1 Supplementary: Materials and Methods

1.1 rvET Algorithm

The real value ET (rvET) [4] scores the evolutionary importance of residues (ET rank) based on the multiple sequence alignment columns in context of the phylogenetic tree. Shannon Entropy is calculated for the entire alignment, and then recalculated for all the subgroups of the alignment selected by the phylogenetic tree. The rank $\rho(i)$ of residue i is calculated as follows:

$$\rho(i) = 1 + \sum_{n=1}^{N-1} \frac{1}{n} \sum_{g=1}^{n} \{- \sum_{a=1}^{20} f_a^g(i) ln f_a^g(i)\} \tag{1}$$

where $f_a^g(i)$ is the frequency of the amino acid of type a within the sub-alignment of group g of the n sub-alignments. The number of possible nodes in the evolutionary tree is $N-1$ where N is the number of sequences in the alignment. The nodes in the phylogenetic tree are numbered in the order of increasing distance from the root.

1.2 Testing LexA Mutation Sensitivity for DNA Damage

All LexA mutants were generated using Qiagen Site Directed Mutagenesis XLII Kit (Stratagene) following manufacturers protocols. Prior to transformation in the expression strain, the incorporation of the mutation(s) was verified by sequencing. The mutant strains were streaked on LBChloramphenicol [25 ug/ml] plates. Individual colonies were picked and grown in LB broth overnight with aeration at 37°C for up to 16 hours. Next morning, fresh subcultures were placed and grown till the OD_{600} reached 0.3-0.4. All strains were normalized to an OD_{600} of 0.2 by dilution. Up to 50ul of the appropriate serial dilutions were then plated on LB-Chloramphenicol [25 ug/ml]$_{final}$ final plates. Using XL-1500 UV Crosslinker [Spectrolinker], the plates were exposed to 16 J/cm^2 of UV. Control plates with no exposure to UV were plated at the same time. All plates were incubated overnight at 37°C in dark [covered with foil] and colonies were counted next morning. The assay was repeated for all strains in triplicates.

1.3 Coverage and accuracy

Coverage and accuracy curves were found by first sorting the predictions for the test set in descending order of the confidence measure (z-score) and then calculating accuracy and coverage as follows $Accuracy = \frac{Total\ True\ Predictions}{Total\ Predictions}$ and $Coverage = \frac{Total\ Predictions}{Size\ of\ Test\ Set}$.

2 Supplementary: Results

2.1 Correlation: MSA size held constant

In order to check the contribution of the number of sequences in the simulation, the sequence simulation was repeated. In the new simulation, we held the number of sequences constant at each iteration where only one-fourth of the alignment was utilized for the rvET analysis. The new analysis showed a strong correlation between smoothing function and overlap measure. The average correlation did decrease by a small amount (-0.65 to -0.50). This shows the majority of correlation between smoothing function and functional site overlap was due to selection of sequences.

3 Supplementary: Pseudocode

3.1 Smoothing Function Simulation

1. **Variables:**

   (a) **MSA** = multiple sequence alignment of protein

   (b) **N** = number of sequences in MSA of protein

   (c) **pdb_structure** = PDB structure for protein

   (d) **epitope** = protein's known functional site

2. **Subroutines:**

   (a) convert_structure_to_Laplacian (pdb_structure) → returns Laplacian matrix based on PDB information (Equation 1 & 2)

   (b) remove_sequences (MSA, random_number) → returns new MSA where random_number of proteins were removed from default MSA

   (c) calculate_smoothing_function (rvET_ranks) → returns value of the smoothing function (Equation 3)

   (d) calculate_z_overlap (rvET_ranks, epitope) → returns value of the functional overlap z-score (Equation 4 & 5)

   (e) Bin_analysis_by_xTLx (X, Y, ave_X_for_bins, Y_for_bins) → returns average value of a subset of X when they fall in bin set by Y

   (f) calculate_correlation_of_binned_analysis (ave_X_for_bins, Y_for_bins) returns correlation of bins

Algorithm 3.1: MAIN CODE($MSA, N, pdb\_structure, epitope$)

```
$Laplacian\_matrix = convert\_structure\_to\_Laplacian(pdb\_structure)$
for each $30,000$ $Iterations$ $(i)$
  do
    $random\_number = 25 + random(N - 25)$
    $new\_set\_of\_proteins = remove\_sequences(MSA, random\_number)$
    $x = rvET\_analysis(new\_set\_of\_proteins)$
    $xTLx[i] = calculate\_smoothing\_function(Laplacian\_matrix, x)$
    $z\_o[i] = calculate\_z\_overlap(x, epitope)$
$Bin\_analysis\_by\_xTLx(xTLx, z\_o, average\_z\_o\_for\_bins, xTLx\_for\_bins)$
$Correlation = calculate\_correlation\_of\_binned\_analysis(average\_z\_o\_for\_bins, xTLx\_for\_bins)$
return $(Correlation)$
```

3.2 piET algorithm

1. **Variables:**
   (a) **MSA** = multiple sequence alignment of protein
   (b) **N** = number of sequence in MSA of protein
   (c) **pdb_structure** = PDB structure for protein

2. **Subroutines:**
   (a) calculate_distance_matrix (MSA, BLOSUM62) → returns distance matrix for MSA
   (b) calculate_adjacency_matrix (pdb_structure) → returns Adjacency matrix based on PDB information (Equation 1)
   (c) UPGMA (Matrix) → returns phylogenetic tree using UPGMA

Algorithm 3.2: MAIN CODE($MSA, N, pdb\_structure$)

```
$Distance\_Matrix = calculate\_distance\_matrix(MSA, BLOSUM62)$
$Tree = calculate\_UPGMA\_tree(Distance\_Matrix)$
for each $Divergence(n) \in Tree$
  do
    for each $Group(g) \in Divergence$
      do
        for each $PossiblePair \in MSA$
          do $\{rho(ij) += calculate\_entropy(MSA, n, g, ij)$
$A = calculate\_adjacency\_matrix(pdb\_structure)$
for each $Residue\ (i) \in Protein$
  do
    for each $Residue\ (j) \in Protein$
      do
        $pi(Residue\ i) = A(i,j) * rho(ij)$
        $number\_contacts(Residue\ i) ++$
for each $Residue\ (i) \in Protein$
  do $pi(residue\ i)/number\_contacts$
return $(pi)$
```

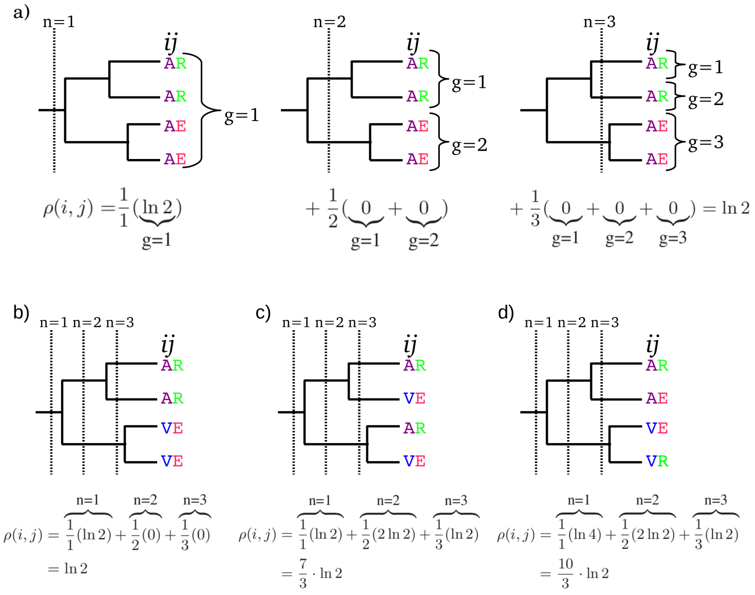
a)
n=1
ij
AR
AR
AE
AE
g=1
$\rho(i,j) = \frac{1}{1}(\underbrace{\ln 2}_{g=1})$
n=2
ij
AR
AR
AE
AE
g=1
g=2
$+\frac{1}{2}(\underbrace{0}_{g=1} + \underbrace{0}_{g=2})$
n=3
ij
AR
AR
AE
AE
g=1
g=2
g=3
$+\frac{1}{3}(\underbrace{0}_{g=1} + \underbrace{0}_{g=2} + \underbrace{0}_{g=3}) = \ln 2$
b)
n=1 n=2 n=3
ij
AR
AR
VE
VE
$\rho(i,j) = \overbrace{\frac{1}{1}(\ln 2)}^{n=1} + \overbrace{\frac{1}{2}(0)}^{n=2} + \overbrace{\frac{1}{3}(0)}^{n=3}$
$= \ln 2$
c)
n=1 n=2 n=3
ij
AR
VE
AR
VE
$\rho(i,j) = \overbrace{\frac{1}{1}(\ln 2)}^{n=1} + \overbrace{\frac{1}{2}(2\ln 2)}^{n=2} + \overbrace{\frac{1}{3}(\ln 2)}^{n=3}$
$= \frac{7}{3} \cdot \ln 2$
d)
n=1 n=2 n=3
ij
AR
AE
VE
VR
$\rho(i,j) = \overbrace{\frac{1}{1}(\ln 4)}^{n=1} + \overbrace{\frac{1}{2}(2\ln 2)}^{n=2} + \overbrace{\frac{1}{3}(\ln 2)}^{n=3}$
$= \frac{10}{3} \cdot \ln 2$

Figure 1: The shared evolutionary importance of a pair of residue neighbors ($\rho(i,j)$) can be measured in the context of the phylogenetic tree. The algorithm applies the standard ET equation to pairs of residues deemed in contact through structural information. The formula first measures the variation for positions i and j in the entire alignment utilizing the entropy term, $\{-\sum_{ab=1}^{400} f_{ab}(i,j) \ln f_{ab}(i,j)\}$, where ab represents the occurrence of one of the 400 possible pairs (AR, AE, etc.). Then the alignment is repeatedly broken down into sets of sub-alignments (labeled n) based on the divergences within the phylogenetic tree. The entropy term is then calculated for these sub-alignments (labeled g). In (a) we have an example of the calculation $\rho(i,j)$. The first entropy term $\{-\sum_{ab=1}^{400} f_{ab}(i,j) \ln f_{ab}(i,j)\} = \ln 2$ for the entire alignment ($n = 1$) since our residue pairs have two events (AR and AE) that occur exactly twice. At $n = 1$ we have variation for residue pair $i : j$ but after the first divergence the sub-alignments are invariant for those residues $i : j$ for the sub-alignments and the entropy term is equal to zero. (b) shows an example where a residue-residue interaction has the same evolutionary score as example (a) even though it appears to have more variation in the individual columns but in the context of the pattern it is equivalent. Example (c) has the same residue-residue pattern but is positioned differently in the subtrees. $\rho(i,j)$ ends up being higher which implies that case (b) would be more evolutionarily important. The last case (d) would be the least important where no event appears more than once.

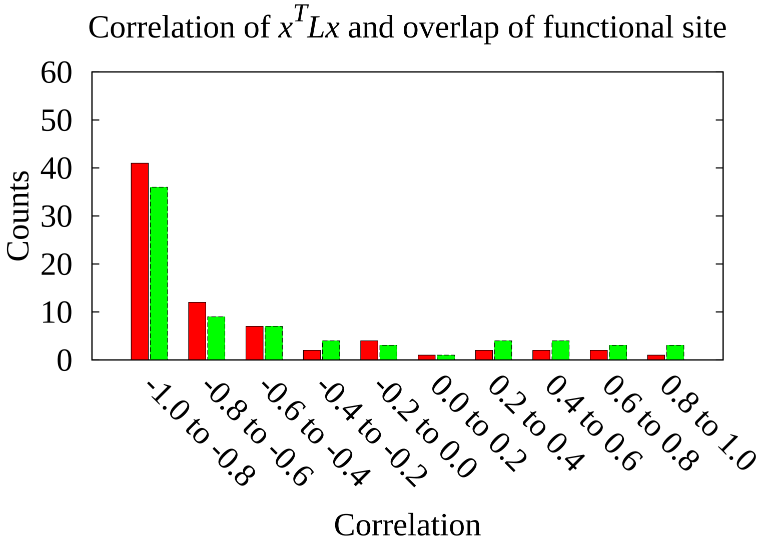

Figure 2: (The orange bars are from the original simulation where number of sequences varied (average correlation equal to -0.65). The green bars are the new simulation where the number of sequences contributing was set to one-fourth of the total sequences in alignment. The average correlation over the 74 proteins was -0.50. The value of the smoothing function $x^T Lx$ for the random sets of sequences correlates with functional site overlap.

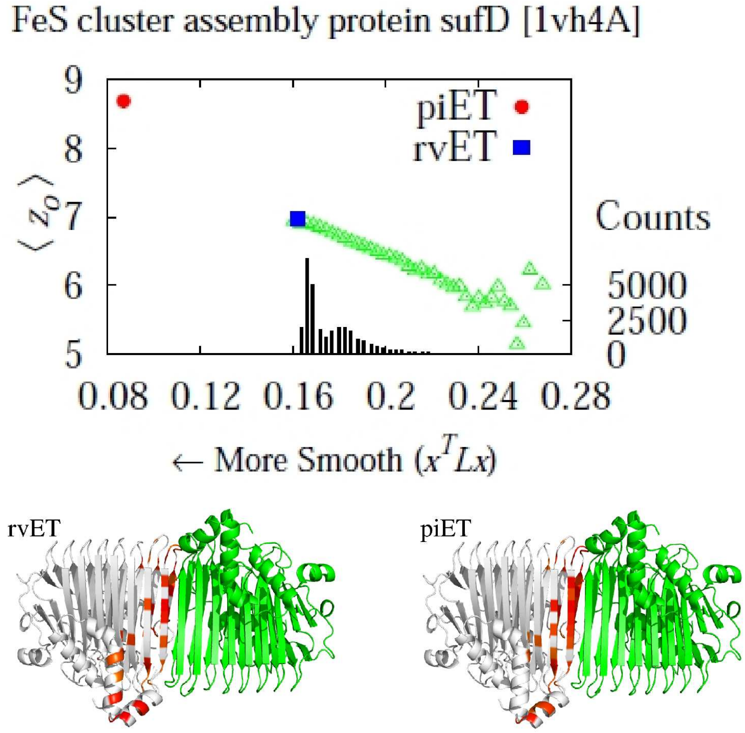

Figure 3: The sufD structure [PDBID 1vh4] with the homodimer interface also shows a slight improvement (top 10% for the piET and rvET are shown). Though the original method rvET picks the interface for this homodimer well, piET shows a statistically significant improvement

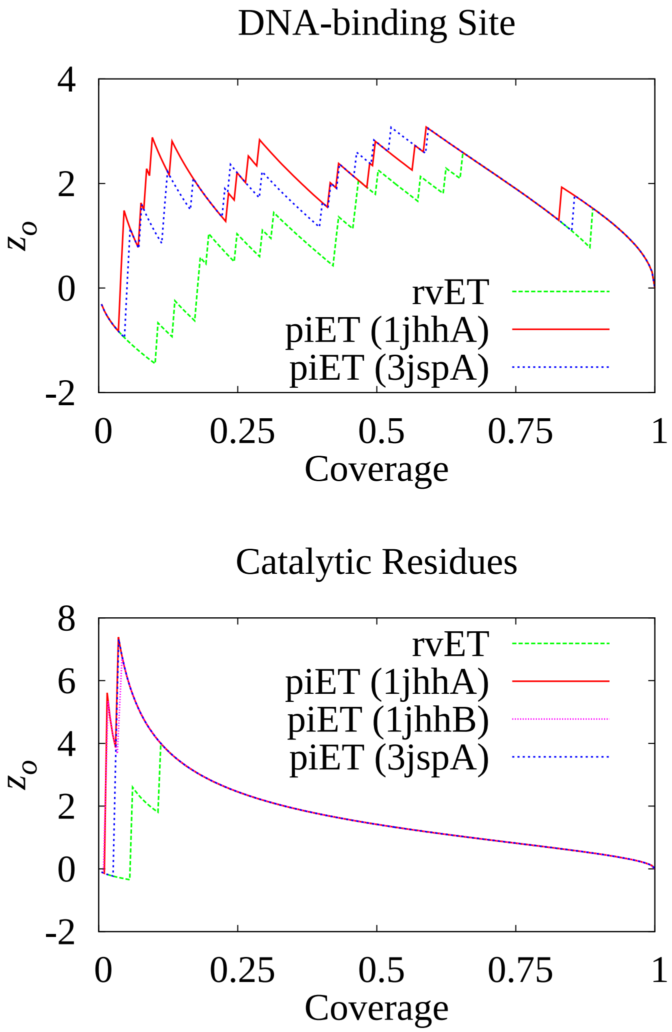

Figure 4: The statistical significance of the prediction is shown for the DNA-bound site and the residues responsible for the catalysis.

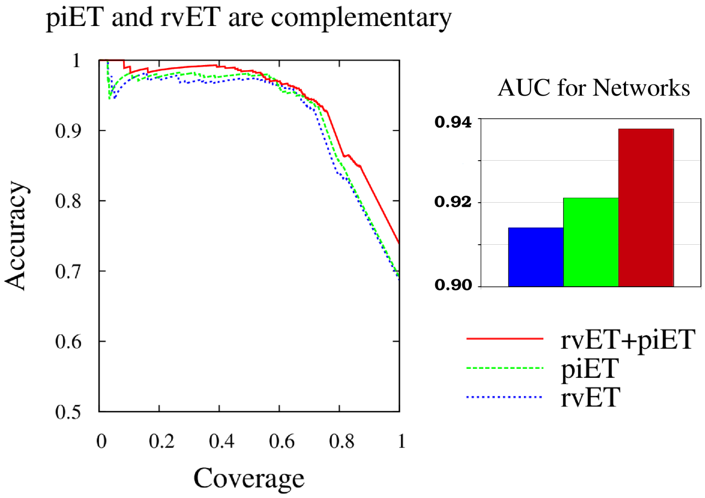

Figure 5: Though the network based on piET has an advantage over rvET, the performance of the combined network based on both measures of evolutionary importance indicate that they provide complementary forms of functional information. The accuracy vs. coverage curves are shown for three methods: rvET, piET and a combined network. The curve for each algorithm is the cumulative accuracy as coverage is increased in order of decreasing prediction confidence. The combined networks features a longer region of 100% accuracy. The test set is made of predictions of 1070 Structural Genomics enzymes with existing annotations.

References

[1] S Altschul, T Madden, A Schaffer, J Zhang, Z Zhang, W Miller, and D Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*, 25:3389–3402, 1997.

[2] U. Hobohm, R. Scharf, M. andchneider, and Sander C. Selection of representative protein data sets. *Protein Sci*, 1:409417, 1992.

[3] R Laskowski, V Chistyakov, and J Thornton. Pdbsum more: new summaries and analyses of the known 3d structures of proteins and nucleic acids. *Nucleic Acids Res*, 33:D266–D268, 2005.

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

[5] R Matthew Ward, Serkan Erdin, T A Tran, D M Kristensen, A M Lisewski, Serkan Erdin, and Olivier Lichtarge. De-orphaning the structural proteome through reciprocal comparison of evolutionarily important structural features. *PLoS ONE*, 7;3(5):e2136, 2008.

[6] A. D. Wilkins, R. Lua, S. Erdin, R. M. Ward, and Lichtarge O. Sequence and structure continuity of evolutionary importance improves protein functional site discovery and annotation. *Protein Science.*, 19(7):1296–1311, 2010.