Supporting online material
for:

HFSP: High speed homology-driven function annotation of proteins

Mahlich, Y.[1,2,3*], Steinegger, M.[2,4,5], Rost, B.[2,3,6,7,8], Bromberg, Y.[1,3,9*]

1 Department of Biochemistry and Microbiology, Rutgers University, 76 Lipman Dr, New Brunswick, NJ 08873, USA
2 Computational Biology & Bioinformatics - i12 Informatics, Technical University of Munich (TUM) Boltzmannstrasse 3 85748 Garching/Munich Germany
3 Institute for Advanced Study, Technische Universität München, Lichtenbergstrasse 2 a, D-85748 Garching, Germany
4 Quantitative and Computational Biology group, Max-Planck Institute for Biophysical Chemistry, Am Fassberg 11, 37077 Göttingen, Germany
5 Department of Chemistry, Seoul National University, Seoul, Korea
6 TUM School of Life Sciences Weihenstephan (WZW), Freising, Germany
7 Columbia University, Department of Biochemistry and Molecular Biophysics, Columbia University, New York, USA
8 New York Consortium on Membrane Protein Structure (NYCOMPS), New York, USA
9 Department of Genetics, Rutgers University, Human Genetics Institute, Life Sciences Building, 145 Bevier Road, Piscataway, NJ 08854, USA

Table of Contents for Supporting Online Material

Supplementary Data 1

Excel file with two sheets, each containing the UniProt ID, EC number and eukaryote/prokaryote mapping of proteins in Swiss-Prot 2002 and Swiss-Prot 2017 respectively

Supplementary Data 2

Excel file includes multiple sheets, each containing the results of individual predictions for proteins in datasets in the manuscript. Each sheet contains 5 columns, protein_reference, ec_reference, protein_prediction, ec_prediction & hfsp_score.

protein_reference & protein_prediction contain the Uniport IDs of the reference protein and the aligned protein, respectively. ec_reference and ec_prediction are the EC numbers of the reference protein and aligned protein, respectively. hfsp_score is the HFSP score for the alignment.

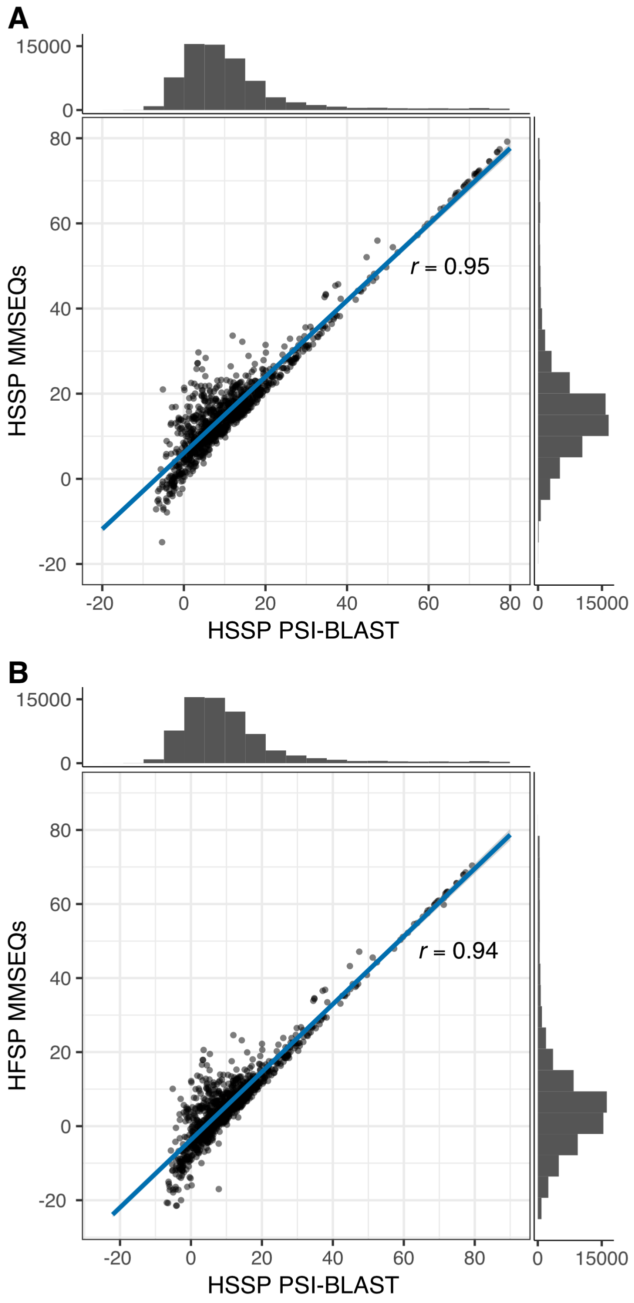
Table 1: 3rd level EC categories with over 50 proteins, sorted according to the number of proteins.

EC3	# Proteins
2.7.11	332
2.7.10	172
1.1.1	136
3.2.1	130
2.7.1	113
2.3.1	111
2.1.1	110
4.1.1	97
2.5.1	97
3.4.21	93
2.4.1	91
3.1.3	85
4.2.1	81
6.1.1	74
2.7.7	66
3.5.1	56
3.1.4	56
3.1.1	55
3.4.22	55

Table 2: F1-scores for each optimization run of HFSP-training.

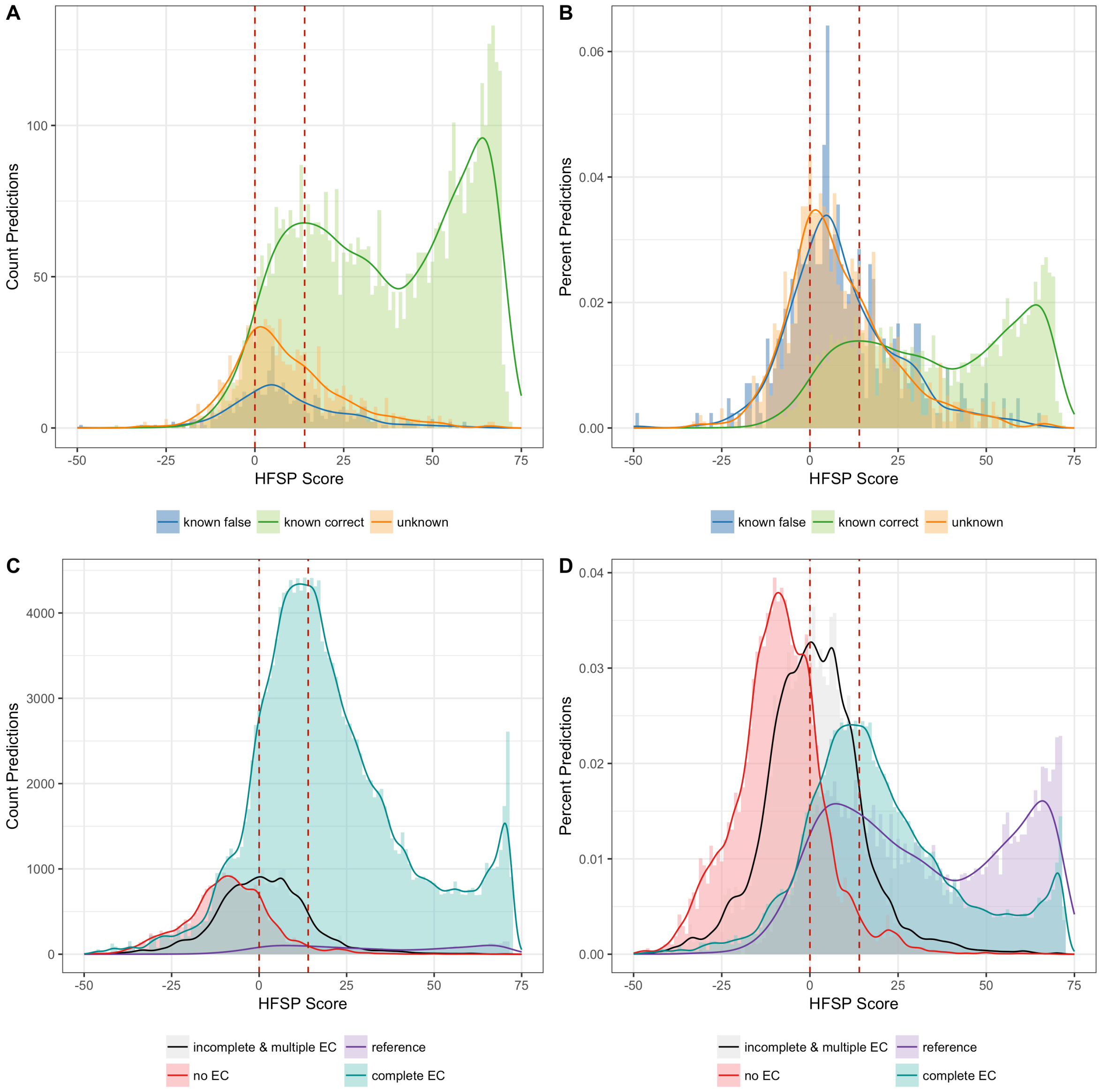
split	F1-score	Exponent	Factor
1	0.75	0.33	770
2	0.74	0.32	660
3	0.74	0.32	660
4	0.74	0.32	658
5	0.74	0.34	823
6	0.74	0.33	770
7	0.74	0.33	770
8	0.73	0.32	660
9	0.74	0.33	770
10	0.73	0.41	1646

Supplementary Figure 1



Supplementary Fig. 1: HSSP scores derived from MMSeqs2 and PSI-BLAST alignments strongly correlate. HSSP scores derived from PSI-BLAST alignments (x-axis) vs. **(A)** HSSP scores and **(B)** HFSP scores derived from MMSeqs2 (y-axis). The histograms display the number of protein pairs in the respective ranges of HSSP scores. In both scenarios HSSP/HFSP scores derived from MMSeqs2 highly correlate with HSSP scores from PSI-BLAST (Pearson-correlation coefficient = 0.95 / 0.94).

Supplementary Figure 2



Supplementary Fig. 2: Newly emerging enzyme functionality difficult to differentiate from incorrect function predictions. Proteins with no known homologs – approximated by experimentally annotated proteins, which have a unique EC number (orange) – show on average smaller highest scoring HFSP hits than proteins with homologs (green – correct predictions, blue – incorrect predictions). **(A/B)** Comparison of HFSP score distributions for highest scoring protein pair for Swiss-Prot 2017, **(A)** showing the distribution of raw counts and **(B)** the corresponding percentages of the respective datasets. **(C/D)**: Panels of counts and percentages as in (A/B), data is the Comparison of HFSP distributions for different subsets of the non-reduced Swiss-Prot: (i) experimentally verified enzymes (reference - purple), (ii) not experimentally verified enzymes with EC annotation complete on all 4 levels (complete EC - teal), (iii) enzymes with incomplete or multiple EC annotations (incomplete & multiple EC – black) and (iv) proteins that are not annotated as enzymes (no EC).