SUPPLEMENTARY MATERIAL

S-1 INTREPID VARIANTS

We define different INTREPID variants based on the positional conservation score \( \text{cons}(S, x) \) which is used to compute the importance score in Equation S-1.

\[
IM_P(x) = \max_x \text{cons}(S, x) - \text{cons}(S)
\]

(S-1)

- INTREPID-JS is the method referred as INTREPID in the main text. In this variant, \( \text{cons}(S, x) \) is the Jensen-Shannon divergence between the amino acid distribution and the background amino acid distribution derived from the Blocks database (Henikoff and Henikoff, 1992) with prior weight = \( \frac{1}{2} \) as in Capra and Singh (2007).
- INTREPID-LO where \( \text{cons}(S, x) \) is the log probability of the most frequent amino acid at position \( x \) within subtree \( S \).
- INTREPID-RE where \( \text{cons}(S, x) \) is computed as the relative entropy between the amino acid distribution within subtree \( S \) of position \( x \) and a background distribution derived from the Blocks database alignments (Henikoff and Henikoff, 1992).

We compare these INTREPID variants to their respective global conservation baselines which we term Global-JS (Capra and Singh, 2007), Global-LO, and Global-RE (the scoring function introduced by Wang and Samudrala (2006)) respectively.

![Graph](image1)

Fig. S-1. Results for catalytic residue prediction on the CSA-100 dataset using rank-based scores. (Left) ROC curve comparing the variants of INTREPID (INTREPID-JS referred to as INTREPID in the main text, INTREPID-RE and INTREPID-LO) and the respective variants of global conservation (Global-JS, Global-RE, and Global-LO). INTREPID-JS and INTREPID-LO are significantly more accurate than Global-JS and Global-LO respectively while INTREPID-RE performs worse than Global-RE. (Right) ROC curves comparing INTREPID, Global-JS, BCMET and ConSurf. The ROC curve shows INTREPID-JS to have the highest sensitivity over the range of specificity followed by Global-JS. BCMET performs better as the specificity decreases.

![Graph](image2)

Fig. S-2. Comparison of the INTREPID variants and the respective global conservation baselines based on the normalized scores on the CSA-100 dataset.
Table S-1. Statistics comparing the variants of INTREPID and the baseline global conservation on the CSA-100 dataset. In the top panel, the ranks of the residues were used; in the bottom panel, the normalized scores are shown. Sensitivity is measured at specificities of 80, 90, and 95% respectively and recall at 10% precision. \( AUC_{x} \), \( x = 80, 90, 95 \) refers to the area under the ROC curve when specificity is at least \( x \% \); \( AUC \) is the area under the entire curve. The p-value refers to the Wilcoxon signed rank p-values between the AUC of the INTREPID variant and the respective global conservation baseline. INTREPID-JS improves over other methods on all metrics. The INTREPID variants, INTREPID-JS and INTREPID-LO improve over their respective baseline while INTREPID-RE performs worse. The confidence intervals on these statistics are reported in Supplementary Table S-2.

| Residue ranks | INTREPID-LO | INTREPID-JS | INTREPID-RE | Global-LO | Global-JS | Global-RE |
|---------------|-------------|-------------|-------------|-----------|-----------|-----------|
| **Sensitivity** | 59.55       | 70.06       | 57.32       | 57.64     | 64.33     | 64.97     |
| Sensitivity90  | 75.80       | 85.03       | 77.39       | 72.29     | 78.66     | 79.94     |
| Sensitivity90  | 90.76       | 93.95       | 91.72       | 85.67     | 90.13     | 90.45     |
| Recall10      | 84.00       | 91.00       | 86.00       | 80.00     | 86.00     | 86.00     |
| **AUC**       | 0.923       | 0.944       | 0.926       | 0.907     | 0.924     | 0.926     |
| AUC95         | 0.019       | 0.024       | 0.018       | 0.019     | 0.022     | 0.021     |
| AUC90         | 0.054       | 0.063       | 0.051       | 0.051     | 0.058     | 0.058     |
| AUC80         | 0.139       | 0.154       | 0.139       | 0.131     | 0.145     | 0.145     |
| p-value       | –           | –           | –           | –         | 3.89 \times 10^{-18} | 3.89 \times 10^{-18} | 0.17 |

Normalized scores

| Residue ranks | Sensitivity95 | Sensitivity90 | Sensitivity80 | Recall10 | **AUC** | AUC95 | AUC90 | AUC80 | p-value |
|---------------|---------------|---------------|---------------|----------|---------|-------|-------|-------|---------|
| **Sensitivity** | 51.91         | 67.83         | 55.10         | 47.77    | 58.28   | 61.15 |
| Sensitivity90  | 73.89         | 85.03         | 75.48         | 69.43    | 76.75   | 80.25 |
| Sensitivity80  | 90.13         | 92.99         | 92.68         | 86.31    | 89.81   | 89.49 |
| Recall10      | 83.00         | 92.00         | 84.00         | 78.00    | 83.00   | 85.00 |
| **AUC**       | 0.911         | 0.935         | 0.918         | 0.891    | 0.910   | 0.916 |
| AUC95         | 0.014         | 0.022         | 0.017         | 0.014    | 0.018   | 0.019 |
| AUC90         | 0.046         | 0.060         | 0.050         | 0.045    | 0.053   | 0.055 |
| AUC80         | 0.130         | 0.149         | 0.136         | 0.124    | 0.137   | 0.141 |
| p-value       | –             | –             | –             | 3.89 \times 10^{-18} | 3.89 \times 10^{-18} | 0.33 |

Fig. S-3. Comparison of INTREPID, Global-JS, BCMET and ConSurf based on the normalized scores on the CSA-100 dataset. The methods have \( AUC \)s of 0.935, 0.910, 0.919, and 0.884 respectively and \( AUC_{90} \) of 0.060, 0.053, 0.036, and 0.048 respectively.

S-2 EXPERIMENTS

S-2.1 Greater evolutionary divergence improves the accuracies of INTREPID and Global-JS

We have shown in Section 3.1.3 of the main paper that greater evolutionary divergence improves the accuracy of INTREPID. Global-JS also improves with the inclusion of additional homologs but appears to be somewhat less robust to sequence divergence (Figure S-4). For instance, at 90% specificity, the sensitivity of Global-JS prediction is 42% on the most restricted alignment (removing sequences with less than 25% identity to the seed) but increases to about 79% on the unrestricted alignment.

Our results are in agreement with previous studies (Panchenko et al., 2004; Aloy et al., 2001; Landgraf et al., 2001). In Landgraf et al. (2001), the recall of their scoring functions improved when the E-value cutoff for homolog inclusion was reduced from \( 10^{-56} \) to \( 10^{-20} \) while (Aloy et al., 2001) observed a considerable improvement in accuracy of their method when their alignments had sequence identity less than 30%. Similar results were also reported by Panchenko et al. (2004), where the performance doubled on alignments with average sequence identity of 20% relative to those with average identity of around 45%. An important difference is that our alignments are highly divergent. In the experiments reported by Panchenko et al. (2004), the least divergent dataset had a minimum percent identity of 25% to the seed. For
Sensitivity is measured at specificities of 10%.

Table S-2. Confidence Intervals for statistics comparing the variants of INTREPID and the baseline global conservation on the CSA-100 dataset. In the top panel, the ranks of the residues were used while in the bottom panel, the normalized scores were used. Sensitivity is measured at specificities of 80, 90, and 95% respectively and recall at 10% precision. $AUC_s, x = 80, 90, 95$ refers to the area under the ROC curve when specificity is at least x%; $AUC_C$ is the area under the entire curve. The 95% confidence intervals are computed from 200 bootstrap replicates.

Table S-3. Confidence Intervals for statistics comparing the different algorithms on the CSA-100 dataset. BCMET refers to the Evolutionary Trace server from Baylor College of Medicine. In the top panel, the ranks of the residues were used while in the bottom panel, the normalized scores were used. Sensitivity is measured at specificities of 80, 90, and 95% respectively and recall at 10% precision. $AUC_s, x = 80, 90, 95$ refers to the area under the ROC curve when specificity is at least x%; $AUC_C$ is the area under the entire curve. The 95% confidence interval are computed from 200 bootstrap replicates.

catalytic residue prediction, we observe that it is beneficial to use highly divergent alignments with minimum percent identities extending as low as 10%.

S-2.2 The accuracy of INTREPID decreases with distance from the seed

To test the effectiveness of INTREPID prediction for proteins not used as seeds for selecting and aligning homologs, we took three families of enzymes from CSA containing at least two members each in the core (manually curated) dataset. One of these sequences for each family was used as a seed for clustering homologs, and we used INTREPID to predict critical residues for all members. We ensured that sequence identities with the seed were not so high as to make the experiment uninformative (i.e., homologous enzymes from CSA were selected with less than 50% identity). Table S-4 compares the ranks of catalytic residues in a sequence that was not used as seed to the ranks when the same sequence was a seed. As sequence identity to the seed decreases, the accuracy of INTREPID also decreases. In the context of predicting catalytic positions in a single protein as opposed to the entire family, these results would apply to other sequence-based methods as well. Based on these limited results, we would recommend ensuring that the sequence of interest has sequence identity > 50% to the seed.
Fig. S-4. Effect of alignment diversity on the accuracy of Global-JS on the CSA-100 dataset: ROC curve for Global-JS on alignments with varying degrees of evolutionary divergence, indicated by the minimum percent identity to the seed. The original alignment with no sequences removed is labelled “Unrestricted”. Global-JS performs significantly better with increasing evolutionary divergence - from 42% sensitivity at 90% specificity and 25% identity trimming to 79% sensitivity when no sequences are removed.

Fig. S-5. Effect of alignment diversity on catalytic residue prediction on the CSA-100 dataset: (Left) ROC curve for INTREPID-LO on alignments with varying degrees of evolutionary divergence, indicated by the minimum percent identity to the seed. The original alignment with no sequences removed is labelled “Unrestricted”. INTREPID-LO performs significantly better with increasing evolutionary divergence - from 42% sensitivity at 90% specificity and 25% identity trimming to 76% sensitivity when no sequences are removed. (Right) Global-LO also benefits from increased sequence diversity though at 90% specificity, it appears to have optimal performance if sequences with identity less than 10% to the seed are excluded.

S-3 EXAMPLES OF INTREPID PREDICTIONS

In this section, we analyze INTREPID predictions on families found in the PhyloFacts resource (http://phylogenomics.berkeley.edu/phylofacts). For the families found in PhyloFacts, homologs were gathered from UniProt (Apweiler et al., 2004) using Floowerpower (Krishnamurthy et al., 2007) (with global-local settings and number of SHMM iterations set to 3) and re-aligned using MUSCLE (Edgar, 2004). For this analysis, we built neighbor-joining trees using the PHYLIP software though other tree construction algorithms may be used in practice. PhyloFacts displays the top 5% of the INTREPID predictions though the cutoff may be varied by the user (we handle tied scores by selecting all residues at a given score).

S-3.1 Dihydronopterin aldolase

Dihydronopterin aldolase catalyzes the conversion of 7,8-dihydronopterin (DHNP) to 6-hydroxymethyl-7,8-dihydropterin (HP) playing an essential role in the folate biosynthesis pathway. Mammals, unlike bacteria, plants, and yeast, lack a complete folate biosynthesis pathway and obtain folate from their diet (Lawrence et al., 2005). Hence, dihydronopterin aldolase, along with other enzymes in the folate biosynthesis pathway, has served as a target for antimicrobial and antibacterial agents (Lawrence et al., 2005).
Fig. S-6. Effect of alignment diversity on catalytic residue prediction on the CSA-100 dataset: (Left) ROC curve for INTREPID-RE on alignments with varying degrees of evolutionary divergence, indicated by the minimum percent identity to the seed. The original alignment with no sequences removed is labelled “Unrestricted”. INTREPID-RE performs significantly better with increasing evolutionary divergence - from 42% sensitivity at 90% specificity and 25% identity trimming to 85% sensitivity when no sequences are removed. (Right) Global-RE also benefits from increased sequence diversity.

Fig. S-7. Comparison of INTREPID-SPEC with and without subtype information on specificity residue prediction. Subtype information leads to a 10% increase in precision across the range of recall values.

| PDB id | Percent id | Rank as non-seed  | Rank as seed  |
|--------|------------|-------------------|---------------|
| 1hti   | 44.4       | 57,72,2,29,30     | 34,67,2,11,29 |
| 1pxv   | 32.9       | 124,124,4,4       | 9,32,7,17     |
| 1azw   | 10.8       | 201,120,64        | 8,2,1         |

Table S-4. INTREPID accuracy decreases with distance from the seed: The table shows the ranks (Rank as non-seed) assigned to the CSA catalytic residues by INTREPID on a sequence which was not the seed. These ranks are compared to the ranks (Rank as seed) when the same sequence was used as the seed. As the sequence identity to the seed decreases, the accuracy decreases as seen from the numerically higher “Rank as non-seed” column.

INTREPID predictions on dihydronopterin aldolase from *Staphylococcus aureus* (PDB Id: 1dhn) are shown in Figure S-12. INTREPID correctly predicts the residues E22 and K100, which are listed as catalytic in the CSA. INTREPID also predicts residues Q27 which forms a hydrogen bond to the substrate in the crystal structure of neopterin(NP), an analog of 7,8-dihydronopterin (DHN) (Wang et al., 2006); Y54 which is known to coordinate catalysis with E22; K74 which influences the affinity of the enzyme for the substrate (Wang et al., 2006; Hennig et al., 1998); and G17 and H16.

S-3.2 Src Homology 2 (SH2) domain

SH2 domains are found in multi-domain intracellular signaling proteins and play key roles in assembling signaling complexes by binding to phosphotyrosine moieties in target proteins. Key residues in SH2 binding pockets determine the specificity of interaction, and thereby the
INTREPID more effectively identifies catalytic residues that are not conserved across the family: Scatter plot comparing the ranks assigned by INTREPID and Global-JS on catalytic residues that are not conserved across the family. Lower ranks correspond to residues that are easily identified as catalytic. The diagonal denotes residues on which both methods do equally well, residues above the diagonal are those for which INTREPID gives better ranks, and residues below the diagonal are those for which Global-JS gives better ranks. Here, INTREPID gives better ranks to 34 residues and Global-JS to 15.

Fig. S-9. Alignment of the family containing enoyl-[acyl-carrier-protein] reductase from *Escherichia Coli* (PDB id: 1mfp) made non-redundant at 45% sequence identity. Positions marked in red correspond to catalytic residues (K163, Y156). Y156 is given a rank of 18 (out of 258 positions) by INTREPID, a rank of 58 by Global-JS, 100 by ConSurf and 31 by BCMET. See Figure S-10 for an expanded view of a subtree containing the seed 1mfp and Section 3.1.4 of the main paper for details.
Fig. S-10. Alignment of sequences in the subtree containing enoyl-[acyl-carrier-protein] reductase from *Escherichia Coli* (PDB id: 1mfp) made non-redundant at 70% sequence identity. Positions marked in red correspond to catalytic residues (K163, Y156). This subtree contains 199 sequences out of the original 833 sequences. Notice that Y156 is conserved within this subtree while it has a frequency of only 25% in the entire family. See Section 3.1.4 of the main paper for details.

Songyang et al. (1993). Amongst the top 6 residue predicted by INTREPID are R32, H58 with the other predictions being W5, G27, F29 and F77.

REFERENCES

Aloy, P., Querol, E., Aviles, F. X., and Sternberg, M. J. (2001). Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. *J Mol Biol*, 311(2), 395–408.
Fig. S-11. Alignment of the family containing Flavocytochrome b2 from Saccharomyces Cerevisiae (PDB id:1fcB) made non-redundant at 40% sequence identity. Positions marked in red correspond to catalytic residues (H373, R376).

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.

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

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

Henikoff, S. and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. Proceedings of the National Academy of Sciences of the United States of America, 89(22), 10915–10919.

Hennig, M., D'Arcy, A., Hampele, I., Page, M., Oefner, C., and Dale, G. (2007). Crystal structure and reaction mechanism of 7,8-dihydroneopterin aldolase from Staphylococcus aureus. Nat. Struct. Biol., 5, 357–362.

Krishnamurthy, N., Brown, D., and Spolander, K. (2007). Flowerpower: clustering proteins into domain architecture classes for phylogenomic inference of protein function. BMC Evolutionary Biology, 7(Suppl 1), S12.
INTREPID predictions for dihydroneopterin aldolase from *Staphylococcus aureus* (PDB Id:2dhn, BPG accession:bpg020587). INTREPID correctly predicts the catalytic residues E22, K100, Q27, K74, and Y54. Of these, only E22 and K100 are listed in the CSA. The non-CSA functional residues refer to INTREPID predictions that are not listed in the CSA but have experimental evidence of being catalytic (Q27, K75, and Y54) (see text).

Lawrence, M., Iliades, P., Fernley, R., Berglez, J., Pilling, P., and Macreadie, I. (2005). The three-dimensional structure of the bifunctional 6-hydroxymethyl-7,8-dihydropterin pyrophosphokinase/dihydropteroate synthase of Saccharomyces cerevisiae. *J. Mol. Biol.*, 348, 655–670.

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

Songyang, Z., Shoelson, S., Chaudhuri, M., Gish, G., Pawson, T., Haser, W., King, F., Roberts, T., Ratnofsky, S., and Lechleider, R. (1993). SH2 domains recognize specific phosphopeptide sequences. *Cell*, 72, 767–778.

Waksman, G., Shoelson, S. E., Pant, N., Cowburn, D., and Kuriyan, J. (1993). Binding of a high affinity phosphotyrosyl peptide to the src sh2 domain: Crystal structures of the complexed and peptide-free forms. *Cell*, 72(5), 779–790.

Wang, K. and Samudrala, R. (2006). Incorporating background frequency improves entropy-based residue conservation measures. *BMC Bioinformatics*, 7(1), 385.

Wang, Y., Li, Y., and Yan, H. (2006). Mechanism of dihydroneopterin aldolase: functional roles of the conserved active site glutamate and lysine residues. *Biochemistry*, 45(51), 15232–15239.