Supplementary Material for: Active Site Prediction using Evolutionary and Structural Information

Sriram Sankararaman[1] Fei Sha[2] Jack Kirsch[3] Michael I. Jordan[1,4]
Kimmen Sjölander[5,6]

[1] Computer Science Division, University of California Berkeley, USA
[2] Computer Science Department, University of Southern California, USA
[3] Department of Molecular and Cell Biology, University of California Berkeley, USA
[4] Department of Statistics, University of California Berkeley, USA
[5] Department of Bioengineering, University of California, Berkeley, USA
[6] Department of Plant and Microbial Biology, University of California, Berkeley, USA

The DISCERN predictor for enzyme active site prediction is a statistical model that incorporates numerous features from sequence and structure to classify residues. DISCERN uses a statistical procedure, L1-regularization, to find a sparse set of features that are jointly predictive of enzyme active sites.

In the main text of this paper, we presented results comparing DISCERN to the best methods for catalytic residue prediction on two challenging manually curated benchmark datasets: a dataset of 140 enzymes from the CATRES dataset (CATRES-FAM) (Bartlett *et al.*, 2002) and a dataset of 423 enzymes from the Catalytic Site Atlas (CSA) (Porter *et al.*, 2004) used to benchmark the FRcons method (Fischer *et al.*, 2008). We showed that DISCERN achieves improvements in recall between 12-20% over the published results of state-of-the-art methods combining sequence and structure information (or inference), and an increase of 50% and higher over methods that make use of only evolutionary conservation signal.

In this supplement, we present details on: (i) the L_1-regularized logistic regression underlying DISCERN, (ii) the features evaluated for catalytic residue prediction, (iii) the experimental setup used to evaluate DISCERN, (iv) results from two additional datasets: one containing 121 enzymes from the CATRES dataset made non-redundant at the SCOP superfamily level (CATRES-SF), and a dataset of 94 enzymes from the Catalytic Site Atlas made non-redundant at the SCOP family level (CSA-FAM), (v) controlled experiments testing the contribution of various aspects of the DISCERN predictor to prediction accuracy, (vii) a case study of DISCERN predictions on *Escherichia coli* Asparagine Synthetase (PDB id:12as) and (viii) a comparison of DISCERN to a Conditional Random Field approach to catalytic residue prediction. We also provide additional details on the comparison of DISCERN to the FRcons method (Fischer *et al.*, 2008).

S-1 L_1-regularized logistic regression

Given an enzyme i with n_i amino acid residues, we denote by $\boldsymbol{x}_j^{(i)}$ the d-dimensional vector of residue-specific features at site j, $j = 1, \ldots, n_i$, by $\boldsymbol{X}^{(i)}$ the $d \times n$ matrix of all such features, and by $z_j^{(i)} \in \{+1, -1\}$ the catalytic label of residue j (whether the residue is catalytic or not). We denote the set of structural neighborhood features by a $dN \times n$ matrix $\boldsymbol{Y}^{(i)}$. Here N refers to the number of structural neighbors of each residue.

We pick the ten residues closest to residue j to form the set of structural neighbors (the distance $d_{j,k}$ between two residues is defined as the minimum of the distance among all pairs of atoms).[1]

We model the conditional distribution of the random variable $Z_j^{(i)} \in \{+1, -1\}$ by a logistic regression

$$\Pr(Z_j^{(i)} = 1 | \boldsymbol{X}^{(i)}, \boldsymbol{Y}^{(i)}, b, \boldsymbol{w}_1, \boldsymbol{w}_2) = \frac{1}{1 + \exp\left(-\left(b + \boldsymbol{w}_1{}' \boldsymbol{x}_j^{(i)} + \boldsymbol{w}_2{}' \boldsymbol{y}_j^{(i)}\right)\right)}. \quad \text{(S-1)}$$

[1] The choice of ten residues as neighbors is arbitrary. It is also possible to treat the size of the structural neighborhood as a parameter and estimate it.

The model has parameters $(b, \boldsymbol{w}_1, \boldsymbol{w}_2)$; b is the intercept term which controls the tradeoff between false positives and false negatives, $\boldsymbol{w}_1$ controls the weights of the residue features while $\boldsymbol{w}_2$ controls the weights of the features from the structural neighbors. Given a training set of enzymes and their catalytic residue annotations, we can estimate the parameters $(b, \boldsymbol{w}_1, \boldsymbol{w}_2)$. To encode a preference for a "sparse" parameter vector, we adopt a regularized maximum likelihood approach in which we maximize the sum of the likelihood and an L_1 penalty term:

$$\max_{\boldsymbol{w}} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \log \Pr(z_j^{(i)} | \boldsymbol{X}^{(i)}, \boldsymbol{Y}^{(i)}, b, \boldsymbol{w}) - \lambda \|\boldsymbol{w}\|_1, \tag{S-2}$$

where $\boldsymbol{w} = (\boldsymbol{w}_1, \boldsymbol{w}_2)$ and where $\|\boldsymbol{w}\|_1 = \sum_k |w_k|$ is the L_1 norm. The non-negative regularization parameter λ controls the sparsity of the estimate of $\boldsymbol{w}$; larger values of λ lead to estimates with increasing numbers of zero components. We chose the value of λ by a cross-validation procedure. The optimization problem is solved using an interior point method as implemented in Koh *et al.* (2007).

Enforcing sparsity on the parameter vector using L_1-regularization not only leads to a more interpretable fitted parameter vector; it also helps to prevent *overfitting*. The problem of overfitting, which is well known in statistics (Hastie *et al.*, 2001), is as follows: when a statistical model contains a large number of parameters relative to the size of the *training set*, the model tends to fit the noise in the training data, leading to high accuracy on the training set but poor performance on novel data. Regularization imposes a constraint on the parameter space (e.g., by limiting the size of the parameters as measured by the L_1 norm) reducing the "effective degrees of freedom" of the model and forcing the model to generalize more effectively.

S-2 Features evaluated for catalytic residue prediction

The DISCERN logistic regression predictor is based on a feature vector having 528 component features. See Table S-5.

S-2.1 Sequence conservation features

Sequence conservation has been observed to be the most important feature for catalytic residue prediction (Gutteridge *et al.*, 2003; Youn *et al.*, 2007). We tested three sequence conservation scores. The first, GLOBAL-JS, is the Jensen-Shannon divergence (Lin and Wong, 1990) between the amino acid distribution at a column and a background distribution (with prior weight = 0.5 as in Capra and Singh (2007)). The other two sequence conservation scores tested make explicit use of the phylogenetic tree topology using the INTREPID algorithm (Sankararaman and Sjölander, 2008). INTREPID has been shown to be sensitive for catalytic residue prediction in general and in particular is able to exploit the information in large divergent families. The two variants used the Jensen-Shannon divergence (INTREPID-JS) and the log frequency of the modal amino acid (INTREPID-LO). Further, the INTREPID scores can be efficiently computed, even for large protein families. The average running time of INTREPID on the CATRES-FAM datasets was 65s. See (Sankararaman and Sjölander, 2008) for details of these scoring functions.

S-2.2 Amino acid properties

Amino acids have varying catalytic propensities. We use the 20 amino acids as separate features and also classify the amino acid into one of three categories—charged (D,E,H,K,R), polar (Q,T,S,N,C,Y) or hydrophobic (A,F,G,I,L,M,P,V,W). We used the classification described in Bartlett *et al.* (2002) with one modification. Tryptophan is included among the class of polar residues in Bartlett *et al.* (2002) but among hydrophobic residues by others (Eisenberg *et al.*, 1982); we use the latter classification.

S-2.3 Structure-based features

For each residue, we compute the residue centrality, the B-factor, solvent accessibility, presence in a cleft and secondary structure as follows. We compute the B-factor, a measure of thermal motion for each residue, as the average of the B-factors of all its atoms (derived directly from its PDB file). We compute a measure of centrality for each residue j as the inverse of the average distance from a residue to all other residues in the enzyme; i.e., $C_j = \frac{n-1}{\sum_{k \neq j} d(k,j)}$ where n is the number of residues in the structure and $d(k, j)$ is the distance from j to k along the contact map. A residue that is located in the center of the protein has smaller average distance to all other residues

and hence a high centrality measure. We use the 7-state secondary structure representation output by DSSP (Kabsch and Sander, 1983). The area of a residue accessible to solvent is obtained from NACCESS (Hubbard and Thornton, 1993). We use LigSite^{csc} (Huang and Schroeder, 2006) to detect the presence of a residue in one of the three largest pockets in the enzyme.

S-3 Details on the computational pipeline

S-3.1 Homolog selection and alignment

For each of the four datasets used in these experiments, PSI-BLAST (Altschul *et al.*, 1997) was run for four iterations against the UniProt database (Apweiler *et al.*, 2004) with an E-value inclusion threshold of 1×10^{-4} from which a maximum of 1000 homologs were retrieved. A multiple sequence alignment (MSA) was estimated using MUSCLE (Edgar, 2004) with MAXITERS set to 2, followed by removing identical sequences and deleting columns in which the seed had a gap.

For CATRES-SF, the set of alignments built contain a minimum of 32 sequences, a maximum of 1033 sequences, and a median of 839 sequences. The average percent identity between the seed sequence and homologs in the alignments varies from 6.42% to 31.14% with a median of 15.22%. Percent identity was computed as the fraction of the alignment columns that have identical characters in the sequence and the seed (i.e., the number of identical columns divided by the number of amino acids in the seed). The low percent identity is partly attributed to the inclusion of many sequences with local alignments in the MSA.

S-3.2 Tree construction

A neighbor-joining tree was built from this alignment using the PROTDIST and NEIGHBOR programs in the PHYLIP package (Felsenstein, 1993). The programs were run with default parameters. We used midpoint rooting (placing the root at the midpoint of the longest span in the tree).

S-4 Experiments

S-4.1 Benchmark datasets

We used four datasets in these experiments, two (CSA-FAM and CSA-Fischer) derived from the manually curated section of the Catalytic Site Atlas (CSA) (Porter *et al.*, 2004) and two (CATRES-FAM and CATRES-SF) derived from the CATRES (Bartlett *et al.*, 2002) resource. We developed these different datasets to allow comparisons between Discern and other methods, and used the Structural Classification of Proteins (SCOP) (Murzin *et al.*, 1995) to filter datasets at different levels of homology.

SCOP is a hierarchical classification of protein domains based on their structural, functional and sequence similarities. Domains in different SCOP folds are unrelated; domains in the same fold but different superfamilies have an uncertain relationship (i.e., although their topologies are similar, there is no other evidence to support homology); domains in the same superfamily are deemed homologous; domains in the same family have very similar functions and structures. The suffix "-FAM" indicates datasets filtered to remove redundancy at the SCOP family level, while the suffix "-SF" indicates datasets filtered at the SCOP superfamily level. Datasets filtered more stringently are naturally more challenging than datasets that include more closely related sequences (i.e., the SCOP superfamily-level datasets are harder than the family-level datasets). In the main text, we described experiments on CATRES-FAM and CSA-Fischer. Additional experiments reported in these Supplementary Materials describe experiments on CATRES-SF and CSA-FAM.

CATRES-FAM consists of 140 enzymes from the CATRES dataset. The CATRES dataset consists of enzymes with PDB structures with catalytic site information assigned from the literature. Subsets of this dataset have been used by previous methods for catalytic residue prediction (Gutteridge *et al.*, 2003; Tong *et al.*, 2008). The original CATRES dataset contains 178 enzymes. We discarded 26 enzymes as unusable in these experiments for various reasons: 21 enzymes presented problems for one or more of our feature extraction programs (18 had catalytic sites spanning multiple sub-units, and three enzymes had non-numeric PDB residue identifiers), one of the enzymes had no annotated catalytic residues, one had only one detectable homolog using PSI-BLAST, MUSCLE crashed on another, and two NMR structures were also discarded as unusable by the structure-based methods. The resulting set of

enzymes was made non-redundant at the SCOP family level by removing an additional 12 enzymes. The resulting dataset contains a total of 472 catalytic residues out of a total of 49,180 residues with a median of three catalytic residues per enzyme.

CATRES-SF consists of 121 enzymes from CATRES made non-redundant at the SCOP superfamily level (i.e., no pair of enzymes belongs to the same SCOP superfamily). This dataset is thus filtered at a more stringent level than CATRES-FAM, presenting a greater challenge to statistical models using this dataset in cross-validation.

CSA-FAM contains 94 enzymes chosen from CSA such that (i) no pair contained domains in the same SCOP family, (ii) no pair had detectable sequence homology (enforced by a BLAST E-value >1), and (iii) each of the sequences had pre-computed results in the Baylor College of Medicine Evolutionary Trace server. (The last requirement was designed to enable a direct comparison with Evolutionary Trace without putting undue load on their servers.)

CSA-Fischer consists of 423 enzymes from the CSA selected by Fischer and colleagues to benchmark FRcons (Fischer *et al.*, 2008), and used in these experiments to evaluate DISCERN relative to FRcons. We used the same protocol established by Fischer *et al.* in performing two-fold cross-validation, and ensuring that no domains from the same SCOP superfamily were found in both the training and test sets for either partition.

S-4.2 Performance measurements

We measure the precision and the recall on the test set where: Precision = $\frac{TP}{TP+FP}$, Recall = $\frac{TP}{TP+FN}$, a true positive (TP) is a residue included in the benchmark dataset that is predicted as catalytic, a false positive (FP) is a residue not listed in the benchmark that is predicted as catalytic, and a false negative (FN) is a catalytic residue in the benchmark which has been missed by a method. The precision-recall curves were averaged over all the cross-validation folds using the code from (Davis and Goadrich, 2006).

S-4.2.1 A note on cross-validation

k-fold cross-validation is a procedure to evaluate the accuracy of a predictor. The data is partitioned into k equal-sized subsets. In each fold, one partition is chosen as the test data and the rest of the data forms the training data; e.g., in 10-fold cross-validation, $9/10^{th}$ of the data would be used to estimate the model parameters, and then tested on the reserved $1/10^{th}$ of the data. In the next fold, a different $1/10^{th}$ is used to test. The accuracy of the predictor, as measured on the *test* dataset, is averaged over the folds to obtain a final estimate of the accuracy.

To assess the performance of Discern, we performed k-fold cross validation over the enzymes in each benchmark dataset. We used k=2 for CSA-Fischer (to allow a direct comparison with Fischer *et al.*), and k=10 for each of the other datasets.

Note that in cross-validation, the characteristics of the dataset can have a major impact on the performance. In particular, the presence of homologs in the dataset can lead to an increase in the *apparent* accuracy (i.e., an overestimate of the expected accuracy of the method when applied to novel data) when these homologs occur in both the training and the test set (see a discussion of this issue in (Youn *et al.*, 2007)). This is the reasoning behind Fischer *et al.*'s ensuring that no enzymes from the same SCOP superfamily were found across both sides of the cross-validation fold. We also observe a similar decrease in accuracy on the CATRES-SF dataset (non-redundant at the SCOP superfamily level) relative to CATRES-FAM (non-redundant at the SCOP family level). The L_1-regularization parameter was estimated by a similar cross-validation within the training set in each fold of the cross-validation.

S-4.3 ConSurf

The ConSurf-DB database of pre-computed results (http://consurfdb.tau.ac.il) was used to obtain results on the CATRES sequences while the ConSurf web server at Tel Aviv University (http://consurf.tau.ac.il) was used to obtain the results on CSA-FAM.

S-4.4 Evolutionary Trace

Evolutionary Trace results were obtained from the pre-computed results of the Evolutionary Trace server at the Baylor College of Medicine (http://mammoth.bcm.tmc.edu/report_maker).

S-4.5 INTREPID

INTREPID results were obtained using the alignments gathered for each dataset (as described in the main text Materials and Methods), using the algorithm described in Sankararaman and Sjölander (2008).

S-4.6 FRcons

The FRcons method (Fischer *et al.*, 2008) uses sequence information only to predict functional residues, but predicts structural features to boost prediction accuracy. FRcons uses a Bayesian network to estimate the probability that a position is functional given its amino acid frequency distribution, conservation score, predicted relative solvent accessibility, and predicted helix and extended sheet propensities. Fischer *et al.* evaluated their method on two tasks: catalytic residue prediction and ligand-binding residue prediction. To benchmark FRcons accuracy at catalytic residue prediction, they used two-fold cross-validation on a dataset of 423 SCOP family domains from the CSA, ensuring that no domains from the same SCOP superfamily appeared in both training and test data. Fischer *et al.* showed a performance of roughly 50% recall at 18.5% precision and 57% recall at 14% precision, i.e., approaching the accuracy of methods that use actual information from both sequence and structure. For comparisons between DISCERN and FRcons, we obtained raw scores from Fischer and colleagues in producing the Precision-Recall curves.

S-4.7 Youn *et al.*

Youn and colleagues (Youn *et al.*, 2007) used a Support Vector Machine (SVM) approach, including features such as amino acid residue type, sequence conservation, the structural environment of each residue represented by 4 shells of thickness 1.875$\AA$, each consisting of 264 atom-based descriptors (Bagley and Altman, 1995), and a structural conservation obtained by comparing the structural environment at each residue. They reported their performance using 10-fold cross-validation on three datasets of enzymes with solved structures from ASTRAL 40v1.65 (Chandonia *et al.*, 2004), filtered to remove redundancy at the SCOP fold, superfamily and family levels respectively. Their dataset contained a total of 987 protein domains, classified into 396 families, 236 superfamilies and 189 folds. Youn *et al.* reported a recall of 57.02% at a precision of 18.51% on the family-level dataset, a recall of 53.93% at a precision of 16.90% on the superfamily-level dataset, and a recall of 51.11% at a precision of 17.13% on the fold-level dataset.

S-4.8 Gutteridge *et al.*

Gutteridge and colleagues (Gutteridge *et al.*, 2003) used a neural network for catalytic residue prediction based on amino acid residue type, sequence conservation features and structural features such as presence in a pocket, B-factor and solvent accessibility. Each residue was classified using the above features computed at the residue alone; i.e., features computed at the structural neighbors were not considered for prediction. The neural network was evaluated by 10-fold cross-validation on 159 enzymes from the CATRES dataset, on which they reported a recall of 56% at a precision of 14%.

S-5 Results

Results on CATRES-FAM and CSA-Fischer are reported in the main text. In this section, we report results on the two other datasets: CATRES-SF and CSA-FAM.

S-5.1 DISCERN performance on CATRES-SF

CATRES-SF was designed to be similar to the Youn *et al.* superfamily-level dataset, and presents a greater challenge to prediction methods than CATRES-FAM due to the more stringent level of homology filtering. At a precision of 17%, DISCERN attains a recall of 65% on CATRES-SF, while Youn *et al.* report a recall of 53.9% at 16.9% precision on their superfamily-level dataset (i.e., DISCERN attains an improvement of 11% at the same level of precision relative to Youn *et al.*).

S-5.2 DISCERN performance on CSA-FAM

The CSA-FAM dataset was designed to enable a direct comparison with Evolutionary Trace (ET) using pre-calculated results from the Baylor College of Medicine ET server (Mihalek *et al.*, 2004). On this dataset, DISCERN achieves a recall of 75% at 18.5% precision (full precision-recall results are available in figure S-4). We also compared DISCERN against INTREPID (Sankararaman and Sjölander, 2008), ConSurf, Youn *et al.*, Gutteridge *et al.*, and a control method, trained identically to DISCERN but which does not make use of INTREPID phylogenomic conservation scores or features computed from structural neighbors, and without the use of L_1-regularization to enforce model sparsity (see Section S-5.3 for additional details). Results for INTREPID, ConSurf and Evolutionary Trace are on the same enzymes. For comparison against Youn *et al.*, we include the reported performance of their method on their SCOP family-level dataset (i.e., similar to CSA-FAM), on which they report 57.02% recall at 18.5% precision. We also include results from Gutteridge *et al.* on the CATRES dataset on which they report a recall of 56% at 14% precision.

Figure S-4 shows that DISCERN attains an improvement in recall over all methods on this dataset. At the same level of precision, DISCERN has 23% greater recall relative to Youn *et al.*, and 21% greater recall relative to the control. Relative to Gutteridge *et al.*, DISCERN shows 19% greater recall and 4.5% greater precision.

S-5.3 Controlled experiments to test the effect of including phylogenomic conservation score, features computed for structural neighbors, and L_1- regularization

The accuracy of the DISCERN predictor depends critically on the inclusion of discriminative features while avoiding model overfitting. To assess the relative contribution of different features we tested the predictive power of statistical models trained identically to DISCERN but withholding certain features. Performance was assessed on the CATRES-FAM dataset using 10-fold cross validation. Table 1 gives details on individual models and Figure S-6 shows full precision-recall curves on the CATRES-FAM dataset. For direct comparison with published results of other methods, we refer in this section to the recall of each method at 18% precision, and to the precision of each method at 50% recall.

Method 0, our control, is an unregularized logistic regression with no features from structural neighbors and no phylogenomic conservation scores (i.e., it uses only GLOBAL-JS, a measure of the family-wide conservation). The control is designed to be similar to methods that exploit information from both sequence and structure but do not use features computed at structural neighbors, do not exploit the phylogenetic information and do not use L_1-regularization to enforce sparsity. The control attains a recall of 48% at 18% precision on the CATRES-FAM dataset.

Method 1 expands on the control through the inclusion of INTREPID phylogenomic conservation scores, achieving a recall of 55% at 18% precision, corresponding to an increase of 7% in recall relative to the control.

Method 2 expands on Method 1 through the inclusion of features computed at structural neighbors but does not include L_1-regularization. Method 2 attains a recall of 41% at a precision of 18%. We see that naively including features from structural neighbors leads to a decrease in performance (reducing recall by 14%), suggestive of model overfitting.

DISCERN expands on Method 2, but also includes L_1-regularization to enforce sparsity. This yields a recall of 69% at 18% precision, corresponding to a 28% improvement in recall relative to Method 2. Relative to the control and Method 1, DISCERN has 21% and 14% greater recall respectively.

Proceeding from the control to DISCERN also shows a dramatic reduction in false positive predictions (residues predicted as catalytic which are not listed in the CATRES dataset). Measuring precision (the fraction of predicted residues that are actually catalytic) at the point where half of the catalytic residues have been detected (i.e., a recall of 50%) shows that the control has precision of 17.0% while DISCERN has 27.3% precision. In other words, DISCERN effectively reduces the ratio of false positives to true positives from 4.1 to 2.8.

S-6 Case Studies

S-6.1 *Escherichia coli* Asparagine Synthetase (PDB id:12as, E.C. number: 6.3.1.1)

L-Asparagine synthetase catalyzes the conversion of L-aspartic acid and ammonia to L-asparagine in the presence of a magnesium ion while hydrolyzing ATP to AMP and pyrophosphate (Meister, 1974). L-Asparagine synthetase

from *Escherichia coli* has three catalytic residues identified in the CATRES dataset—D46, R100 and Q116 (Nakatsu *et al.*, 1998).

Figure S-11 presents DISCERN predictions at the point where all the catalytic residues listed in CATRES were selected, based on model parameters derived when 12as was in the test set of the cross-validation (i.e., not used in training). The number of residues selected by DISCERN is thus equal to the worst rank (16) it gives to a catalytic residue listed in CATRES.

We separately examined the 20 top-ranked residues for DISCERN (see Table S-1 in Supplementary Materials). In addition to the three CATRES-selected catalytic residues, DISCERN identifies seven residues (K77, E120, D219, D235, E248, S251, and R255) which have been shown or inferred to play functional roles (Nakatsu *et al.*, 1998). Of the ten remaining residues, many are found in clusters with residues that have been functionally characterized. These form three sequence motifs that are proximal in structure but separate in primary sequence. Motif 1 includes H71, K75 and K77. Of these, K77 has been proposed, based on homology with the catalytic domain of yeast class II aspartyl-tRNA synthetase, to interact with the β-carboxylate group of L-aspartic acid (Nakatsu *et al.*, 1998). Motif 2 includes D115, Q116, D118, W119 and E120; all lie on a single beta strand that lines the active site cleft (referred to as β-6). Of these, Q116 is included in CATRES, and E120 has been proposed to interact with the β-carboxylate group of L-aspartic acid (Nakatsu *et al.*, 1998). Motif 3 includes R214, Y218, D219 and D220. Of these, the side chain carboxyl group of D219 has been observed to interact with the amino group of the L-asparagine through a water molecule (Nakatsu *et al.*, 1998).

S-7 Conditional Random Field for catalytic residue prediction

The logistic regression model in DISCERN exploits the structural context by combining features from the structural neighbors but still makes independent predictions of the catalytic label at each residue. In this section, we describe an alternate model based on the framework of Conditional Random Fields (CRFs) (Lafferty *et al.*, 2001). CRFs allow us to capture contextual information by coupling the labels of the structural neighbors and making a joint prediction across all the residues. In principle CRFs can capture more complex dependencies than a model that treats each residue independently. A dependency of the form *structurally proximal residues X and Y tend to be in the same cleft if they are both catalytic* is one example since it is a function of the features and the residue labels (which need to be inferred).

We define a CRF for the catalytic residue prediction problem as follows:

$$\begin{aligned}
&\log \Pr(\boldsymbol{z}^{(i)}|\boldsymbol{X}^{(i)}, b, \boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{w}_3) = \boldsymbol{w}'\phi(\boldsymbol{z}, \boldsymbol{X}^{(i)}) - Z^{(i)}(b, \boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{w}_3) \\
&= b + \sum_{j=1}^{n_i} \left(z_j^{(i)} \boldsymbol{w}_1{}' \boldsymbol{x}_j^{(i)} + z_j^{(i)} \boldsymbol{w}_2{}' \boldsymbol{y}_j^{(i)} + \boldsymbol{w}_3{}' \sum_{k \in N^{(i)}(j)} \psi(z_j^{(i)}, z_k^{(i)}, \boldsymbol{X}^{(i)}) \right) \\
&- Z^{(i)}(b, \boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{w}_3),
\end{aligned} \tag{S-3}$$

where $\boldsymbol{w} = (b, \boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{w}_3)$ and $Z^{(i)}(b, \boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{w}_3) = \log\left(\sum_{\boldsymbol{z}} \exp\left(\boldsymbol{w}'\phi(\boldsymbol{z}, \boldsymbol{X}^{(i)})\right)\right)$ is the log normalizer. Here, in addition to the features used in the logistic regression model, we have extra interaction features ψ to capture dependencies between the labels of two neighboring catalytic residues z_j, z_k. Setting $\boldsymbol{w}_3$ to zero in Equation S-3 results in the logistic regression model discussed earlier.

To predict the labels of all the residues jointly, we would like to obtain the labeling $\boldsymbol{z}^{(i)*}$ with highest posterior probability.

$$\boldsymbol{z}^{(i)*} = \arg\max_{\boldsymbol{z}} \log \Pr(\boldsymbol{z}|\boldsymbol{X}^{(i)}, b, \boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{w}_3). \tag{S-4}$$

The configuration $\boldsymbol{z}^{(i)*}$ can be computed efficiently provided the interaction features ψ are chosen carefully. We use a maximum margin approach to estimate the parameters $\boldsymbol{w}$.

S-7.1 Maximum Margin Parameter Estimation for the CRF

For general interaction features ψ, the problem of computing the maximum *a posteriori* (MAP) configuration z^* of the CRF described in Equation S-2 is NP-hard (Boykov *et al.*, 2001). Efficient algorithms based on graph cuts exist for computing z^* when the interaction features are sub-modular; i.e., $\psi(0,0,x)+\psi(1,1,x) \geq \psi(0,1,x)+\psi(1,0,x)$ (Boykov

et al., 2001; Kolmogorov and Zabih, 2002; Boykov and Kolmogorov, 2004). We therefore restrict the model to sub-modular interaction features ψ which take values in $\{0, 1\}$—this restriction allows us to estimate the parameters $\boldsymbol{w}$ that respect the sub-modularity constraint for all inputs.

We use a maximum margin approach to estimate the parameters $\boldsymbol{w}$ of the CRF. The maximum margin framework leads to the following optimization problem

$$\begin{aligned}
&\min_{\boldsymbol{w}} \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^{m}\xi_i \qquad \text{such that} \\
&\boldsymbol{w}'\left(\phi(\boldsymbol{z}^{(i)}, \boldsymbol{X}^{(i)}) - \phi(\boldsymbol{z}, \boldsymbol{X}^{(i)})\right) \geq L(\boldsymbol{z}^{(i)}, \boldsymbol{z}) - \xi_i, \qquad \forall i = 1, \ldots, m, \forall \boldsymbol{z} \in \{+1, -1\}^{n_i} \\
&\xi_i \geq 0, \qquad \forall i = 1, \ldots, m \\
&\boldsymbol{w}_3\left(\psi(0, 0, x) + \psi(1, 1, x) - \psi(1, 0, x) - \psi(0, 1, x)\right) \geq 0 \qquad \forall x.
\end{aligned}$$

The first constraint requires the model to give the highest score to the true labeling $\boldsymbol{z}^{(i)}$. All other labelings are assigned scores lower than the score for the true labeling; the difference in the scores depends on a cost function $L(\boldsymbol{z}^{(i)}, \boldsymbol{z})$. We use the Hamming distance as the cost function—a labeling that is very different from the true labeling should be assigned a lower score than one that is more similar. To handle nonlinearly separable data, we introduce the non-negative slack variables $\xi_i, i = 1 \ldots, m$. The final constraint ensures that the fitted model has no non-sub-modular interaction features so that z^* can be efficiently computed.

We can replace the first constraint with the equivalent

$$\boldsymbol{w}'\phi(\boldsymbol{z}^{(i)}, \boldsymbol{X}^{(i)}) \geq \boldsymbol{w}'(\phi(\hat{\boldsymbol{z}}^{(i)}, \boldsymbol{X}^{(i)})) + L(\boldsymbol{z}^{(i)}, \hat{\boldsymbol{z}}^{(i)})) - \xi_i, \forall i = 1, \ldots, m,$$

where $\hat{\boldsymbol{z}}^{(i)} = \arg\max_{\boldsymbol{z}} \boldsymbol{w}'(\phi(\boldsymbol{z}, \boldsymbol{X}^{(i)})) + L(\boldsymbol{z}^{(i)}, \boldsymbol{z}))$. The Hamming distance loss does not affect any of the interaction features so that $\hat{\boldsymbol{z}}^{(i)}$ can be computed efficiently. The original optimization problem now reduces to

$$\begin{aligned}
&\min_{\boldsymbol{w}} \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^{m}\boldsymbol{w}'\left(\phi(\hat{\boldsymbol{z}}^{(i)}, \boldsymbol{X}^{(i)}) + L(\boldsymbol{z}^{(i)}, \hat{\boldsymbol{z}}^{(i)}) - \phi(\boldsymbol{z}^{(i)}, \boldsymbol{X}^{(i)})\right) \\
&\boldsymbol{w}_3(\psi(0, 0, x) + \psi(1, 1, x) - \psi(1, 0, x) - \psi(0, 1, x)) \geq 0 \qquad \forall x.
\end{aligned}$$

This is a convex program with a non-differentiable objective function which we solve using a subgradient method. In practice, we use the L_1-regularized logistic regression to estimate the parameters $(b, \boldsymbol{w}_1, \boldsymbol{w}_2)$, discard the zero weights and only estimate the interaction parameter vectors $(\boldsymbol{w}_2, \boldsymbol{w}_3)$.

S-7.2 Features used in the CRF

In addition to the features used in the logistic regression, we compute three additional feature functions for the CRF (described by the ψ terms in Equation S-2). Each of these feature functions operates on pairs of neighboring residues; i.e., a pair is predicted as catalytic if they share one of these features: charged, polar or conserved. (Recall that $z_j = 1$ if residue j is predicted catalytic.) The first two feature functions couple two neighboring residues if they are both polar or both charged. The last feature function couples two neighboring residues that are both highly conserved (the INTREPID scores are normalized to have zero mean and unit variance for each enzyme).

$$\psi_1(z_j, z_k, x) = \begin{cases} 1 & \text{if } z_j = z_k = 1 \text{ \& } j, k \text{ are polar} \\ 0 & \text{otherwise} \end{cases}$$

$$\psi_2(z_j, z_k, x) = \begin{cases} 1 & \text{if } z_j = z_k = 1 \text{ \& } j, k \text{ are charged} \\ 0 & \text{otherwise} \end{cases}$$

$$\psi_3(z_j, z_k, x) = \begin{cases} 1 & \text{if } z_j = z_k = 1 \text{ \& INTREPID scores for j,k} > 1 \\ 0 & \text{otherwise} \end{cases}$$

S-7.3 Comparison of CRF to DISCERN

We see from Table S-3 that the CRF has very similar accuracies to DISCERN with no change in recall on the CATRES-FAM dataset. The extra structural features used in the CRF attained low weights with the highest weight (0.122) being assigned to the feature that enforces agreement between two structural neighbors if each appears conserved. This is likely a result of the small number of catalytic sites observed in the dataset so that the new features introduced by the CRF do not capture any dependencies in addition to those captured at the feature level by the logistic regression model.

S-8 Discussion

In the main text, we presented the DISCERN predictor and reported its performance on the CATRES-FAM and CSA-Fischer datasets. We showed that DISCERN has superior accuracy relative to other methods using information from sequence and structure, and also (as expected) to methods that are restricted to evolutionary conservation signal only.

In these Supplementary Materials, we show that DISCERN retains its improved performance relative to other methods on a dataset of enzymes made non-redundant at the superfamily level (CATRES-SF). Results on CATRES-SF also highlight the decrease in accuracy for all methods as datasets are filtered at increasing levels of stringency. For instance, at a precision of 17%, DISCERN attains a recall of 70% on CATRES-FAM (filtered to remove homologs from the same SCOP family, but retaining superfamily members) but a recall of 65% on CATRES-SF (filtered to remove both SCOP family and superfamily members).

We separately evaluated the contribution of individual elements to the accuracy of DISCERN and demonstrated the importance of controlling model complexity using L_1-regularization. Paradoxically, the inclusion of many features is intended to improve a predictor's accuracy, but can reduce its ability to generalize to recognize novel data. This problem is called *model overfitting*, and often arises when the ratio of model parameters to training data is large (Hastie *et al.*, 2001). Thus, while Youn and colleagues improved significantly upon the state-of-the-art in catalytic site prediction by including information from structural neighbors (Youn *et al.*, 2007), the additional model complexity may have reduced its ability to generalize successfully. We also built a model in which the features from the structural neighbors were averaged as a function of their distance. This model attained accuracies similar to DISCERN (data not shown).

We considered an extension to logistic regression, based on the framework of Conditional Random Fields (CRF). CRF methods go beyond a simple logistic regression to allow the coupling of catalytic labels for different residues, enabling us to capture more complex dependencies and to make a joint prediction of the residue labels. In practice, we find that the accuracy of the CRF is virtually indistinguishable from DISCERN.

The data used in these experiments—i.e., the multiple sequence alignments, phylogenetic trees and PDB files—are available for download from our website (http://phylogenomics.berkeley.edu/discern/supplement.html).

References

Alterovitz, R., Arvey, A., Sankararaman, S., Dallett, C., Freund, Y., and Sjölander, K. (2009). Resboost: characterizing and predicting catalytic residues in enzymes. *BMC Bioinformatics*, **10**(1), 197.

Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**(17), 3389–3402.

Apweiler, R., Bairoch, A., Wu, C., Barker, W., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M., Natale, D., O'Donovan, C., Redaschi, N., and Yeh, L. (2004). UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.*, **32**, D115–119.

Bagley, S. C. and Altman, R. B. (1995). Characterizing the microenvironment surrounding protein sites. *Protein Sci.*, **4**(4), 622–635.

Bartlett, G. J., Porter, C. T., Borkakoti, N., and Thornton, J. M. (2002). Analysis of catalytic residues in enzyme active sites. *J. Mol. Biol.*, **324**(1), 105–121.

Berna, P. P., Mrabet, N. T., Van Beeumen, J., Devreese, B., Porath, J., and Vijayalakshmi, M. A. (1997). Residue accessibility, hydrogen bonding, and molecular recognition: metal-chelate probing of active site histidines in chymotrypsins. *Biochemistry*, **36**, 6896–6905.

Birktoft, J. J., Kraut, J., and Freer, S. T. (1976). A detailed structural comparison between the charge relay system in chymotrypsinogen and in alpha-chymotrypsin. *Biochemistry*, **15**, 4481–4485.

Boykov, Y. and Kolmogorov, V. (2004). An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Trans. Pattern Anal. Mach. Intell.*, **26**(9), 1124–1137.

Boykov, Y., Veksler, O., and Zabih, R. (2001). Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **23**, 2001.

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

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

Chandonia, J. M., Hon, G., Walker, N. S., Lo Conte, L., Koehl, P., Levitt, M., and Brenner, S. E. (2004). The ASTRAL Compendium in 2004. *Nucleic Acids Res.*, **32**(Database issue).

Davis, J. and Goadrich, M. (2006). The relationship between precision-recall and roc curves. ICML '06: Proceedings of the 23rd International Conference on Machine Learning, pages 233–240, New York. ACM.

Edgar, R. C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**(1).

Eisenberg, D., Weiss, R. M., Terwilliger, T. C., and Wilcox, W. (1982). Hydrophobic moments and protein structure. *Faraday Symp. Chem. Soc.*, **17**, 109–120.

Felsenstein, J. (1993). PHYLIP (Phylogeny Inference Package) version 3.5c. *Distributed by the author. Department of Genetics, University of Washington, Seattle.*

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

Frey, P., Whitt, S., and Tobin, J. (1994). A low-barrier hydrogen bond in the catalytic triad of serine proteases. *Science*, **264**(5167), 1927–1930.

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

Hastie, T., Tibshirani, R., and Friedman, J. H. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York.

Hedstrom, L., Szilagyi, L., and Rutter, W. (1992). Converting trypsin to chymotrypsin: the role of surface loops. *Science*, **255**(5049), 1249–1253.

Huang, B. and Schroeder, M. (2006). LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC Struct. Biol.*, **6**, 19.

Hubbard, S. and Thornton, J. (1993). A computer algorithm to calculate surface accessibility. Department of Biochemistry and Molecular Biology, University College, London.

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

Koh, K., Kim, S.-J., and Boyd, S. (2007). An interior-point method for large-scale L1-regularized logistic regression. *J. Mach. Learn. Res.*, **8**, 1519–1555.

Kolmogorov, V. and Zabih, R. (2002). What energy functions can be minimized via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **26**, 65–81.

Kraut, J. (1977). Serine proteases: structure and mechanism of catalysis. *Annu. Rev. Biochem.*, **46**, 331–358.

Lafferty, J., McCallum, A., and Pereira, F. (2001). Probabilistic models for segmenting and labeling sequence data. Proc. 18th International Conf. on Machine Learning, pages 282–289, San Francisco, CA. Morgan Kaufmann.

Lin, J. and Wong, S. K. M. (1990). A new directed divergence measure and its characterization. *Int. J. Gen. Syst.*, **17**(1), 73–81.

Meister, A. (1974). *The Enzymes*, volume 10. Academic Press, New York, 3rd edition.

Mihalek, I., Res, I., and Lichtarge, O. (2004). A family of evolution-entropy hybrid methods for ranking protein residues by importance. *J. Mol. Biol.*, **336**(5), 1265–1282.

Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**(4), 536–540.

Nakatsu, T., Kato, H., and Oda, J. (1998). Crystal structure of asparagine synthetase reveals a close evolutionary relationship to class II aminoacyl-tRNA synthetase. *Nat. Struct. Biol.*, **5**, 15–19.

Perona, J. J. and Craik, C. S. (1995). Structural basis of substrate specificity in the serine proteases. *Protein Sci.*, **4**, 337–360.

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

Sankararaman, S. and Sjölander, K. (2008). INTREPID–INformation-theoretic TREe traversal for Protein functional site IDentification. *Bioinformatics*, **24**(21), 2445–2452.

Stratton, J. R., Pelton, J. G., and Kirsch, J. F. (2001). A novel engineered subtilisin BPN' lacking a low-barrier hydrogen bond in the catalytic triad. *Biochemistry*, **40**, 10411–10416.

Tong, W., Williams, R. J., Wei, Y., Murga, L. F., Ko, J., and Ondrechen, M. J. (2008). Enhanced performance in prediction of protein active sites with THEMATICS and support vector machines. *Protein Sci.*, **17**(2), 333–341.

Vrallyay, E., Lengyel, Z., Grf, L., and Szilgyi, L. (1997). The role of disulfide bond C191-C220 in trypsin and chymotrypsin. *Biochem. Biophys. Res. Commun.*, **230**, 592–596.

Youn, E., Peters, B., Radivojac, P., and Mooney, S. D. (2007). Evaluation of features for catalytic residue prediction in novel folds. *Protein Sci.*, **16**(2), 216–226.

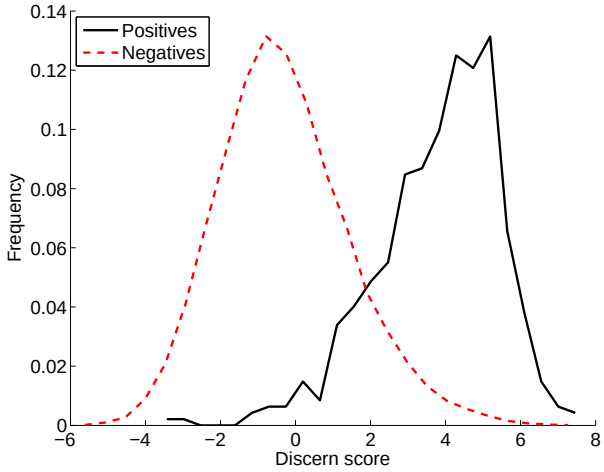

Figure S-1: **Distribution of DISCERN scores for residues listed by CATRES as catalytic (positives) and for the remaining residues (negatives)**. These scores were predicted for each residue in the enzymes belonging to the CATRES-FAM dataset. The scores were predicted when each enzyme was present in a test set during the cross-validation. Catalytic residues tend to have higher scores than the remaining residues.

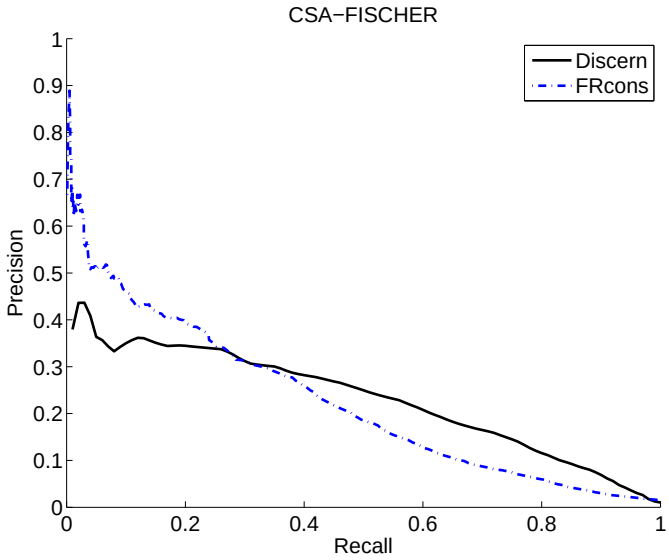

Figure S-2: **Comparison of Discern against FRcons on the CSA-Fischer dataset**. DISCERN shows improved precision relative to FRcons at recall values exceeding 30%. Analysis of the area under the precision-recall curve, termed PR-AUC, shows that the PR-AUC of FRcons is 0.1 compared to 0.23 for Discern. These results were obtained on a set of 423 enzymes from the Catalytic Site Atlas used by Fischer *et al.* in the evaluation of FRcons.

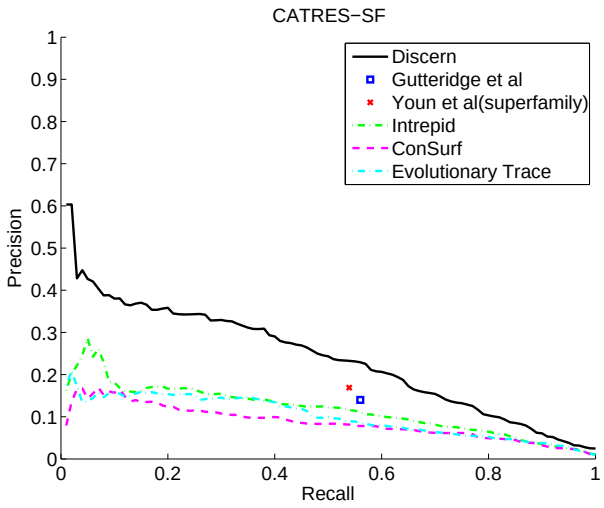

Figure S-3: **Results on the CATRES-SF benchmark dataset comparing** DISCERN **against Youn *et al.*, Gutteridge *et al.*, INTREPID and ConSurf**. DISCERN achieves a recall (the fraction of catalytic residues identified) of 65% at 17% precision (the fraction of predicted catalytic residues that are actually catalytic) on this dataset. Youn *et al.* results shown are from their reported performance on their SCOP superfamily-level dataset (i.e., similar in composition to CATRES-SF) on which they report a recall of 53.93% at a precision of 16.90%. Gutteridge *et al.* results are from their reported performance on the CATRES dataset, which includes sequences from the same SCOP family (i.e., an easier dataset), on which they report 56% recall at 14% precision. These results show that DISCERN attains an improvement in recall of 11% over Youn *et al.* superfamily-level results at the same precision, an improvement in recall of 16% over Gutteridge *et al.* results at 14% precision, and an improvement of 34% over INTREPID at 18% precision. ConSurf does not reach 18% precision on this dataset.

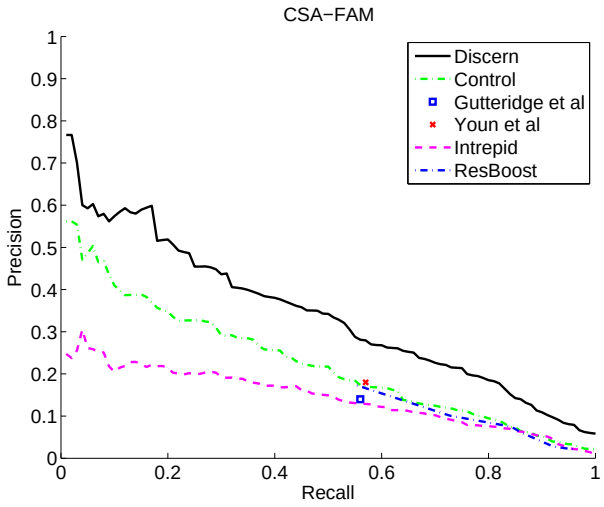

Figure S-4: **Results on the CSA-FAM benchmark dataset comparing** DISCERN **against Youn *et al.* Gutteridge *et al.*, INTREPID, ResBoost and a control**. CSA-FAM is filtered at the SCOP family level, and used to provide a comparison against datasets filtered at the same level of homology by Youn *et al.* and Gutteridge *et al.* Results from Youn *et al.* are from their reported performance on a dataset containing single representatives from SCOP families, for which they report 57.02% recall at 18.5% precision. Gutteridge *et al.* results are from their reported performance on the CATRES dataset, which includes sequences from the same SCOP family, on which they report 56% recall at 14% precision. ResBoost results are shown on this dataset for the range of recall values reported in (Alterovitz *et al.*, 2009). The control was trained identically to DISCERN but did not make use of INTREPID scoring functions or structural neighbors, and did not use L_1-regularization to enforce model sparsity (see Section S-5.3 and main text, Table 1). These results show that DISCERN attains an improvement in recall of 23% over the Youn *et al.* family-level results (achieving a recall of 75% at 18.5% precision relative to a recall of 57.02% reported by Youn *et al.* at the same precision), an improvement in recall of 26% over the Gutteridge *et al.* results at 14% precision, and an improvement of 39% over INTREPID at 18% precision. DISCERN also shows an improvement of 21% over the control at a precision of 18%.

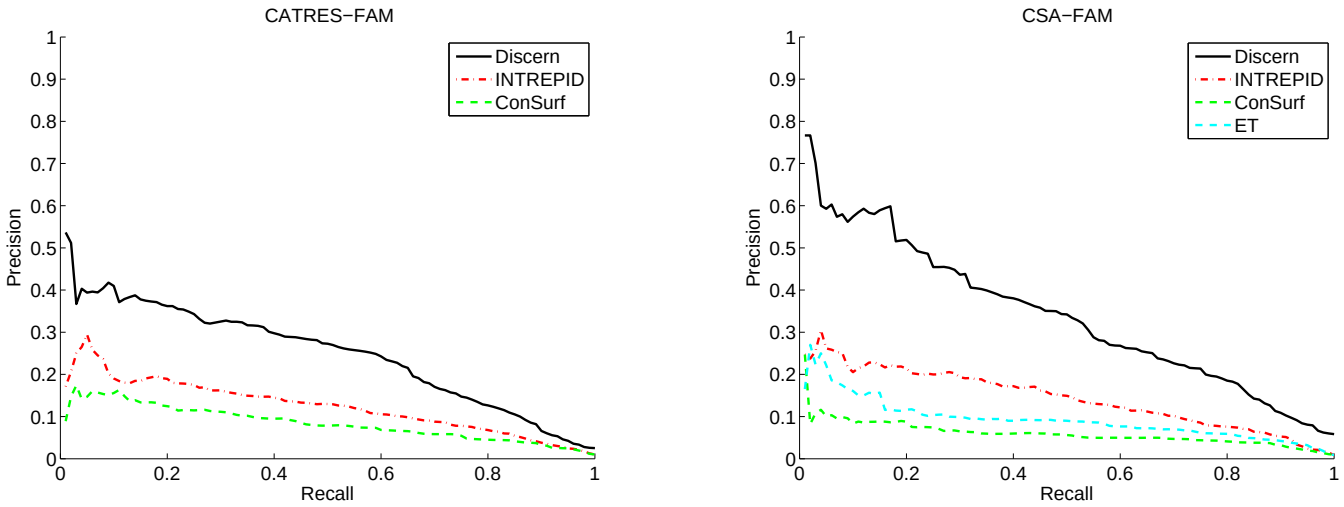

Figure S-5: **Comparison of** DISCERN **to methods that rely only on sequence conservation information.** Of methods shown here, only DISCERN makes use of structural information, giving it a significant advantage in these experiments. **Left:** On the CATRES-FAM dataset, at 18% precision, DISCERN has 69% recall and INTREPID has 19% recall while ConSurf does not attain a precision of 18%. At a lower precision of 10%, DISCERN obtained a recall of 87% compared to a recall of 64% and 35% by INTREPID and ConSurf respectively. At 50% recall, DISCERN, INTREPID, and ConSurf attain precisions of 27.27%, 12.96% and 7.86%. **Right:** On the CSA-FAM dataset, at a precision of 10%, DISCERN has 90% recall while INTREPID, ConSurf and Evolutionary Trace (ET) have 71%, 3% and 31% recall respectively. At 50% recall, DISCERN, INTREPID, and ConSurf attain precisions of 28.25%, 14.93% and 5.61%. ET results were obtained from the Baylor College of Medicine Evolutionary Trace server. ConSurf results were obtained from the ConSurf server DataBase (http://consurf.tau.ac.il).

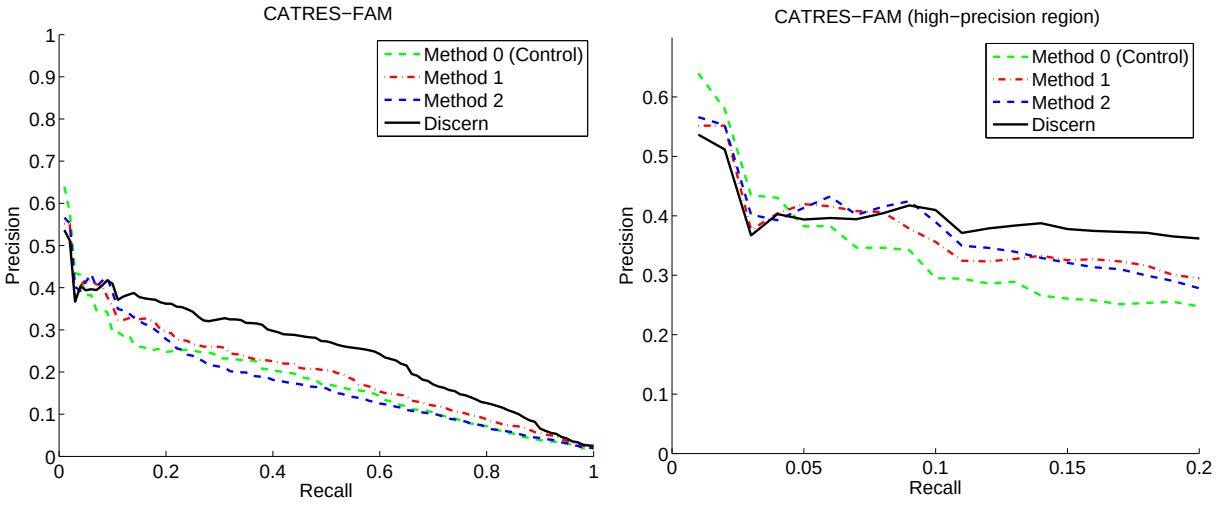

Figure S-6: **Precision-Recall curves comparing different logistic regression predictors on the CATRES-FAM dataset**. **Left:** Full Precision-recall curves, comparing DISCERN against logistic regressions trained using strict subsets of the DISCERN model. **Right:** Precision-recall curves for the high precision region (note that the axes are drawn to different scales). We evaluated several logistic regressions, varying the inclusion of features for structural neighbors and phylogenomic conservation scores from INTREPID and the use of L_1-regularization to control model complexity and prevent overfitting. The control uses non-phylogenetic conservation scores, while other methods use INTREPID. DISCERN is more accurate than the other variants over the range of recalls, except between a recall of 0.05 and 0.1 where Method 2 is most accurate. Further, since the control has very similar accuracies to Youn *et al.* and Gutteridge *et al.* (as shown in the main text), the improvement of DISCERN over these methods is significant and is unlikely to be an artifact of the dataset. See Section S-5.3 in this Supplementary Materials, and Table 1 in the main text for details on each variant and a comparison at fixed points of precision and recall.

Table S-1: **Top 20 residues predicted by different methods on *Escherichia coli* Asparagine Synthetase (PDB id:12as).** The three catalytic residues listed in CATRES (D46, R100 and Q116) are marked with *. Residues with a proposed functional role that are not listed in CATRES are marked with †. DISCERN detects all three catalytic residues in these top 20, INTREPID detects one, and ConSurf detects two of the three. Residues among these top 20 that are also described as functional in the literature but are not listed in CATRES include P35, K77, E120, D219, D235, E248, S251, R255, and I295. Of these 10 residues, seven are found among the top 20 for DISCERN, one is found by INTREPID and two are found by ConSurf. See Figure S-8 for the DISCERN predictions plotted onto the structure of asparagine synthetase. Figure S-10 shows an MSA for 12as and homologs. Refer to Section S-6.1 for a detailed analysis of these predictions.

DISCERN	Intrepid	Consurf
R214	W76	S72
D219†	W119	S111
D115	W318	S250
D235†	W117	S251†
K77†	H309	S298
D46 *	W221	I201
R100 *	H211	N233
E248†	M252	I291
E120†	M96	I295†
R255†	Q264	A74
Y218	M302	A98
H71	H110	V32
D118	Y218	V55
R78	Q297	V70
K75	N233	V114
Q116 *	P35†	V137
S251†	Q116 *	V256
S250	F197	I12
W119	H279	M96
D220	P288	M252

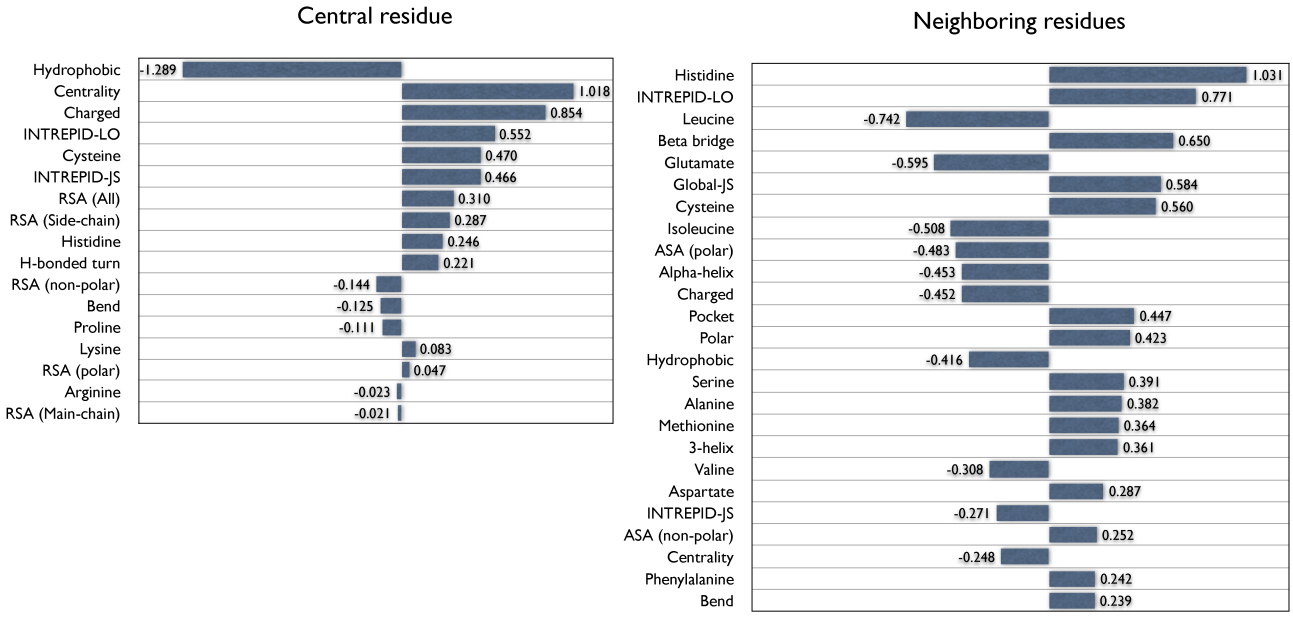

Figure S-7: **Features selected by** Discern**.** Shown here are weighted features based on fitting the Discern logistic regression to the entire CATRES-FAM dataset and displaying the features with the largest weights. Positive weights indicate positive correlation with putative catalytic residues; negative weights imply negative correlation. The magnitude of the weight is indicative of a feature's relative importance. **Left: Features computed at the residue of interest.** For features at the central residue, only 17 had non-zero weights. The feature with the largest weight (-1.289) is *hydrophobicity*; the negative weight is consistent with the observation that hydrophobic residues are rarely catalytic. The next highest-ranked feature is *residue centrality* with a weight of 1.018; high values for this feature indicate that a residue is located in the core of the enzyme 3D structure. *INTREPID-LO*, and *INTREPID-JS*, information-theoretic measures of the evolutionary conservation of a residue, jointly have a weight of 1.018, the same weight as centrality. Residue charge comes next with a weight of 0.854, followed by presence of a cysteine (0.470). *Relative solvent accessibility*, a measure of the fraction of a residue exposed to solvent averaged over all the atoms (RSA(All)) and over the side-chain atoms (RSA(Side-chain)), comes next with weights of 0.310 and 0.246 respectively. **Right: Features summed over residues that are nearby in the 3D structure.** The top 25 features with largest absolute weights are displayed. The feature with consistently large weights are the evolutionary conservation scores (*INTREPID-LO* and *GLOBAL-JS*). *INTREPID-LO* and *GLOBAL-JS* (a measure of sequence conservation across the family that does not use the phylogenetic tree) have a combined total weight of 1.255. The feature with the next largest weight (1.031) is the presence of a neighboring *histidine*. Two features with significant weights for residues in the structural neighborhood were *negatively* correlated with catalycity: presence of leucine (-0.742), glutamate (-0.595), and isoleucine (-0.508) and polar *absolute solvent accessibility* (ASA(polar)) (-0.483), i.e., solvent accessibility computed over all oxygens and nitrogens in the sidechain. ASA has large values for amino acids with large absolute surface areas, whereas RSA is normalized by the total surface area in the sidechain. Thus glycine could presumably have a large RSA under some circumstances, but will not have large ASA. The negative correlation of ASA at neighboring positions was unexpected; we hypothesize that this may be due to the function of a catalytic residue being inhibited by the presence of a nearby sidechain protruding into the cleft. The presence of a beta-bridge in the vicinity is indicative of a catalytic residue while an alpha-helix is negatively correlated. Note that the feature weights are summed over the structural neighbors.

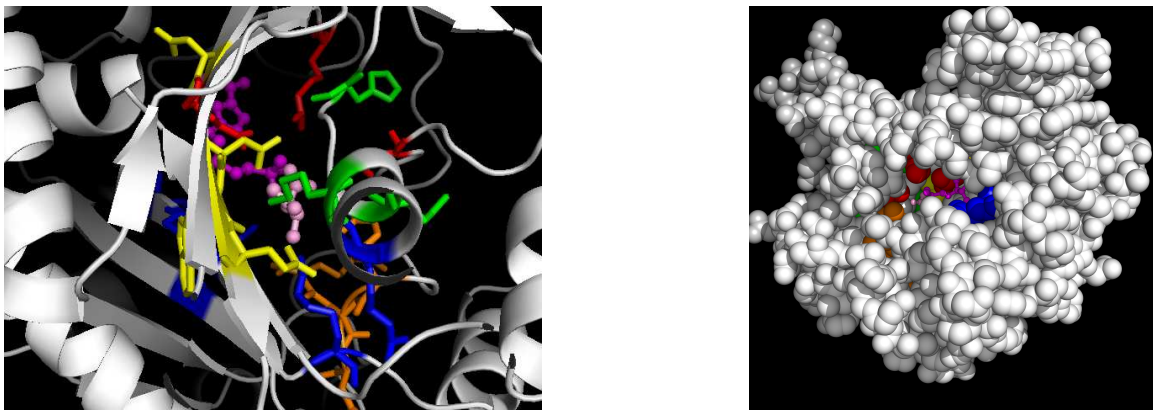

Figure S-8: **Top 20 residues predicted by** DISCERN **on *Escherichia coli* Asparagine Synthetase (PDB id:12as). Left:** Detailed view of the active site. Red indicates residues listed in CATRES (D46, R100, Q116). Green, yellow and orange indicate residues in motifs 1 (H71, K75 and K77), 2 (D115, Q116, D118, W119 and E120), and 3 (R214, Y218, D219, and D220) respectively. Other predicted residues are shown in blue. Also shown are the AMP and L-asparagine molecules. **Right:** DISCERN predictions shown in space-fill representation. See Table S-4 for a list of these residues.

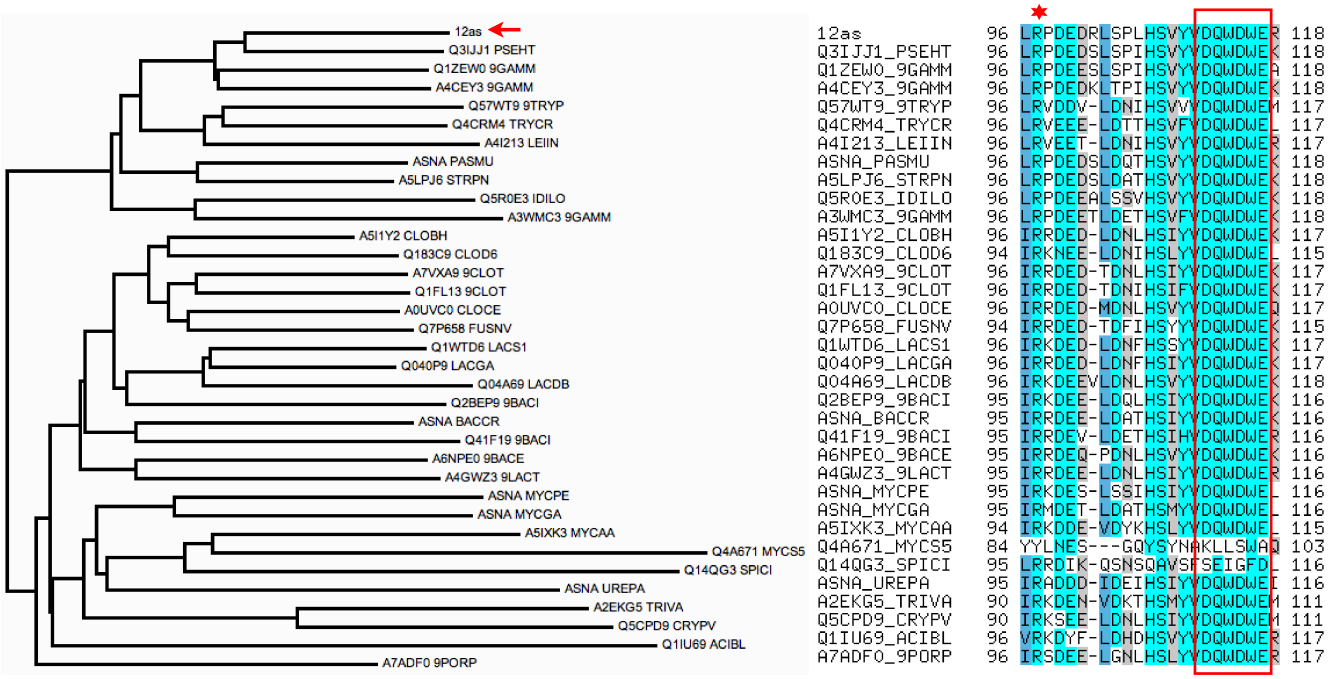

Figure S-9: **Tree and alignment of the homologs for *Escherichia coli* Asparagine Synthetase (PDB id:12as).** The seed sequence is indicated by an arrow. A neighbor-joining tree and multiple sequence alignment were derived by making the original alignment non-redundant at 70% identity relative to the seed. The positions in the seed sequence correspond to the residue number in PDB minus 3, e.g., the arginine at position 97 corresponds to R100 in the PDB record. R100 is marked with a star because it is listed as catalytic in CATRES. Note that not all sequences contain an arginine at this position. Positions in motif 2 (D115, Q116, D118, W119 and E120) have been boxed. The branch lengths of Q4A671_MYCS5 and Q14QG3_SPICI have been truncated from their original lengths of 1.089 and 1.181 respectively to a value of 0.5 for better visualization.

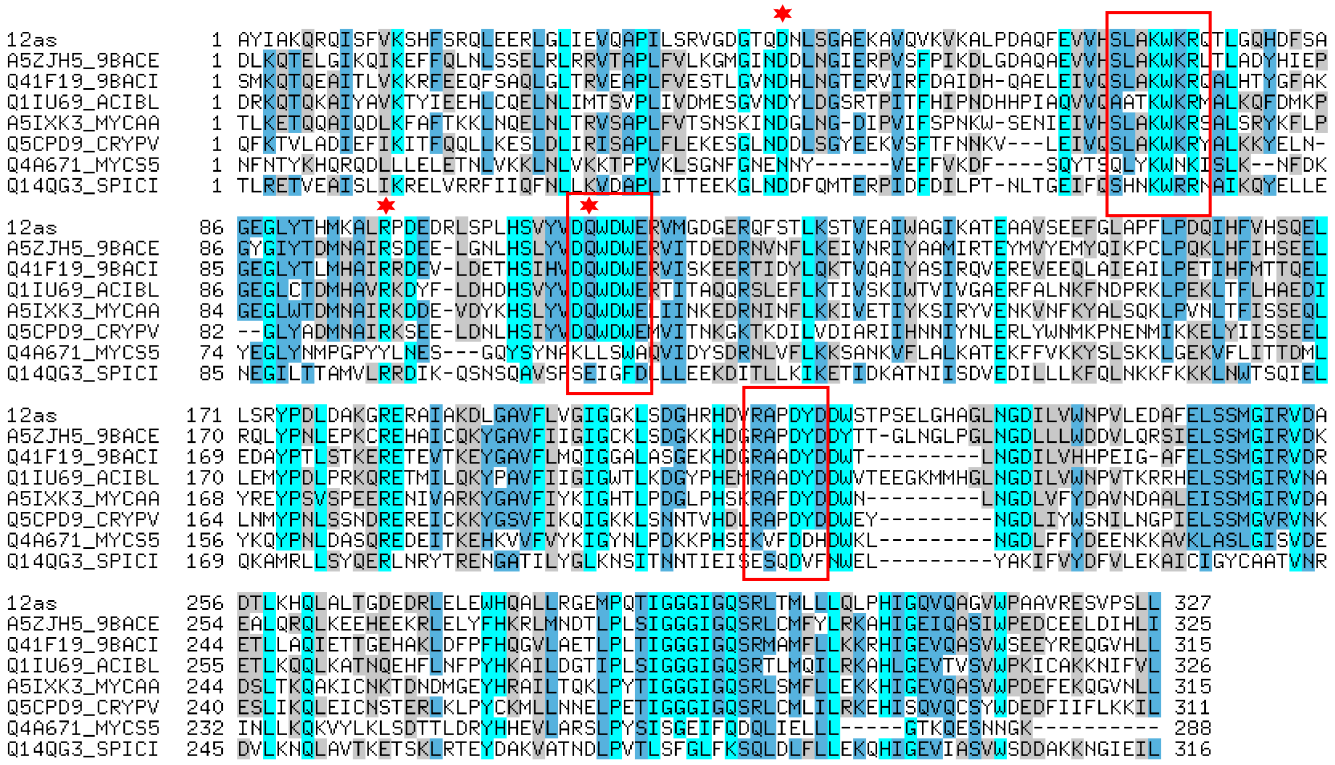

Figure S-10: **Multiple sequence alignment of selected homologs for *Escherichia coli* Asparagine Synthetase (PDB id:12as).** The displayed alignment was derived by making the original alignment non-redundant at 50% identity. Residues listed as catalytic in CATRES (D46, R100 and Q116) are marked with a star while positions that form motifs based on their DISCERN scores have been boxed. See Table S-1 for the list of predicted residues. Note that none of the catalytic residues are perfectly conserved in this dataset, reflecting a limitation of the use of simple global conservation scores.

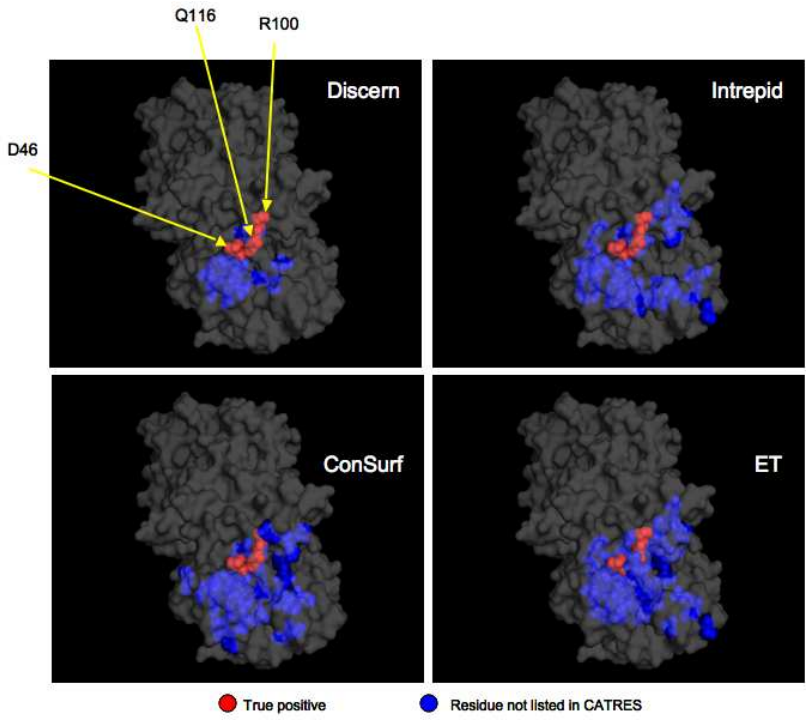

Figure S-11: Comparison of DISCERN, INTREPID, ConSurf and Evolutionary Trace (ET) predictions on *Escherichia coli* Asparagine Synthetase (PDB id:12as): The predictions from all methods are shown at a recall of 100%; i.e., when all the catalytic residues listed in CATRES have been selected. DISCERN predicts the three catalytic residues listed in CATRES (D46, R100, and Q116) and 13 additional residues (R214, D115, Y218, D219, D118, E120, H71, K75, K77, R78, D235, E248 and R255) of which seven have been proposed to play functional roles on the basis of structural studies Nakatsu *et al.* (1998). In contrast, INTREPID, ConSurf and ET require a total of 33, 44, and 50 residues respectively to achieve perfect recall. Note that the catalytic residues predicted by the methods are sometimes visually obscured by the false positives. See Table S-1 in for more details on these predictions.

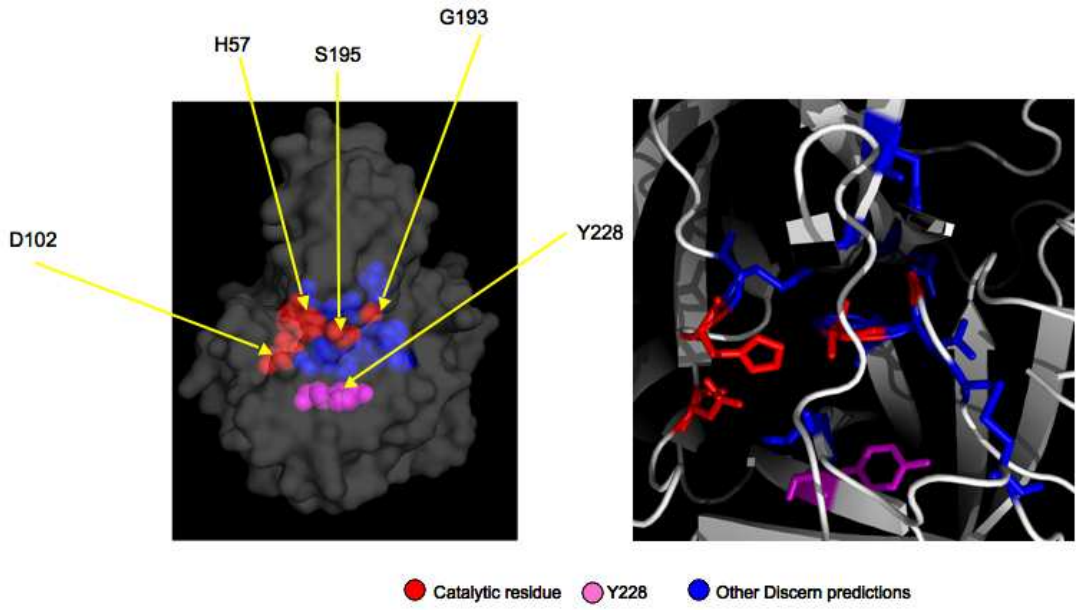

Figure S-12: DISCERN **predictions on *Bovine* α-Chymotrypsin (PDB id:1acb).** Red indicates the catalytic residues (H57, D102, S195, G193). Fuchsia indicates Y228. Blue indicates all the other DISCERN predictions (D194, C191, C42, C58, Q30, C220, S214, G197, H40 and G196). **Left:** The top 15 DISCERN predictions. DISCERN predicts the catalytic triad H57, D102, and S195, with ranks 6, 4, and 1, respectively. The catalytic glycine, G193, is predicted with rank 13. Y228 (DISCERN rank 10) is found in the S1 specificity pocket, but its functional role is unknown. The roles of Y228 and other residues (D194, C191, C42, C58, Q30, C220, S214, G197, H40 and G196) are described in Table S-2 in the Supplementary Materials. **Right:** Closeup of the active site.

Table S-2: **Top 15 residues predicted by** DISCERN **on *Bovine* α-Chymotrypsin (PDB id:1acb, E.C. number: 3.4.121.1).** The catalytic triad H57, D102 and S195, and the catalytic glycine G193, are marked with *. Residues with a proposed or known functional role are marked with †. DISCERN detects all three catalytic residues in these top 15. Refer to Section 2.1 for a detailed analysis of these predictions.

Rank	Residue	Score	Role
1	S195 *	5.725	Catalysis (Perona and Craik, 1995; Carter and Wells, 1988; Stratton *et al.*, 2001)
2	D194†	5.050	S1 pocket, Proenzyme activation (Birktoft *et al.*, 1976)
3	C191†	4.784	S1 pocket, Disulfide bridge (with C220) (Vrallyay *et al.*, 1997)
4	D102 *	4.646	Catalysis (Perona and Craik, 1995; Carter and Wells, 1988; Stratton *et al.*, 2001; Frey *et al.*, 1994)
5	C42†	4.529	Disulfide bridge (with C58), binding (Hedstrom *et al.*, 1992)
6	H57 *	4.316	Catalysis (Perona and Craik, 1995; Carter and Wells, 1988; Stratton *et al.*, 2001)
7	C58†	3.748	Disulfide bridge (with C42), binding (Hedstrom *et al.*, 1992)
8	Q30†	3.586	Proenzyme activation (Birktoft *et al.*, 1976)
9	C220†	3.551	S1 pocket, Disulfide bridge (with C191) (Vrallyay *et al.*, 1997)
10	Y228	3.439	S1 pocket, role unknown
11	S214†	3.426	S1 pocket (Hedstrom *et al.*, 1992)
12	G197†	3.370	β-bulge
13	G193 *	3.358	Catalysis (Kraut, 1977)
14	H40†	3.283	Proenzyme activation (Berna *et al.*, 1997; Birktoft *et al.*, 1976)
15	G196†	3.067	β-bulge

Table S-3: **Comparison of** DISCERN **and the CRF.** Precision_{50} reports the precision at 50% recall, and Recall_{18} reports the recall at 18% precision (these precision and recall points were selected to allow direct comparison to the results reported in Youn *et al.* (2007)). DISCERN and CRF results are statistically indistinguishable.

Method	CATRES-FAM	
	Precision_{50}	Recall_{18}
DISCERN	27.3%	69%
CRF	26.9%	69%

Table S-4: **Comparison of** DISCERN**, INTREPID and ConSurf.** Precision_{50} reports the precision at 50% recall, and Recall_{10} reports the recall at 10% precision (ConSurf does not achieve a precision of 18% on CATRES-FAM).

Method	CATRES-FAM		CSA-FAM	
	Precision_{50}	Recall_{10}	Precision_{50}	Recall_{10}
DISCERN	27.3%	86%	28.3%	90%
INTREPID	13.0%	64%	14.9%	70%
ConSurf	7.9%	35%	5.6%	6%

Table S-5: **Features evaluated for catalytic residue prediction:** This set of features are evaluated at a residue and each of its ten structural neighbors resulting in $48 \times 11 = 528$ features. RSA and ASA refer to the relative and absolute solvent accessibility respectively. Refer to Section S-2 for detailed descriptions.

Type of feature	Description
Sequence conservation features	INTREPID-JS, INTREPID-LO, GLOBAL-JS
Amino acid properties	{Charged, Polar, Hydrophobic}, {20 amino acid sidechains}
Structure-based features	B-factor, Centrality, Secondary structure element (Alpha helix, Beta bridge, Strand, 3-helix, pi-helix, H-bonded turn, Bend) RSA (All atoms, Side chain, Main chain, Non polar, Polar), ASA (All atoms, Side chain, Main chain, Non polar, Polar), Presence in each of three largest pockets