

Supplementary material: “The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens”

**List of Figures**

S1	Head-to-head comparison between top five CAFA3 versus CAFA1 methods . . . . .	3
S2	$F_{\max}$ curves for the top-performing methods in <i>partial</i> evaluation mode for (A) Molecular Function ontology, (B) Biological Process ontology, (C) Cellular Component ontology on <i>No Knowledge</i> benchmark. . . . .	4
S3	$F_{\max}$ curves for the top-performing methods on <i>limited knowledge</i> benchmarks for (A) Molecular Function ontology, (B) Biological Process ontology, (C) Cellular Component ontology in <i>full</i> evaluation mode . . . . .	5
S4	Weighted precision-recall curves for the top-performing methods for (A) Molecular Function ontology, (B) Biological Process ontology, (C) Cellular Component ontology on <i>No Knowledge</i> benchmark and <i>full</i> evaluation mode . . . . .	6
S5	Normalized RU-MI curves for the top-performing methods for (A) Molecular Function ontology, (B) Biological Process ontology, (C) Cellular Component ontology on <i>No Knowledge</i> benchmark and <i>full</i> evaluation mode . . . . .	7
S6	Top 10 $F_{\max}$ in <i>Homo Sapiens</i> . . . . .	9
S7	Top 10 $F_{\max}$ in <i>Arabidopsis thaliana</i> . . . . .	10
S8	Top 10 $F_{\max}$ in <i>Mus musculus</i> . . . . .	11
S9	Top 10 $F_{\max}$ in <i>Rattus norvegicus</i> . . . . .	12
S10	Top 10 $F_{\max}$ in <i>Escherichia coli K12</i> . . . . .	13
S11	Top 10 $F_{\max}$ in <i>Drosophila melanogaster</i> . . . . .	14
S12	Top 10 $F_{\max}$ in <i>Dictyostelium discoideum</i> . . . . .	15
S13	Top 10 $F_{\max}$ in <i>Danio rerio</i> . . . . .	16
S14	Top 10 $F_{\max}$ in <i>Candida albicans (strain SC5314 / ATCC MYA-2876)</i> . . . . .	17
S15	Distribution of benchmark depth in leaf nodes. A leaf node is defined if any descendent nodes are not included as benchmark. . . . .	21
S16	Frequency of total information content of benchmark proteins for (a) Molecular Function ontology, (b) Biological Process ontology, and (c) Cellular Component ontology. Data include all benchmark proteins and all experimentally annotated proteins at the point of benchmark collection $t_1$ . The red point indicates the value of information content for the predicted annotation using to the Naïve model. . . . .	22

## List of Tables

S1	Number of experimental annotations in UniProt-GOA for biofilm formation (GO:0042710) . .	18
S2	Number of experimental annotations in UniProt-GOA for cilium or flagellum-dependent cell motility (GO:0001539) . . . . .	19
S3	Number of experimental annotations in UniProt-GOA for long-term memory (GO:0007616) .	20

## Contents

<b>Mathematical definitions of protein-centric metrics</b>	<b>23</b>
<b>List of CAFA3 Keywords</b>	<b>25</b>

Fig S1. Head-to-head comparison between top five CAFA3 versus CAFA1 methods

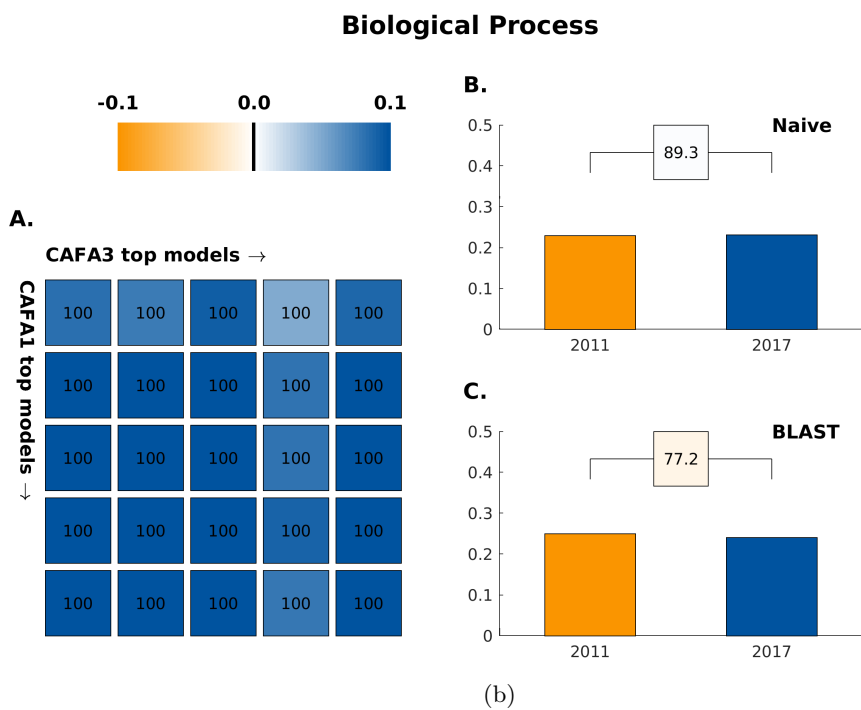
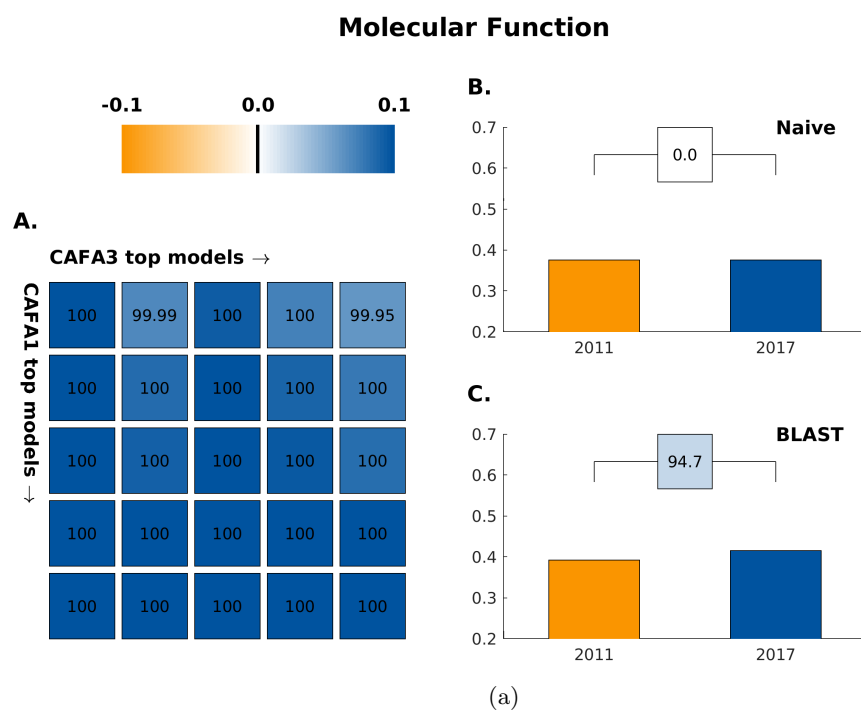
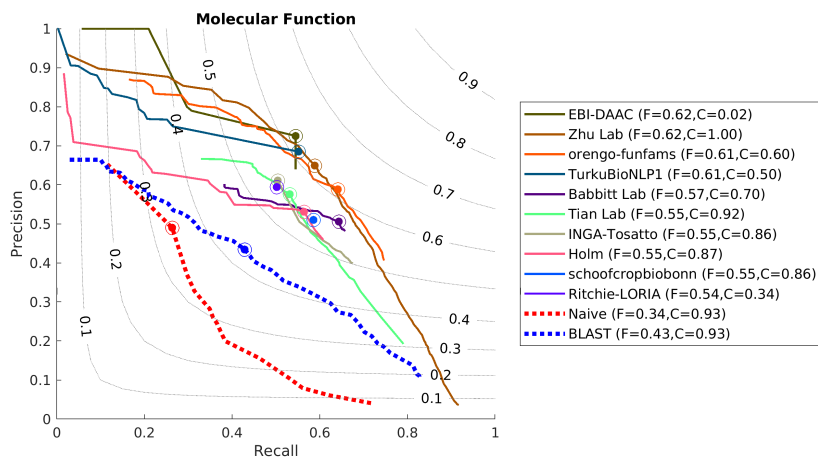
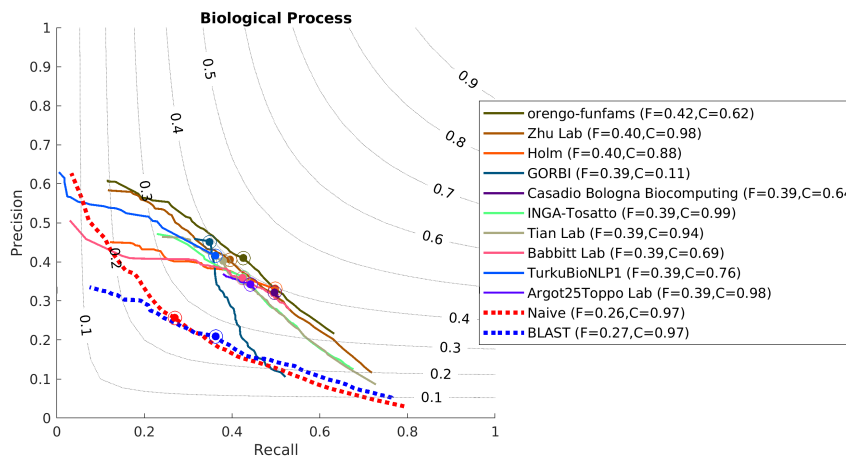


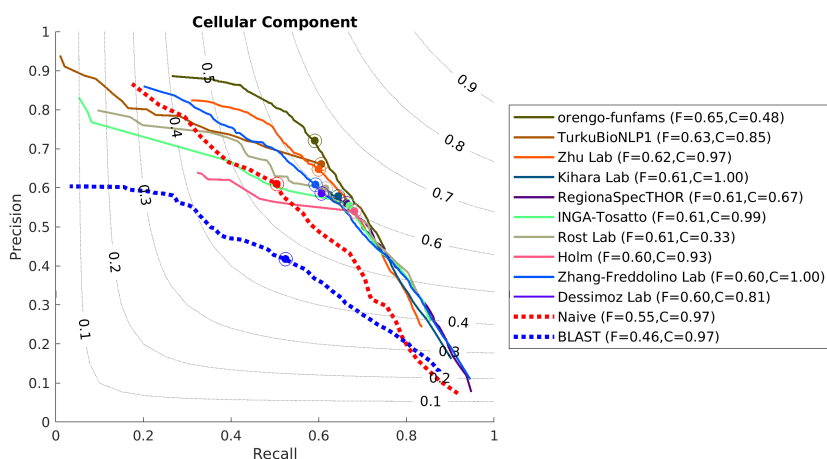
Fig S2.  $F_{\max}$  curves for the top-performing methods in *partial* evaluation mode for (A) Molecular Function ontology, (B) Biological Process ontology, (C) Cellular Component ontology on *No Knowledge* benchmark.



(a)

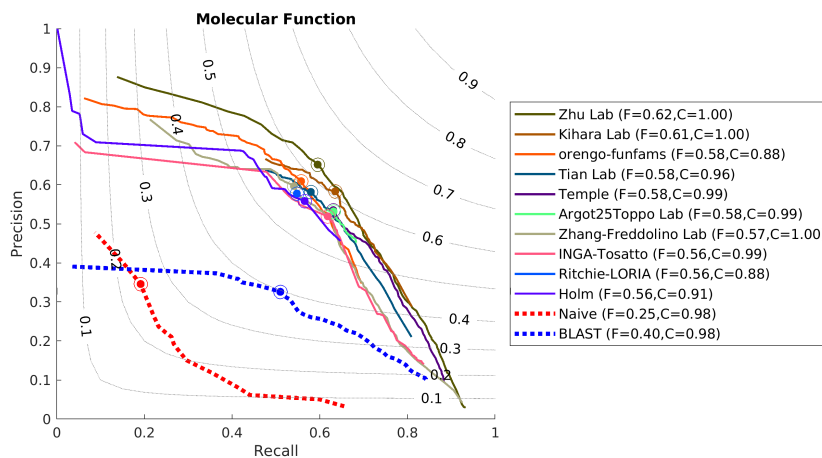


(b)

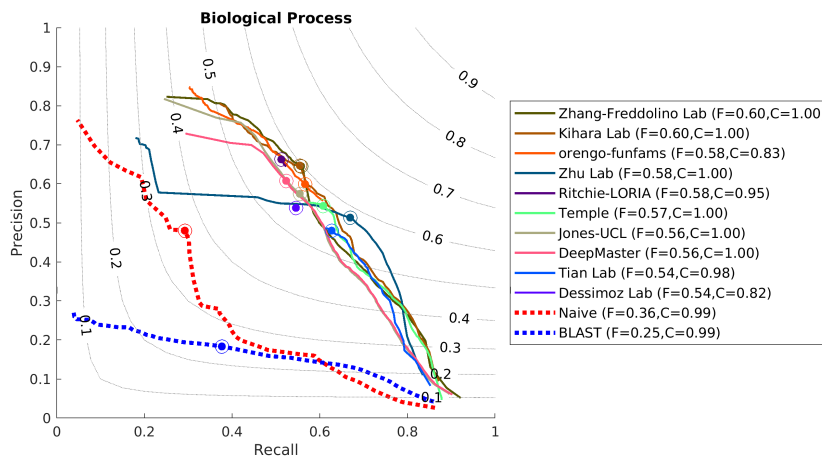


(c)

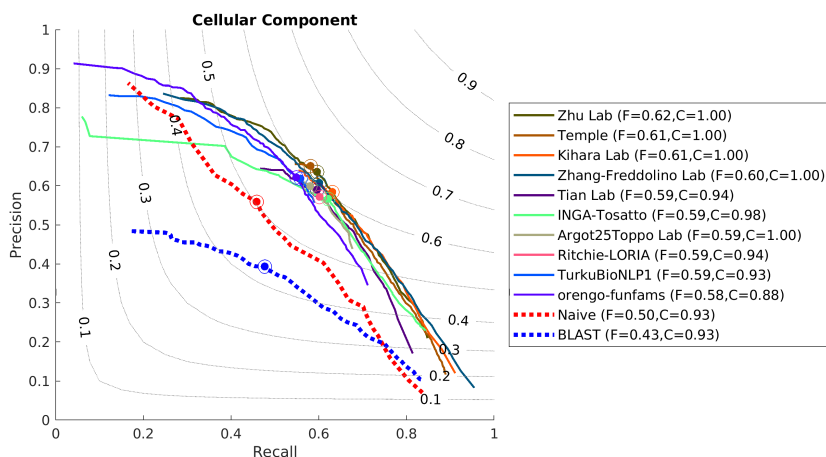
Fig S3.  $F_{\max}$  curves for the top-performing methods on *limited knowledge* benchmarks for (A) Molecular Function ontology, (B) Biological Process ontology, (C) Cellular Component ontology in *full* evaluation mode



(a)

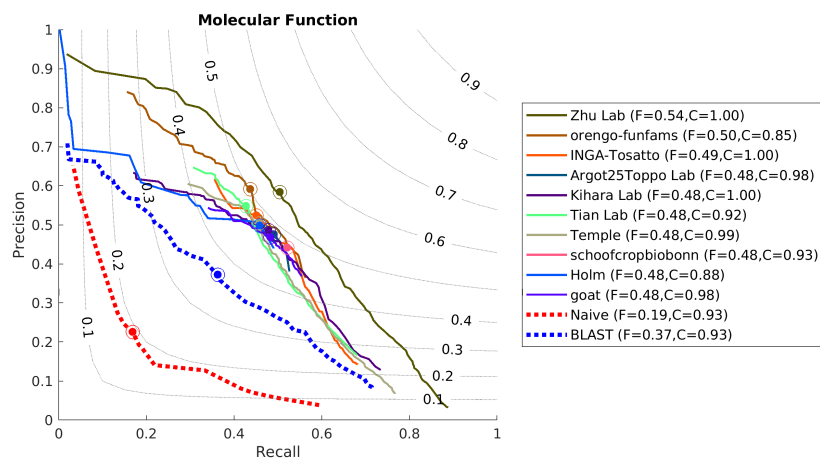


(b)

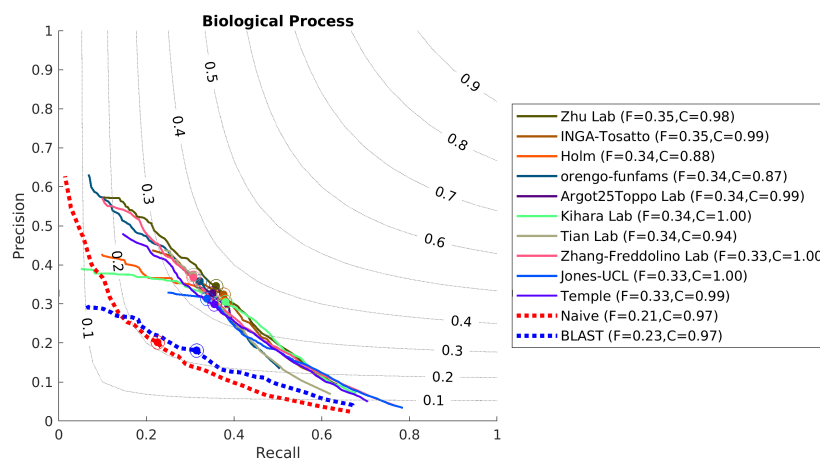


(c)

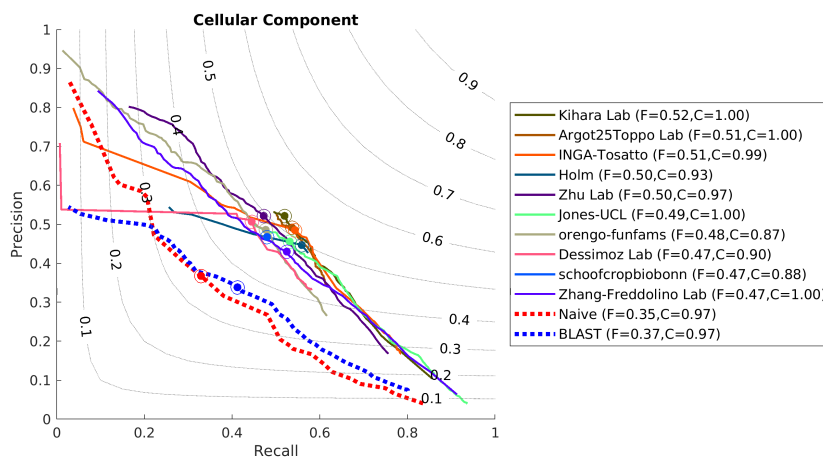
Fig S4. Weighted precision-recall curves for the top-performing methods for (A) Molecular Function ontology, (B) Biological Process ontology, (C) Cellular Component ontology on *No Knowledge* benchmark and *full* evaluation mode



(a)

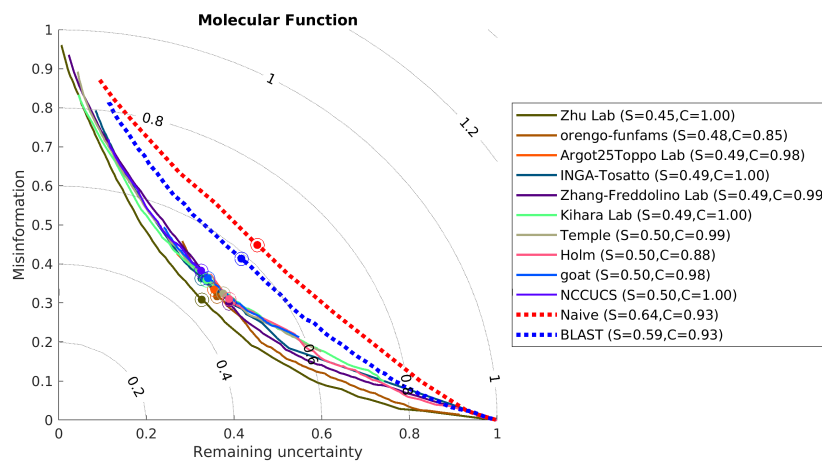


(b)

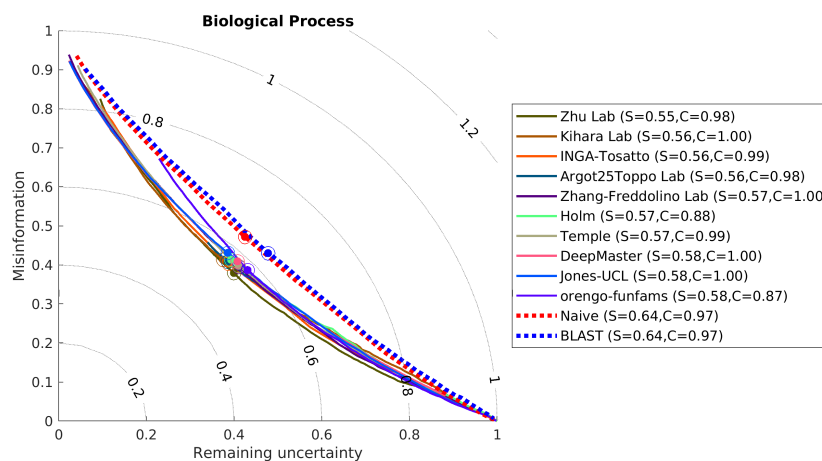


(c)

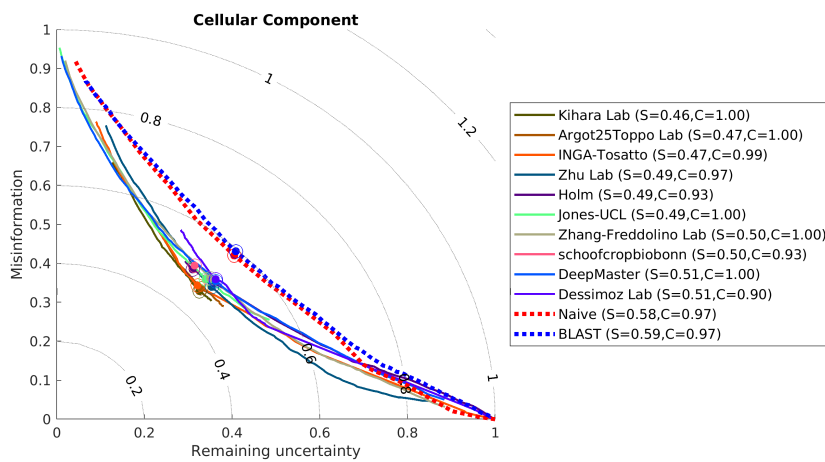
Fig S5. Normalized RU-MI curves for the top-performing methods for (A) Molecular Function ontology, (B) Biological Process ontology, (C) Cellular Component ontology on *No Knowledge* benchmark and *full* evaluation mode



(a)



(b)

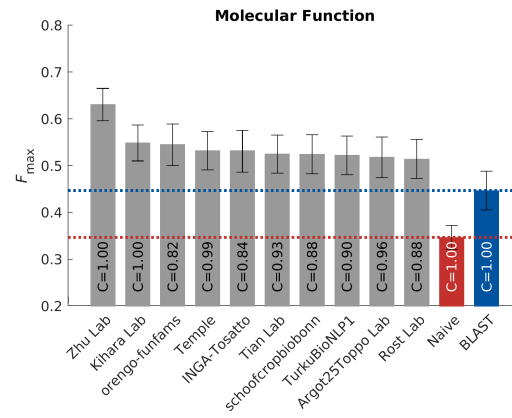


(c)

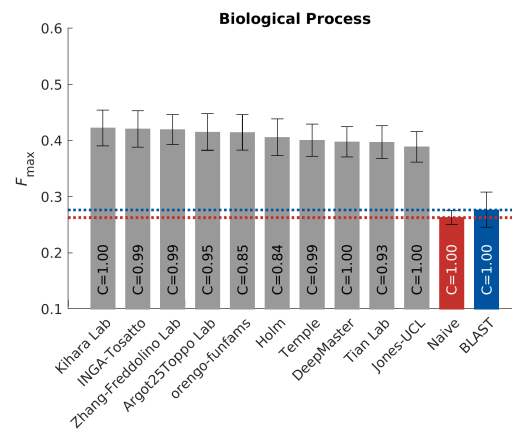
The followings figures S6, S7, S8, S9, S10, S11, S12, S13, S14 shows evaluations based on the *No Knowledge* benchmarks in the *full* evaluation mode.



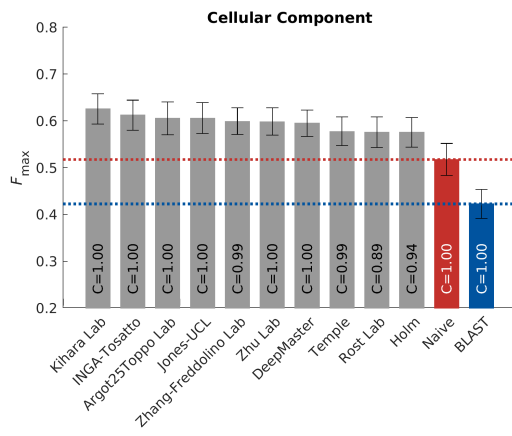
Fig S6. Top 10  $F_{\max}$  in *Homo Sapiens*



(a)

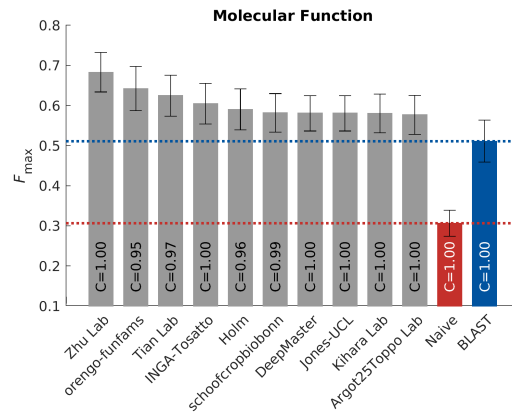


(b)

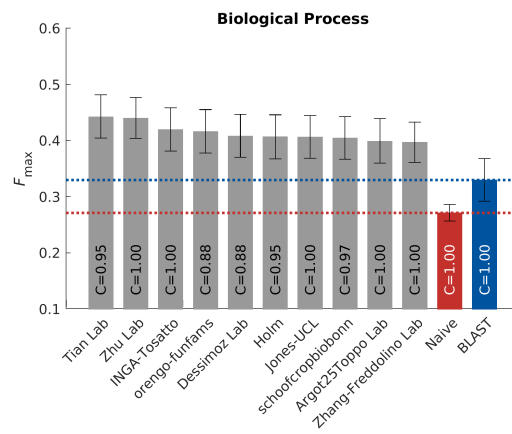


(c)

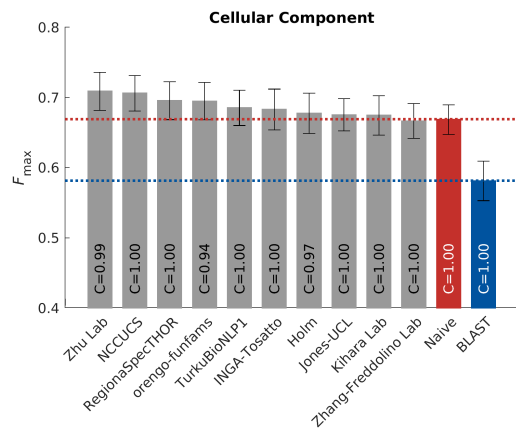
Fig S7. Top 10  $F_{\max}$  in *Arabidopsis thaliana*



(a)

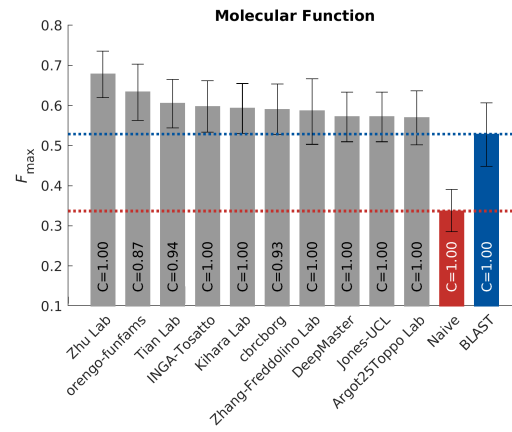


(b)

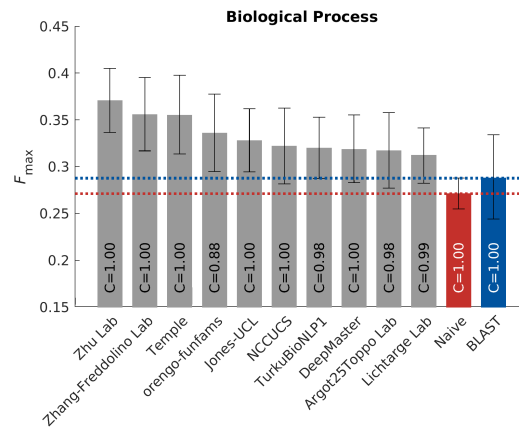


(c)

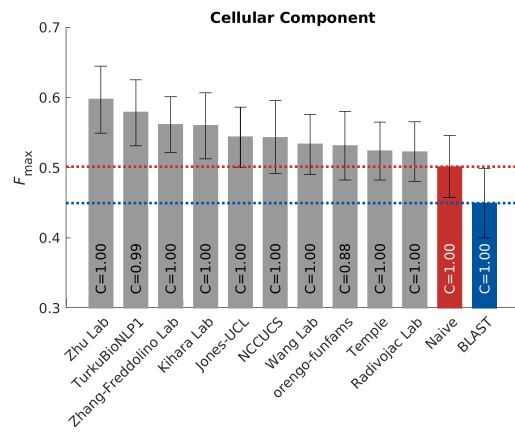
Fig S8. Top 10  $F_{\max}$  in *Mus musculus*



(a)

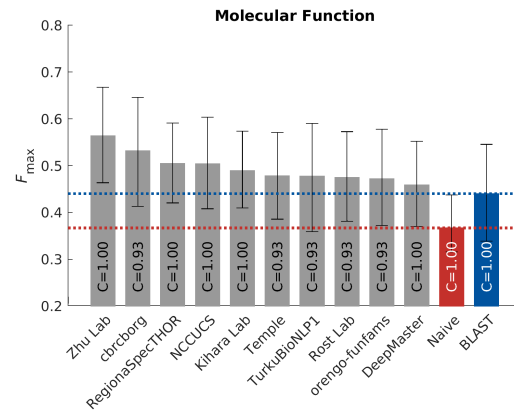


(b)

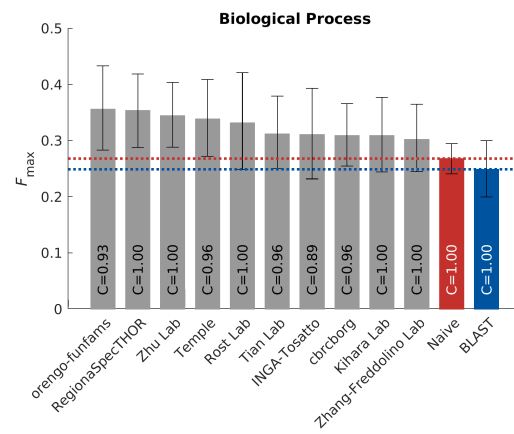


(c)

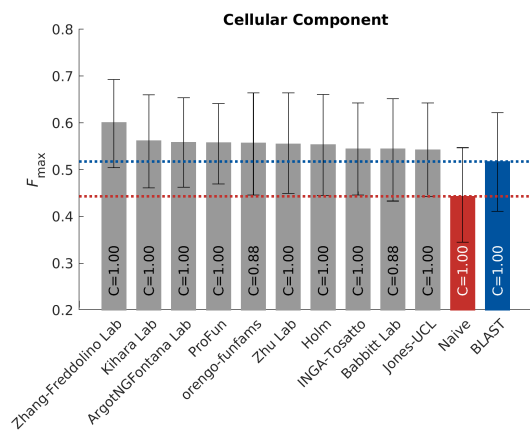
Fig S9. Top 10  $F_{\max}$  in *Rattus norvegicus*



(a)

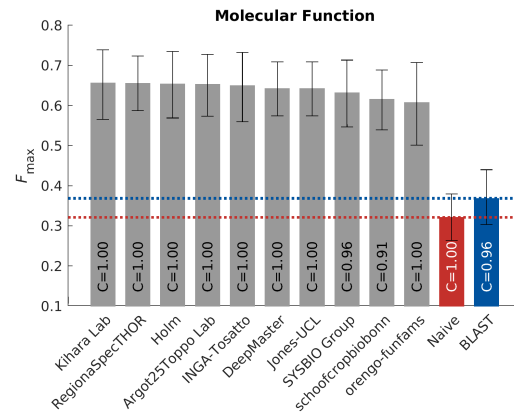


(b)

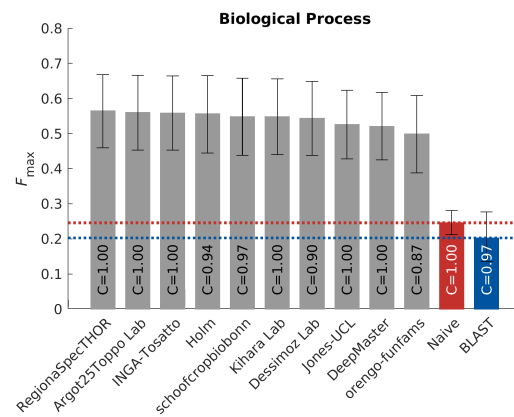


(c)

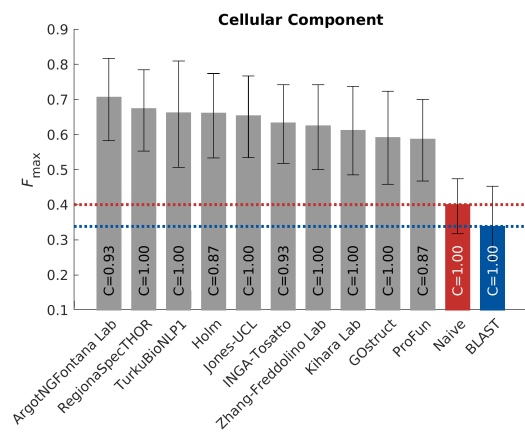
Fig S10. Top 10  $F_{\max}$  in *Escherichia coli* K12



(a)

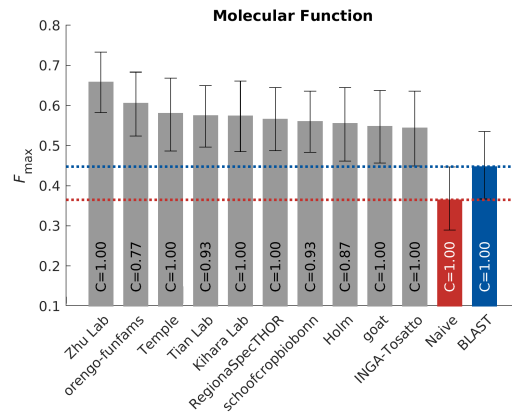


(b)

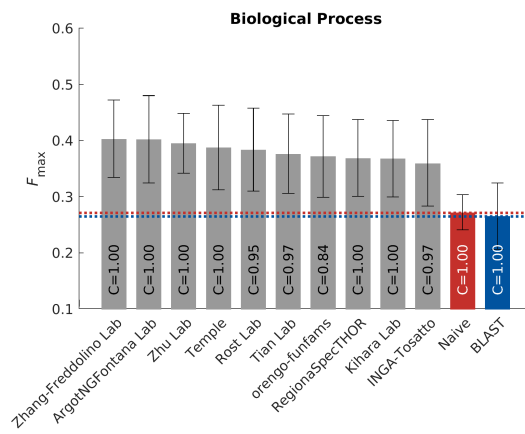


(c)

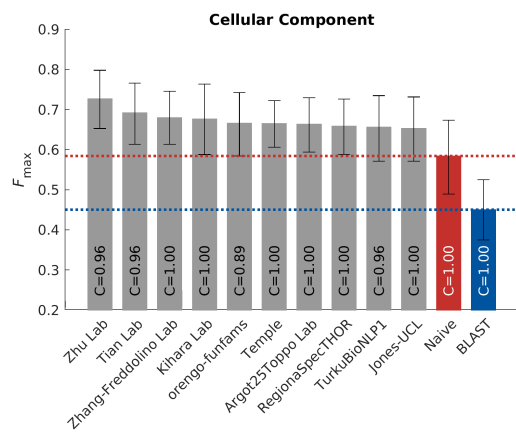
Fig S11. Top 10  $F_{\max}$  in *Drosophila melanogaster*



(a)

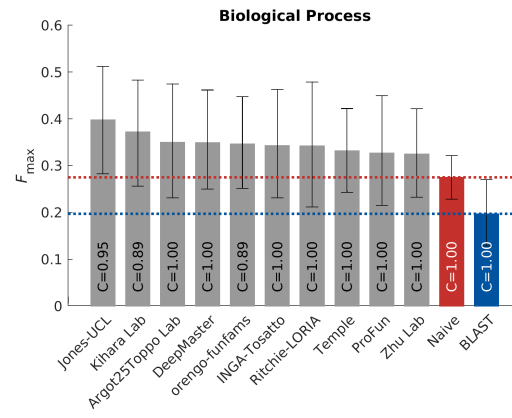


(b)

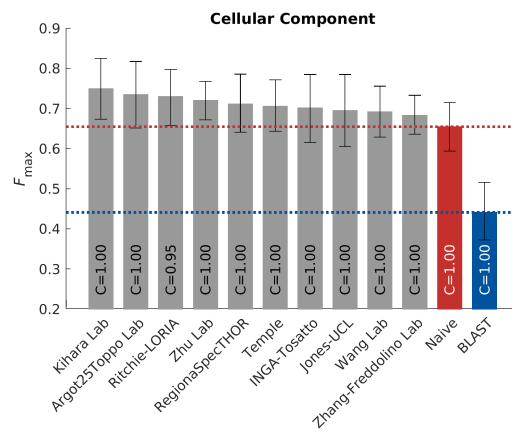


(c)

Fig S12. Top 10  $F_{\max}$  in *Dictyostelium discoideum*

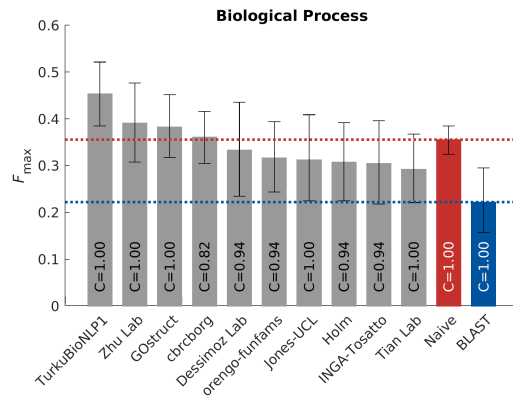


(a)



(b)

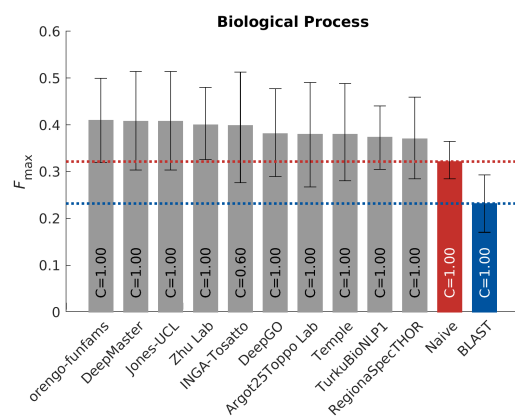
Fig S13. Top 10  $F_{\max}$  in *Danio rerio*



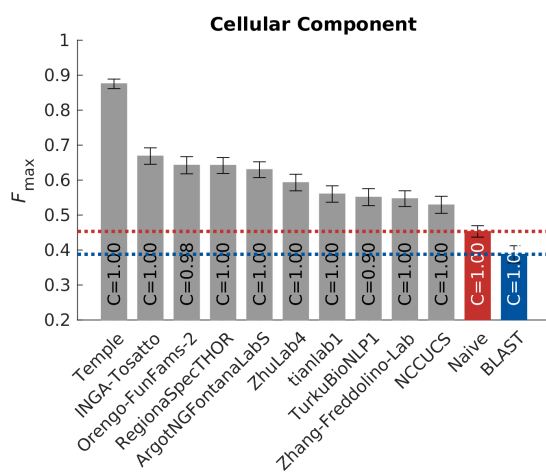
(a)



Fig S14. Top 10  $F_{\max}$  in *Candida albicans* (strain SC5314 / ATCC MYA-2876)



(a)



(b)

Table S1: Number of experimental annotations in UniProt-GOA for biofilm formation (GO:0042710)

Taxonomy ID	Scientific Name	Number of experimental annotations
237561	<i>Candida albicans</i> SC5314	188
208964	<i>Pseudomonas aeruginosa</i> PAO1	54
83333	<i>Escherichia coli</i> K-12	46
559292	<i>Saccharomyces cerevisiae</i> S288C	10
284593	[ <i>Candida</i> ] <i>glabrata</i> CBS 138	5
578454	<i>Candida parapsilosis</i> CDC317	4
190486	<i>Xanthomonas citri</i> pv. <i>citri</i> str. 306	2
290339	<i>Cronobacter sakazakii</i> ATCC BAA-894	2
216592	<i>Escherichia coli</i> 042	2
5476	<i>Candida albicans</i>	2
224308	<i>Bacillus subtilis</i> subsp. <i>subtilis</i> str. 168	1
1314	<i>Streptococcus pyogenes</i>	1
330879	<i>Aspergillus fumigatus</i> Af293	1
210007	<i>Streptococcus mutans</i> UA159	1
243277	<i>Vibrio cholerae</i> O1 biovar <i>El Tor</i> str. N16961	1
227321	<i>Aspergillus nidulans</i> FGSC A4	1
93061	<i>Staphylococcus aureus</i> subsp. <i>aureus</i> NCTC 8325	1
316273	<i>Xanthomonas campestris</i> pv. <i>vesicatoria</i> str. 85-10	1
1280	<i>Staphylococcus aureus</i>	1
100226	<i>Streptomyces coelicolor</i> A3(2)	1

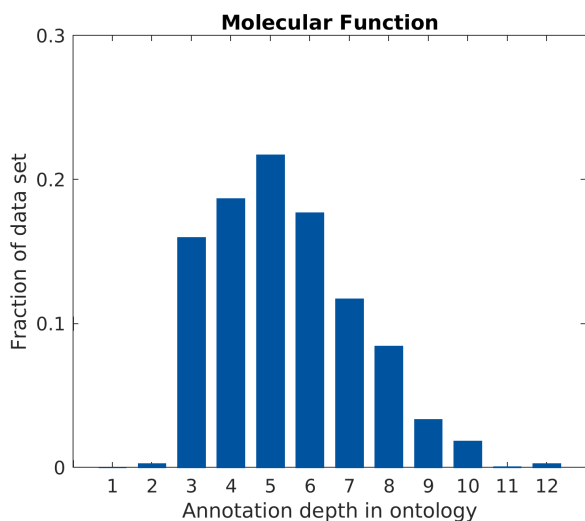
Table S2: Number of experimental annotations in UniProt-GOA for cilium or flagellum-dependent cell motility (GO:0001539)

Taxonomy ID	Scientific Name	Number of experimental annotations
208964	<i>Pseudomonas aeruginosa</i> PAO1	41
3055	<i>Chlamydomonas reinhardtii</i>	40
83333	<i>Escherichia coli</i> K-12	30
185431	<i>Trypanosoma brucei brucei</i> TREU927	26
7227	<i>Drosophila melanogaster</i>	16
10090	<i>Mus musculus</i>	7
7955	<i>Danio rerio</i>	5
9606	<i>Homo sapiens</i>	4
287	<i>Pseudomonas aeruginosa</i>	4
85962	<i>Helicobacter pylori</i> 26695	3
224308	<i>Bacillus subtilis subsp. subtilis</i> str. 168	2
189518	<i>Leptospira interrogans serovar Lai</i> str. 56601	1
5664	<i>Leishmania major</i>	1
246197	<i>Myxococcus xanthus</i> DK 1622	1
9913	<i>Bos taurus</i>	1
529507	<i>Proteus mirabilis</i> HI4320	1
9615	<i>Canis lupus familiaris</i>	1
99287	<i>Salmonella enterica subsp. enterica serovar Typhimurium</i> str. LT2	1
8090	<i>Oryzias latipes</i>	1
31286	<i>Trypanosoma brucei rhodesiense</i>	1

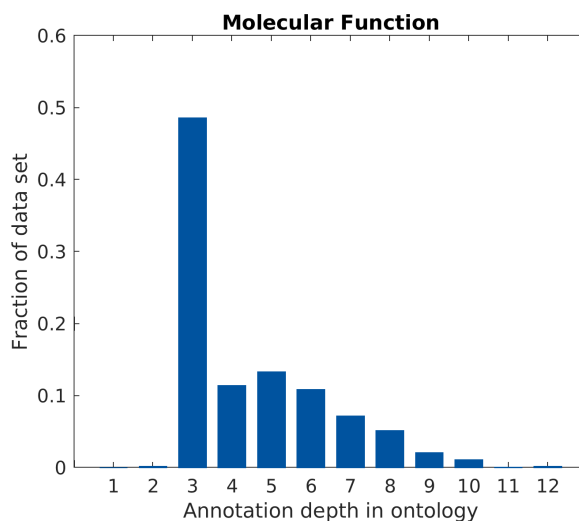
Table S3: Number of experimental annotations in UniProt-GOA for long-term memory (GO:0007616)

Taxonomy ID	Scientific Name	Number of experimental annotations
7227	<i>Drosophila melanogaster</i>	217
10090	<i>Mus musculus</i>	23
10116	<i>Rattus norvegicus</i>	22
6239	<i>Caenorhabditis elegans</i>	13
9606	<i>Homo sapiens</i>	7
381128	<i>Lehmannia marginata</i>	1

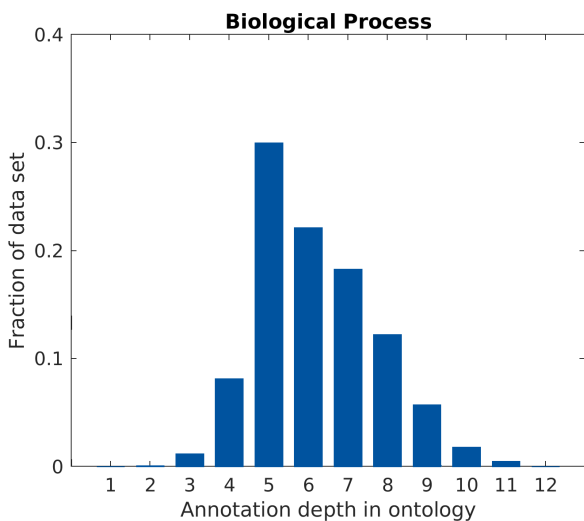
Fig S15. Distribution of benchmark depth in leaf nodes. A leaf node is defined if any descendent nodes are not included as benchmark.



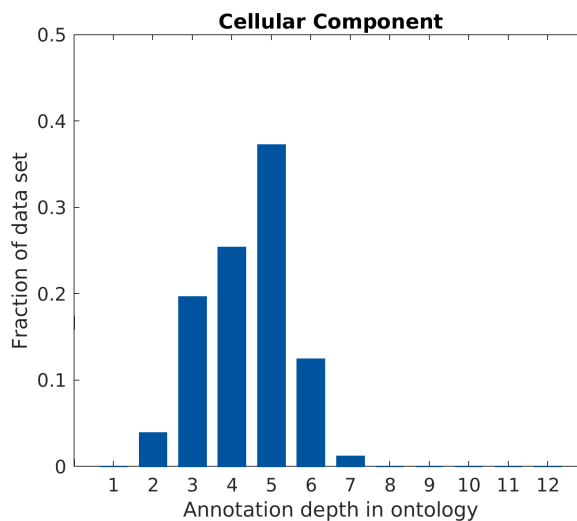
(a) After removing single protein-binding annotations



(b) Before removing single protein-binding annotations

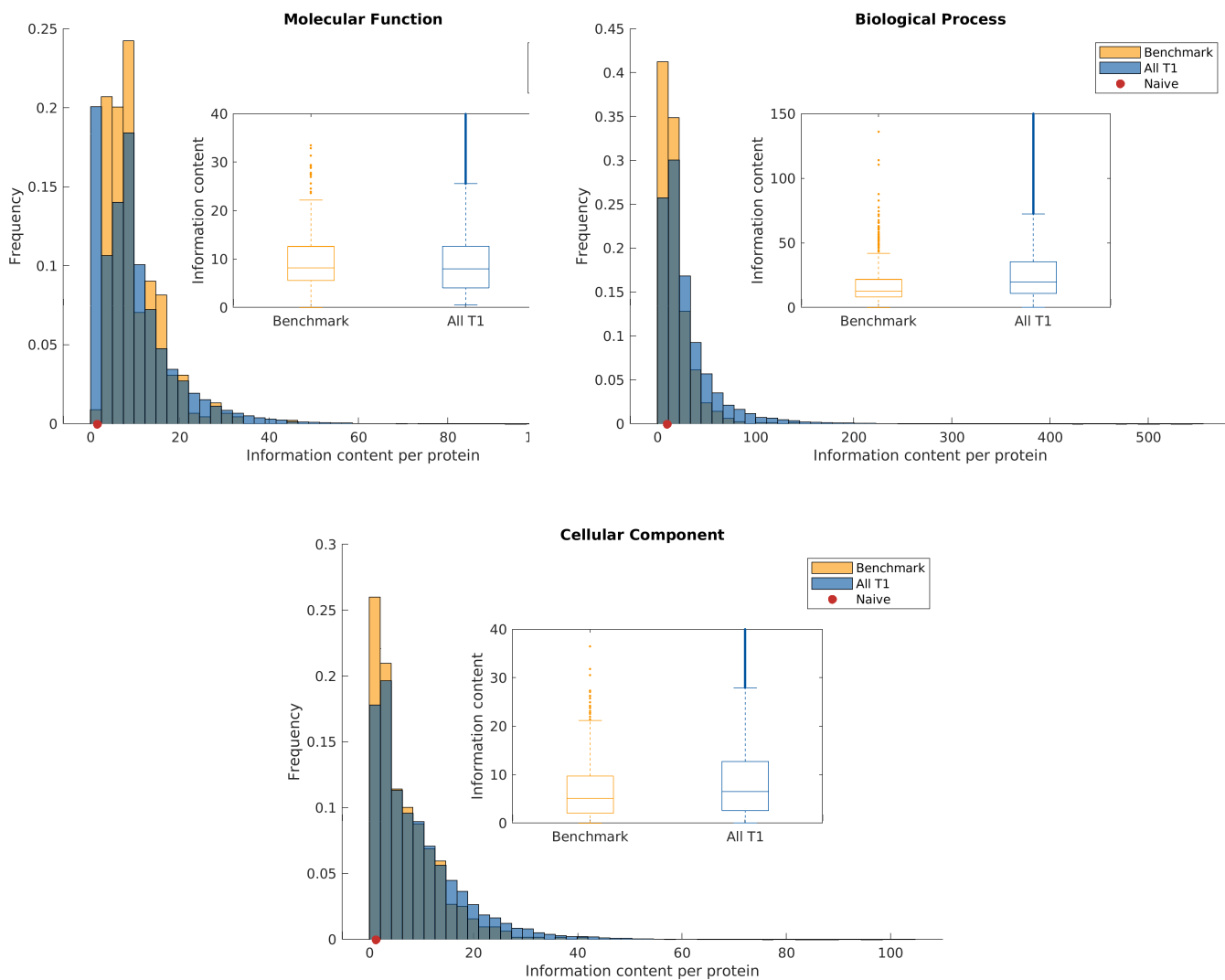


(c)



(d)

Fig S16. Frequency of total information content of benchmark proteins for (a) Molecular Function ontology, (b) Biological Process ontology, and (c) Cellular Component ontology. Data include all benchmark proteins and all experimentally annotated proteins at the point of benchmark collection  $t_1$ . The red point indicates the value of information content for the predicted annotation using to the Naïve model.



## Mathematical definitions of protein-centric metrics

- Precision Recall

$$\begin{aligned}
 pr(\tau) &= \frac{1}{m(\tau)} \sum_{i=1}^{m(\tau)} \frac{\sum_f \mathbb{1}(f \in P_i(\tau) \wedge f \in T_i)}{\sum_f \mathbb{1}(f \in P_i(\tau))}, \\
 rc(\tau) &= \frac{1}{n_e} \sum_{i=1}^{n_e} \frac{\sum_f \mathbb{1}(f \in P_i(\tau) \wedge f \in T_i)}{\sum_f \mathbb{1}(f \in T_i)}, \\
 F_{\max} &= \max_{\tau} \left\{ \frac{2 \cdot pr(\tau) \cdot rc(\tau)}{pr(\tau) + rc(\tau)} \right\},
 \end{aligned}$$

where  $P_i(\tau)$  denotes the set of terms that have predicted scores greater than or equal to  $\tau$  for a protein sequence  $i$ ,  $T_i$  denotes the corresponding ground-truth set of terms for that sequence,  $m(\tau)$  is the number of sequences with at least one predicted score greater than or equal to  $\tau$ ,  $\mathbb{1}(\cdot)$  is an indicator function and  $n_e$  is the number of targets used in a particular mode of evaluation. In the full evaluation mode  $n_e = n$ , the number of benchmark proteins, whereas in the partial evaluation mode  $n_e = m(0)$ ; i.e., the number of proteins which were chosen to be predicted using the particular method. For each method, we refer to  $\frac{m(0)}{n}$  as the *coverage* because it provides the fraction of benchmark proteins on which the method made any predictions.

- Remaining Uncertainty and Missing Information

$$\begin{aligned}
 ru(\tau) &= \frac{1}{n_e} \sum_{i=1}^{n_e} \sum_f ic(f) \cdot \mathbb{1}(f \notin P_i(\tau) \wedge f \in T_i), \\
 mi(\tau) &= \frac{1}{n_e} \sum_{i=1}^{n_e} \sum_f ic(f) \cdot \mathbb{1}(f \in P_i(\tau) \wedge f \notin T_i), \\
 S_{\min} &= \min_{\tau} \left\{ \sqrt{ru(\tau)^2 + mi(\tau)^2} \right\},
 \end{aligned}$$

where  $ic(f)$  is the information content of the ontology term  $f$  [?]. It is estimated in a maximum likelihood manner as the negative binary logarithm of the conditional probability that the term  $f$  is present in a protein's annotation given that all its parent terms are also present. Note that here,  $n_e = n$  in the full evaluation mode and  $n_e = m(0)$  in the partial evaluation mode applies to both  $ru$  and  $mi$ .

- Weighted Precision Recall

$$wpr(\tau) = \frac{1}{m(\tau)} \sum_{i=1}^{m(\tau)} \frac{\sum_f ic(f) \cdot \mathbb{1}(f \in P_i(\tau) \wedge T_i(\tau))}{\sum_f ic(f) \cdot \mathbb{1}(f \in P_i(\tau))}, \quad \text{and}$$

$$wrc(\tau) = \frac{1}{n_e} \sum_{i=1}^{n_e} \frac{\sum_f ic(f) \cdot \mathbb{1}(f \in P_i(\tau) \wedge T_i(\tau))}{\sum_f ic(f) \cdot \mathbb{1}(f \in T_i(\tau))},$$

where  $P_i(\tau)$  is the set of predicted terms for protein  $i$  with score no less than threshold  $\tau$  and  $T_i$  is the set of true terms for protein  $i$ ,  $m(\tau)$  is the number of sequences with at least one predicted score greater than or equal to  $\tau$ , and  $n_e$  is the number of proteins used in a particular mode of evaluation. In the full evaluation mode  $n_e = n$ , the number of benchmark proteins, whereas in the partial evaluation mode  $n_e = m(0)$ .

- Normalized Remaining Uncertainty and Missing Information

$$nru(\tau) = \frac{1}{n_e} \sum_{i=1}^{n_e} \frac{\sum_f ic(f) \cdot \mathbb{1}(f \notin P_i(\tau) \wedge f \in T_i)}{\sum_f ic(f) \cdot \mathbb{1}(f \in P_i(\tau) \vee f \in T_i)}, \quad \text{and}$$

$$nmi(\tau) = \frac{1}{n_e} \sum_{i=1}^{n_e} \frac{\sum_f ic(f) \cdot \mathbb{1}(f \in P_i(\tau) \wedge f \notin T_i)}{\sum_f ic(f) \cdot \mathbb{1}(f \in P_i(\tau) \vee f \in T_i)},$$

where  $P_i(\tau)$  is the set of predicted terms for protein  $i$  with score no less than threshold  $\tau$  and  $T_i$  is the set of true terms for protein  $i$ , and  $n_e$  is the number of proteins used in a particular mode of evaluation. In the full evaluation mode  $n_e = n$ , the number of benchmark proteins, whereas in the partial evaluation mode  $n_e$  is the number of proteins that have at least one positive predicted score.



## List of CAFA3 Keywords

sequence alignment, sequence-profile alignment, profile-profile alignment, phylogeny, sequence properties, physicochemical properties, predicted properties, protein interactions, gene expression, mass spectrometry, genetic interactions, protein structure, literature, genomic context, synteny, structure alignment, comparative model, predicted protein structure, de novo prediction, machine learning, genome environment, operon, ortholog, paralog, homolog, hidden Markov model, clinical data, genetic data, natural language processing, other functional information