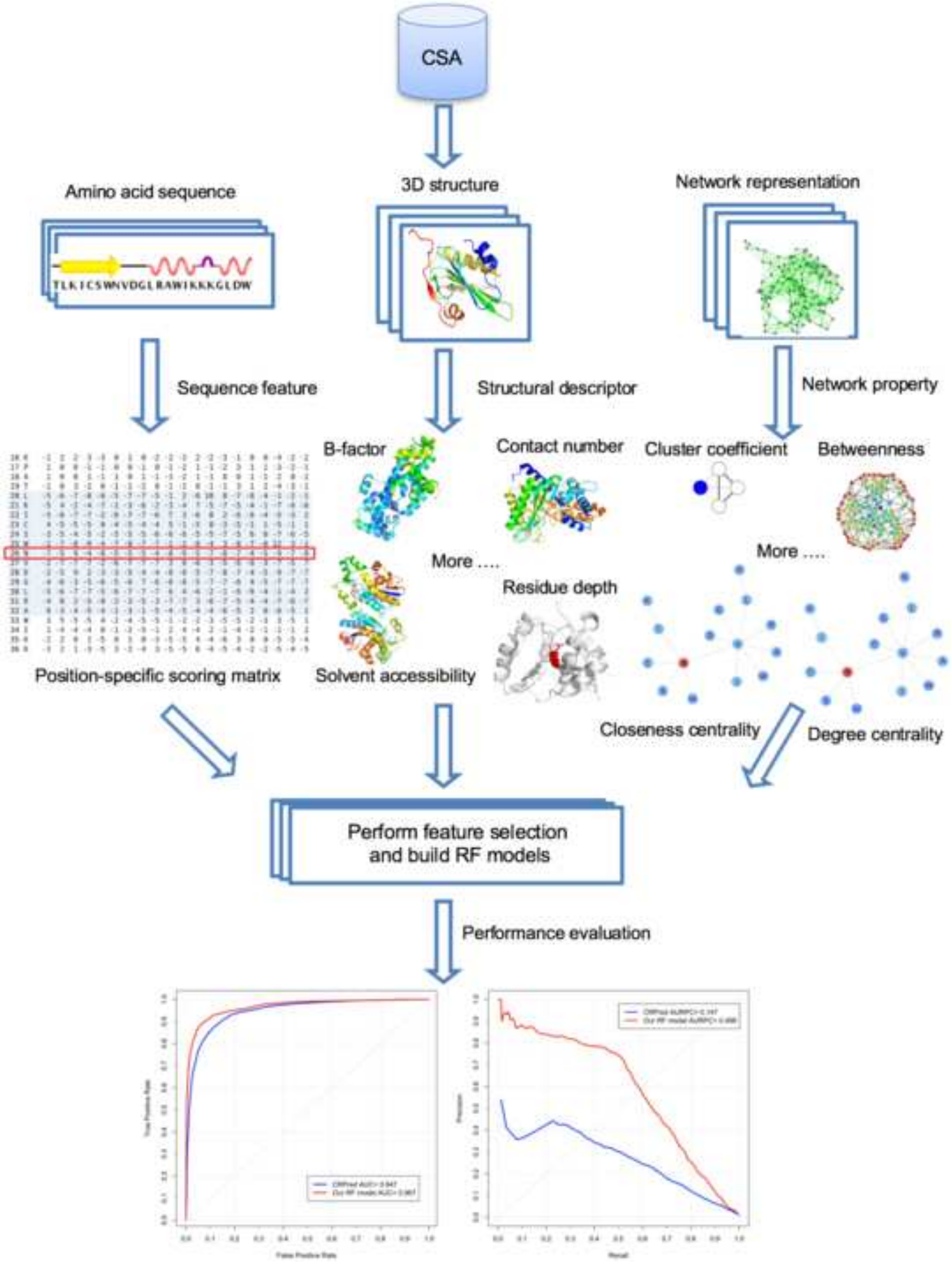
Figure 4. Boxplots of the mean and standard deviations of the four representative structural and residue-contact-network features based on the unpaired two-sample *t* test.

Figure 5. ROC and RPC for the RF-based PREvaIL method and the SVM-based CRpred method. AUC and AURPC values were provided to quantify the performance.

Figure 6. Examples of the predicted catalytic residues mapped onto the original PDB structures. (A and B) Anabolic ornithine transcarbamylase from *Escherichia coli* (PDB: 1AKM, chain A) (Jin *et al.*, 1997). (C and D) Mitochondrial creatine kinase (PDB: 1CRK, chain A) (Fritz-Wolf *et al.*, 1996). Different prediction catalogs are represented by different colors: TP, red; TN, grey; FP, blue; FN, yellow.

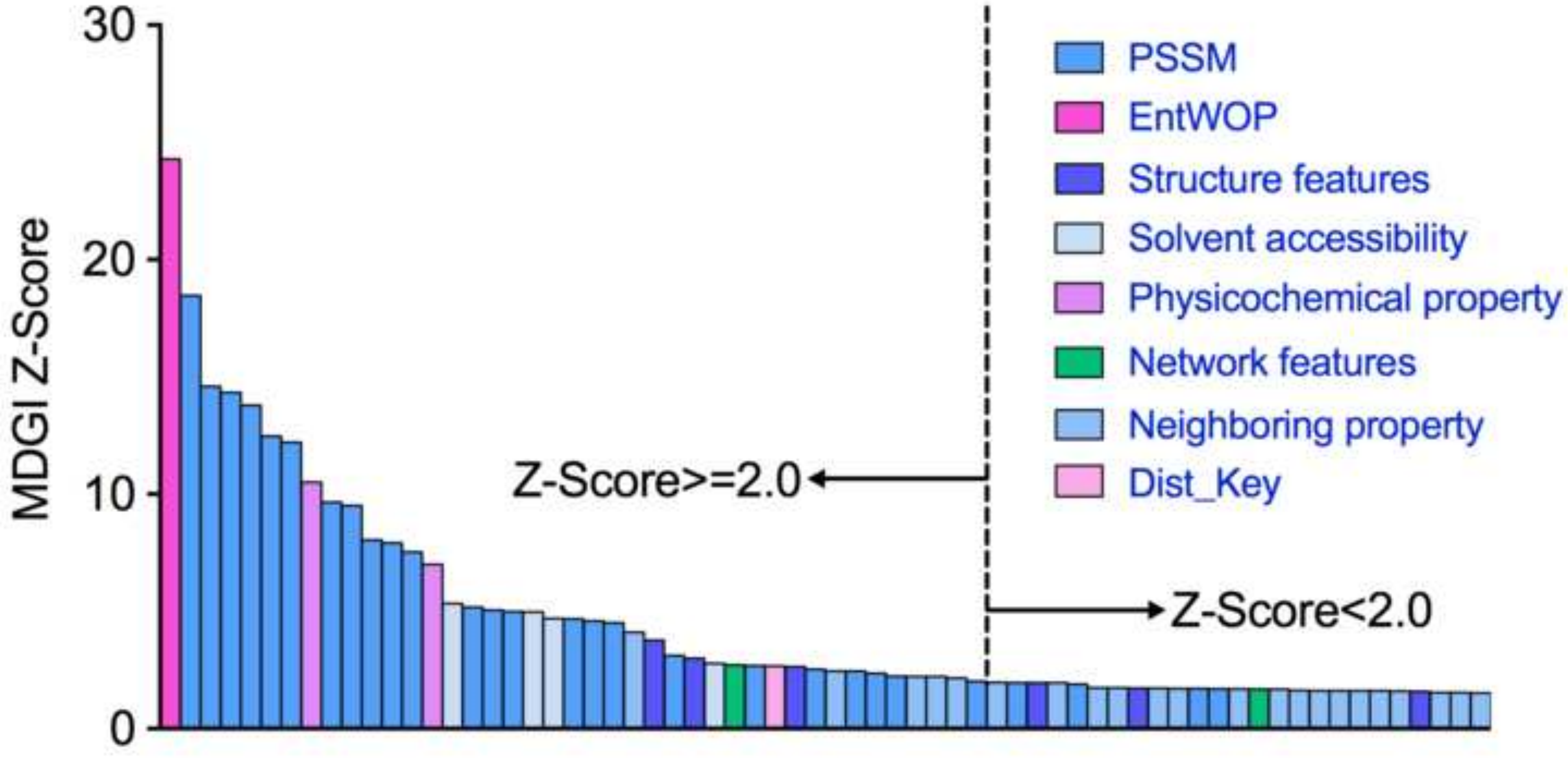
4. Figure 1

[redacted]

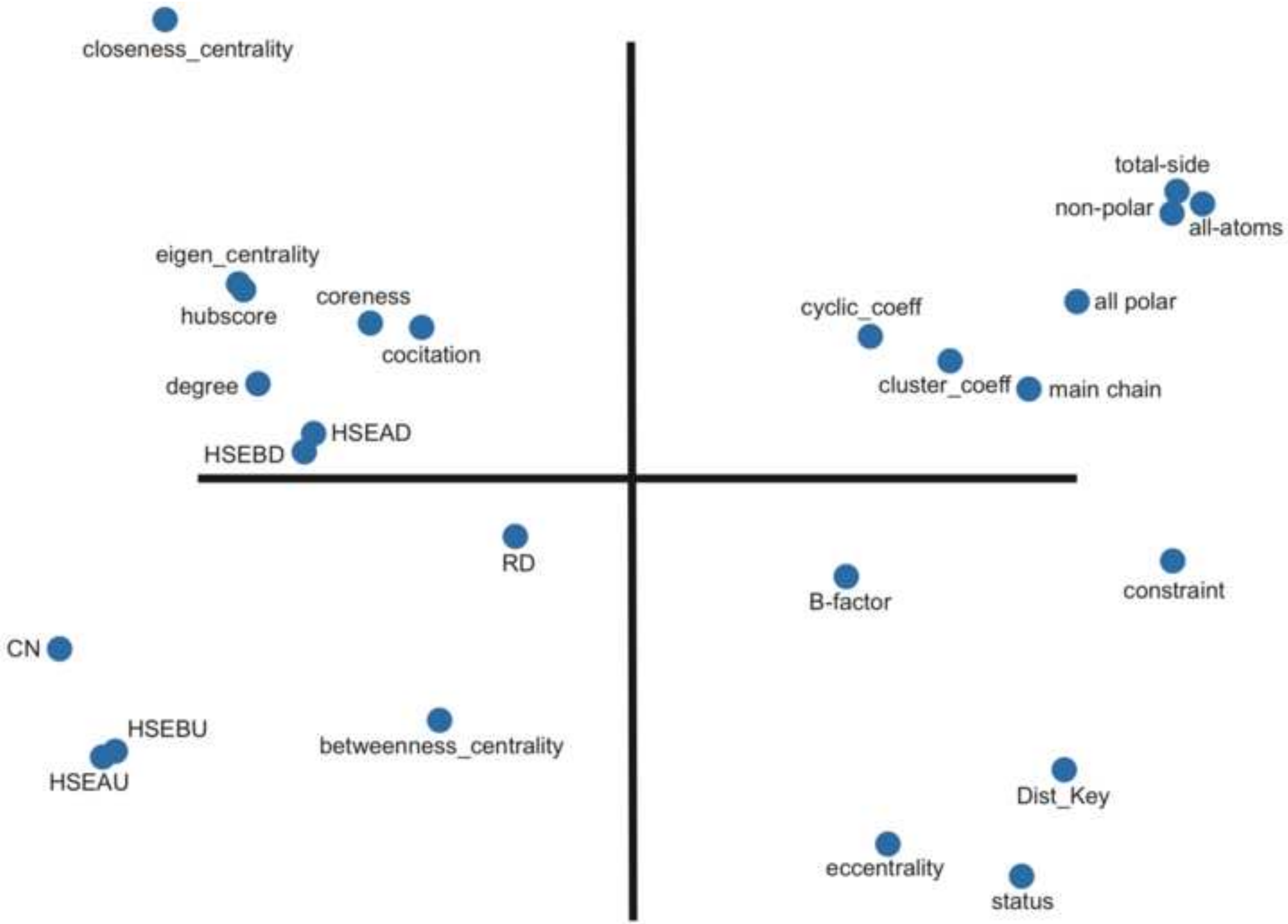


4. Figure 2

[redacted]

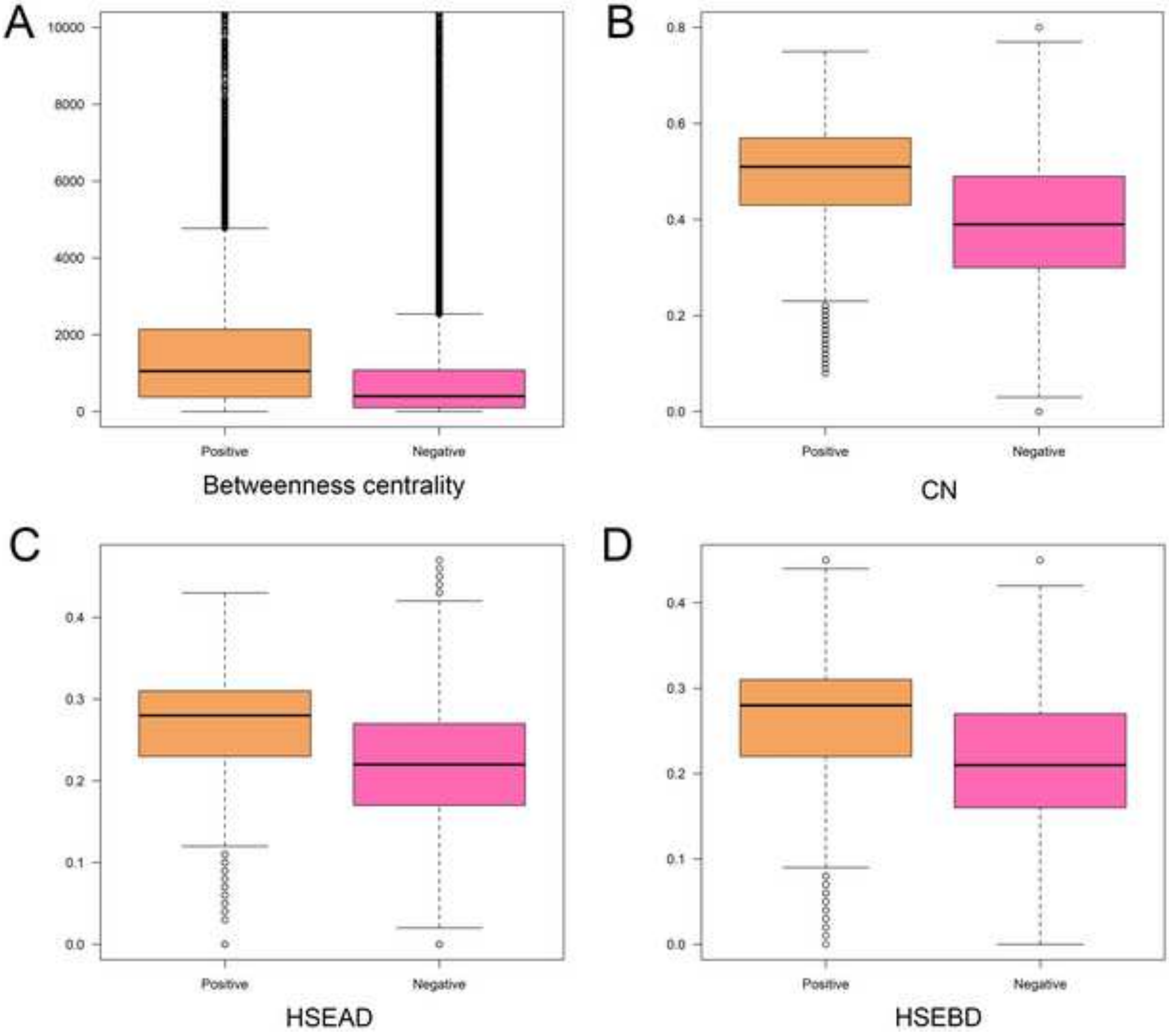


4. Figure 3


closeness_centrality
total-side
non-polar
all-atoms
eigen_centrality
hubscore
coreness
cocitation
cyclic_coeff
all polar
cluster_coeff
main chain
degree
HSEAD
HSEBD
RD
B-factor
constraint
CN
HSEBU
HSEAU
betweenness_centrality
Dist_Key
eccentrality
status

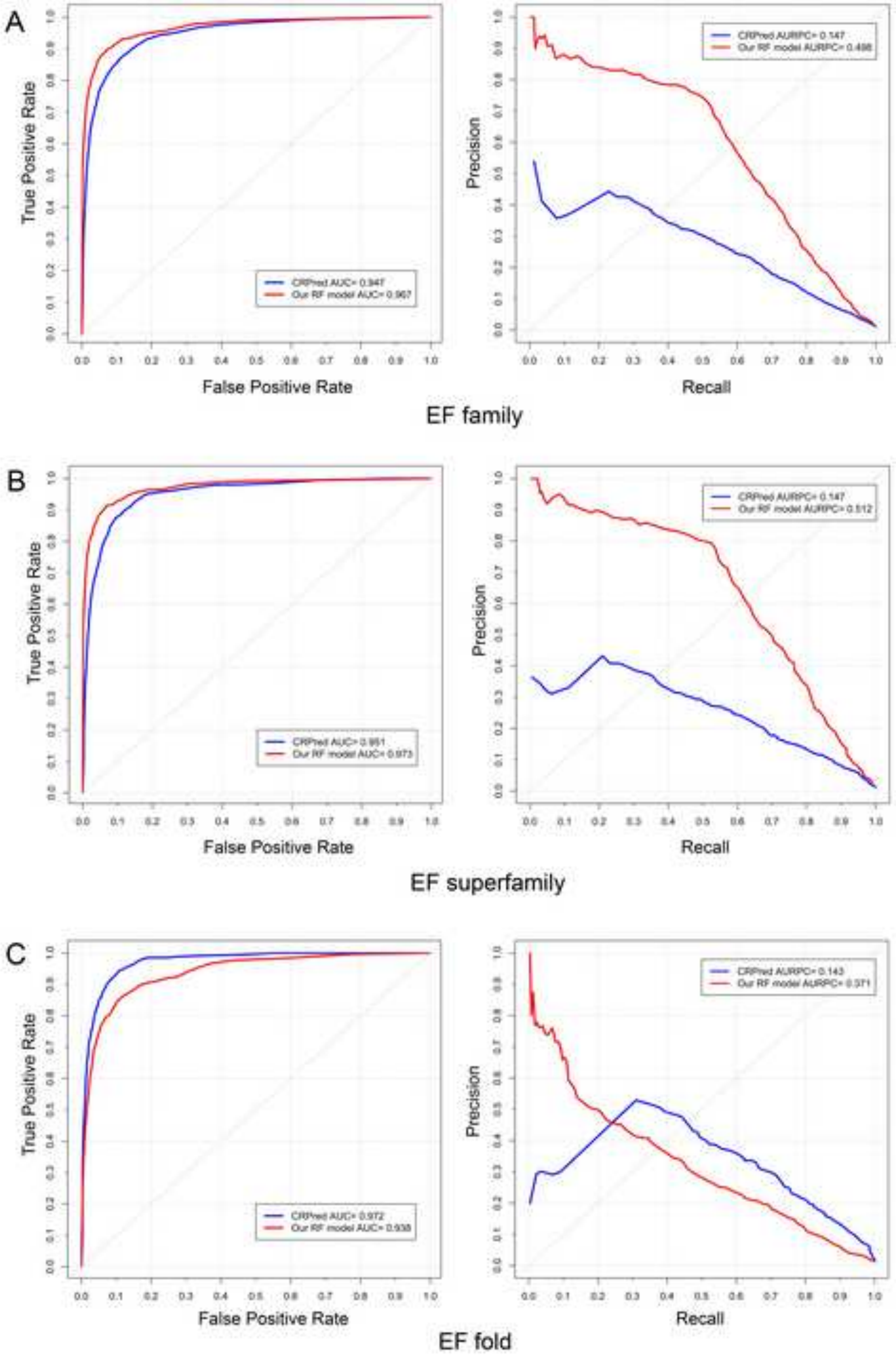
4. Figure 4

[redacted]



4. Figure 5

[redacted]

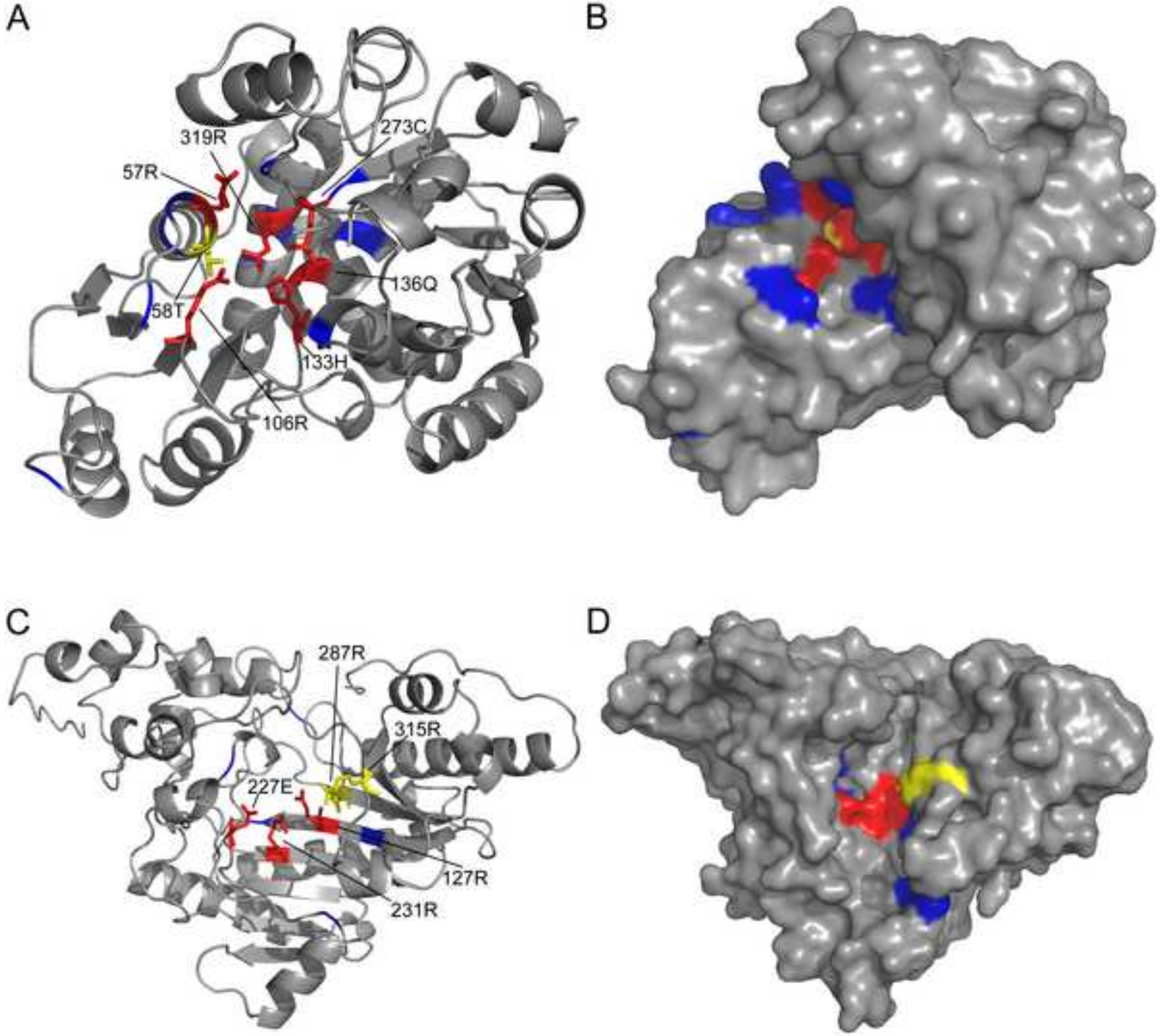


EF family

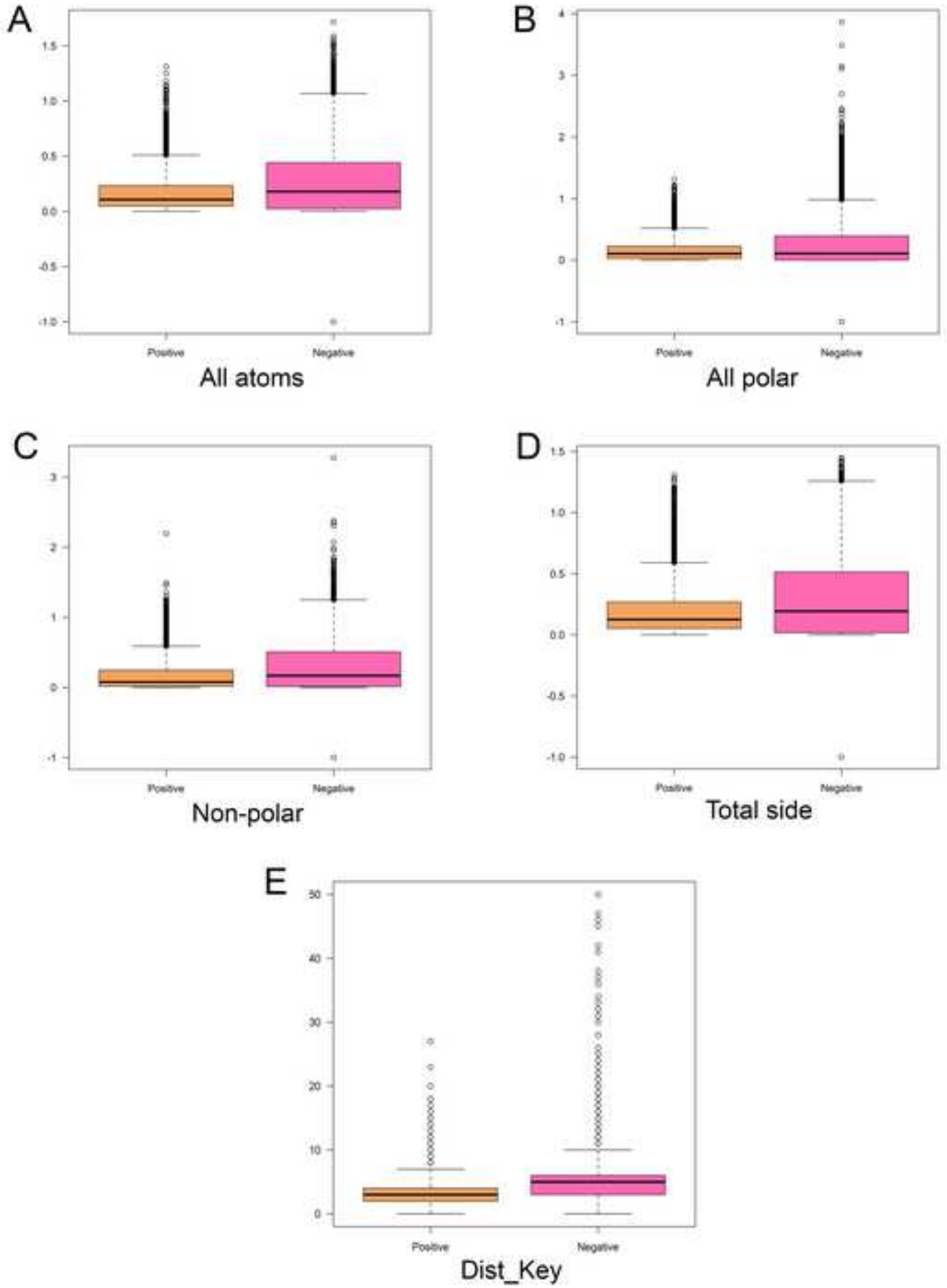
EF superfamily

EF fold

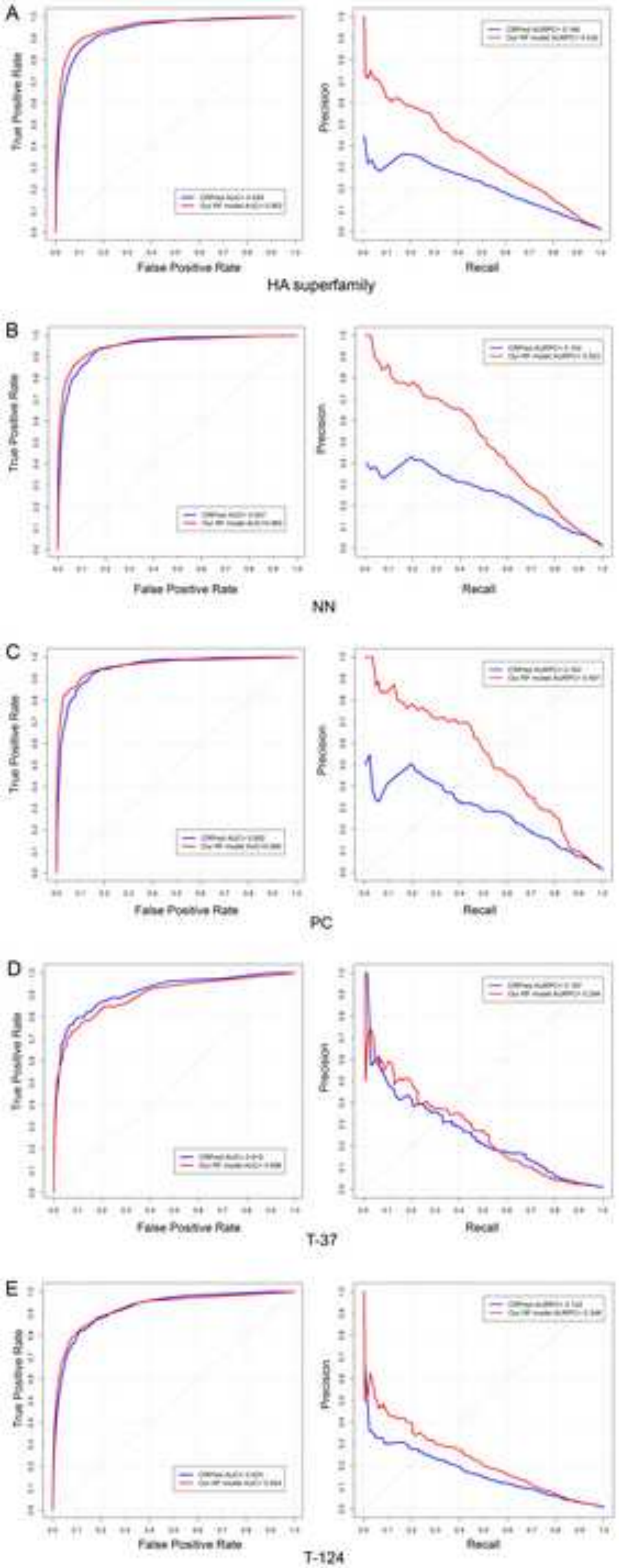
4. Figure 6


A
319R
273C
57R
136Q
58T
133H
106R
B
C
287R
315R
227E
127R
231R
D

4. Figure S1


A
1.5
1.0
0.5
0.0
-0.5
-1.0
Positive
Negative
All atoms
B
4
3
2
1
0
-1
Positive
Negative
All polar
C
3
2
1
0
-1
Positive
Negative
Non-polar
D
1.5
1.0
0.5
0.0
-0.5
-1.0
Positive
Negative
Total side
E
50
40
30
20
10
0
Positive
Negative
Dist_Key

4. Figure S2



HA superfamily

NN

PC

T-37

T-124

Tables

Table 1. A comprehensive summary of sequence, structure, and network features used in this study.

Feature category	**Dimensionality**	**Software /Database**	**References**	**Description**
Network features	12	iGraph	(Csardi and Nepusz, 2006)	Residue-contact network features include degree, closeness, status, hubscore, clustering coefficient, cyclic coefficient, constraint, betweenness, eigenvector, cocitation, coreness and eccentrality
Dist_Key	1	In-house		Relative sequential distance between catalytic residues
Network neighboring properties	3050	iGraph	(Csardi and Nepusz, 2006)	The residue-contact network features were used to describe the local spatial environment of catalytic residues
PSSM	260	PSI-BLAST	(Altschul, *et al.*, 1997; Jones, 1999)	Evolutionary information in the form of PSSM
EntWOP	1	PSI-BLAST	(Zhang, *et al.*, 2008)	Shannon entropy-based weighted observed percentage (WOP) calculated using PSI-BLAST of the catalytic residue of interest
Structure descriptors	6	Biopython	(Cock, *et al.*, 2009)	Structure descriptors include residue depth, contact number, HSEAU, HSEAD, HSEBU, and HSEBD
B-factor	1	PDB	(Rose, *et al.*, 2017)	B-factor or temperature factor
Solvent accessibility	5	Naccess	(Hubbard and Thornton, 1993)	This feature group include solvent accessibilities of all-atoms, Total-side, Main-chain, Non-polar and All-polar.
Secondary structure features	8	DSSP	(Kabsch and Sander, 1983)	Eight secondary structure types annotated by DSSP
Physicochemical property	3	BioJava	(Prlic, *et al.*, 2012)	This feature group include charged, hydrophobic and polar and is calculated from sequences.
CRPair	76	In-house	(Zhang, *et al.*, 2008)	A CRPair is a pair of catalytic residues in the protein. The sequence distances between any two adjacent catalytic residues are all calculated and collectively encode for the central catalytic residue. A total of 76 CRPairs were extracted.

Table 2. Performance comparison between LR and RF models of PREvaIL, CRpred, and other competing methods. All performance results were evaluated based on 10-fold cross-validation tests using the six datasets EF_fold, EF_superfamily, EF_family, HA_family, NN, and PC.

Method	**Performance measure**	**Performance evaluated on different datasets** (%)							
		EF_fold	**EF_superfamily**	**EF_family**	**HA_superfamily**	**NN** (without clustering)	**NN** (with clustering)	**PC**	**NN**
Competing methods	TP-rate	51.1[a]	53.9[a]	57.0[a]	29.3[b]	56.0[c]	68.0[d]	90[e]	50.0[f]
	Precision	17.1[a]	16.9[a]	18.5[a]	16.5[b]	14.0[c]	16.0[d]	7.0[e]	13.0[f]
CRpred	TP-rate	48.2[g]	52.1[g]	58.3[g]	54.0[g]	57.1[g]	57.1[g]	53.7[g]	57.1[g]
	Precision	17.0[g]	17.0[g]	18.6[g]	14.9[g]	17.8[g]	17.8[g]	17.5[g]	17.8[g]
Our LR model	TP-rate	52.5	53.1	55.0	46.1	50.0	50.0	53.7	50.0
	Precision	17.1	17.0	16.9	17.0	17.0	17.0	17.0	17.0
Our RF model	TP-rate	56.5	59.4	60.2	57.9	58.9	58.9	58.1	58.9
	Precision	17.0	17.0	17.0	17.0	17.0	17.0	17.0	17.0

[a] Performance results assessed on the EF_fold, EF_superfamily and EF_family datasets by Youn *et al.*, 2007, respectively;

[b] Performance results assessed on the HA_family dataset by Chea and Livesay, 2007;

[c] Performance results assessed on the NN dataset by using the structure-based method without spatial clustering by Gutteridge *et al.*, 2003;

[d] Performance results assessed on the NN dataset by using the structure-based method with spatial clustering by Gutteridge *et al.*, 2003;

[e] Performance results assessed on the PC dataset by Petrova and Wu, 2006;

[f] Performance results assessed on the NN dataset by using the sequence-based method by Gutteridge *et al.*, 2003;

[g] Performance results assessed on all the six datasets by using the sequence-based CRpred method by Zhang *et al.*, 2008.

Table 3. Performance comparison of different methods on the T-124 independent test dataset.

Method	**TP**	**FN**	**FP**	**TN**	**TP-rate**	**Precision**
CRpred[a] (all residues)	190	189	1131	47503	50.1	14.4
CRpred[b] (residues with coordinates)	190	189	1103	46017	50.1	14.7
HA[c] (residue identity filter)	105	274	549	46571	27.7	16.1
HA[d] (combination filter)	91	288	553	46567	24.0	14.1
Our LR model	183	185	1119	44339	49.7	14.9
Our RF model	229	139	1311	44147	62.2	14.9

[a] Performance results of CRpred by Zhang *et al.* (2008) based on all residues;

[b] Performance results of CRpred by Zhang *et al.* (2008) based on residues with coordinates;

[c] Performance results of the HA method by Chea and Livesay (2007) based on residue identity filter;

[d] Performance results of the HA method by Chea and Livesay (2007) based on combination filter.