HPO2GO: prediction of human phenotype ontology term associations using cross ontology annotation co-occurrences

Tunca Doğan [Corresp. 1, 2, 3]

[1] Department of Health Informatics, Graduate School of Informatics, Middle East Technical University, Ankara, Turkey

[2] Cancer Systems Biology Laboratory (CanSyL), Graduate School of Informatics, Middle East Technical University, Ankara, Turkey

[3] European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridge, UK

Corresponding Author: Tunca Doğan
Email address: tdogan@metu.edu.tr

Analysing the relationships between biomolecules and the genetic diseases is a highly active area of research, where the aim is to identify the genes and their products that cause a particular disease due to functional changes originated from mutations. Biological ontologies are frequently employed in these studies, which provides researchers with extensive opportunities for knowledge discovery through computational data analysis. In this study, a novel approach is proposed for the identification of relationships between biomedical entities by automatically mapping phenotypic abnormality defining HPO terms with biomolecular function defining GO terms, where each association indicates the occurrence of the abnormality due to the loss of the biomolecular function expressed by the corresponding GO term. The proposed HPO2GO mappings were extracted by calculating the frequency of the co-annotations of the terms on the same genes/proteins, using already existing curated HPO and GO annotation sets. This was followed by the filtering of the unreliable mappings that could be observed due to chance, by statistical resampling of the co-occurrence similarity distributions. Furthermore, the biological relevance of the finalized mappings were discussed over selected cases, using the literature. The resulting HPO2GO mappings can be employed in different settings to predict and to analyse novel gene/protein - ontology term - disease relations. As an application of the proposed approach, HPO term – protein associations (i.e., HPO2protein) were predicted. In order to test the predictive performance of the method on a quantitative basis, and to compare it with the state-of-the-art, CAFA2 challenge HPO prediction target protein set was employed. The results of the benchmark indicated the potential of the proposed approach, as HPO2GO performed better than all of the models from 38 participating groups (with Fmax = 0.402), by a margin of 12.6% compared to the top performer. The automated cross ontology mapping approach developed in this work may be extended to other ontologies as well, to identify unexplored relation patterns at

the systemic level. The datasets, results and the source code of HPO2GO are available for download at: https://github.com/cansyl/HPO2GO.

1 **HPO2GO: Prediction of Human Phenotype Ontology Term**

2 **Associations Using Cross Ontology Annotation Co-occurrences**

4 Tunca Doğan[1,2,3,*]

6 [1] Cancer Systems Biology Laboratory (KanSiL), Graduate School of Informatics, METU, Ankara,

7 06800, Turkey

8 [2] Department of Health Informatics, Graduate School of Informatics, METU, Ankara, 06800,

9 Turkey

10 [3] European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI),

11 Hinxton, Cambridge, CB10 1SD, UK

12 * corresponding author email address: tdogan@metu.edu.tr

14 **ABSTRACT**

15 Analysing the relationships between biomolecules and the genetic diseases is a highly active area

16 of research, where the aim is to identify the genes and their products that cause a particular disease

17 due to functional changes originated from mutations. Biological ontologies are frequently

18 employed in these studies, which provides researchers with extensive opportunities for knowledge

19 discovery through computational data analysis.

20 In this study, a novel approach is proposed for the identification of relationships between

21 biomedical entities by automatically mapping phenotypic abnormality defining HPO terms with

22 biomolecular function defining GO terms, where each association indicates the occurrence of the

23 abnormality due to the loss of the biomolecular function expressed by the corresponding GO term.

24 The proposed HPO2GO mappings were extracted by calculating the frequency of the co-

25 annotations of the terms on the same genes/proteins, using already existing curated HPO and GO

26 annotation sets. This was followed by the filtering of the unreliable mappings that could be

27 observed due to chance, by statistical resampling of the co-occurrence similarity distributions.

28 Furthermore, the biological relevance of the finalized mappings were discussed over selected

29 cases, using the literature.

30 The resulting HPO2GO mappings can be employed in different settings to predict and to analyse

31 novel gene/protein - ontology term - disease relations. As an application of the proposed approach,

32 HPO term – protein associations (i.e., HPO2protein) were predicted. In order to test the predictive

33 performance of the method on a quantitative basis, and to compare it with the state-of-the-art,

34 CAFA2 challenge HPO prediction target protein set was employed. The results of the benchmark

35 indicated the potential of the proposed approach, as HPO2GO performance was among the best

36 ($Fmax$ = 0.35). The automated cross ontology mapping approach developed in this work may be

37 extended to other ontologies as well, to identify unexplored relation patterns at the systemic level.

38 The datasets, results and the source code of HPO2GO are available for download at:

39 https://github.com/cansyl/HPO2GO.

40   **1. INTRODUCTION AND BACKGROUND**

41   Systematic definition of biomedical entities (e.g., diseases, abnormalities, symptoms, traits, gene

42   and protein attributes, activities, functions and etc.) is crucial for computational studies in

43   biomedicine. Ontological systems, composed of standardized controlled vocabularies, are

44   employed for this purpose. Human Phenotype Ontology (HPO) system annotates disease records

45   (i.e., terms and definitions about diseases together with related information) with a standardized

46   phenotypic vocabulary (Robinson *et al.*, 2008; Köhler *et al.*, 2017). HPO is composed of five

47   independent sub-ontologies namely, *phenotypic abnormality* (i.e., the main sub-ontology defining

48   the basic qualities of diseases), *mode of inheritance* (i.e., annotates diseases in terms of mendelian

49   or non-mendelian principles), *mortality / aging* (i.e., information related to age of death due to the

50   corresponding disease), *frequency* (i.e., frequency of the disease in a patient cohort) and the *clinical*

51   *modifier* (i.e., additional disease characterization such as lethality, severity, etc.). Within each sub-

52   ontology, all terms are related to each other with a parent-child relationship, where each child term

53   defines a specific aspect of its parent. HPO has a directed acyclic graph (DAG) structure. The

54   sources of the disease information in HPO are Orphanet (Rath *et al.*, 2012), DECIPHER (Firth *et*

55   *al.*, 2009), and OMIM (Amberger *et al.*, 2014) databases. Each term in the *phenotypic abnormality*

56   sub-ontology define a specific type of abnormality encountered in human diseases (e.g.,

57   HP:0001631 - atrial septal defect). The generation of HPO terms (and their associations with

58   diseases) are carried out via both manual curation efforts and automated procedures (e.g., text

59   mining). The curation job is usually done by experts by reviewing the relevant literature

60   publications along with the disease centric information at various biomedical data resources. For

61   each association between a disease term and an HPO term, there is an evidence code tag to specify

62   the source of the information (i.e., curated or automated). The evidence codes used in HPO are

63   IEA (inferred from electronical annotation), PCS (published clinical study), ICE (individual

64   clinical experience), ITM (inferred by text mining), TAS (traceable author statement). As of

65   January 2018, the growing library of HPO contains nearly 12,000 phenotype terms, providing

66   more than 123,000 annotations to 7,000 different rare (mostly Mendelian) diseases and the newly

67   added 132,000 annotations to 3,145 common diseases (Groza *et al.*, 2015). A long-term goal of

68   the HPO project is for the system to be adopted for clinical diagnostics. This will both provide a

69   standardized approach to medical diagnostics and present structured machine readable biomedical

70   data for the development of novel computational methods. Apart from phenotype-disease

71 associations, which is the main aim of the HPO project, HPO also provides phenotype-gene
72 associations by using the known rare disease - gene relations (i.e., the information which is in the
73 form of: "certain mutation(s) in *Gene X* causes the hereditary *Disease Y*"), directly using the
74 abovementioned disease centric resources (e.g., Orphanet and OMIM). The disease-gene
75 associations in the source databases are produced by expert curation from the publications of
76 clinical molecular studies. The associations between HPO terms and biomolecules, together with
77 the downstream analysis of these associations, help in disease gene identification and prioritization
78 (Köhler *et al.*, 2009). With the mapping of phenotypes to human genes, HPO currently (January
79 2018) provides 122,166 annotations between 3,698 human genes and 6,729 HPO terms.

80 The Gene Ontology (GO) is an ontological system to define gene/protein attributes with an
81 extensive controlled vocabulary (GO Consortium, 2014). Each GO term defines a unique aspect
82 of biomolecular attributes. Similar to other ontological systems, GO has a directed acyclic graph
83 (DAG) structure, where terms are related to each other mostly with "is_a" or "part_of"
84 relationships. GO is composed of three categories (i.e., aspects) in terms of the type of the defined
85 gene product / protein attribute such as: *(i)* molecular function – MF (i.e., the fundamental function
86 of the protein at the molecular level; e.g., GO:0016887 - ATPase activity), *(ii)* biological process
87 - BP (i.e., the high level process, in which the protein plays a role; e.g., GO:0005975 - carbohydrate
88 metabolic process), and *(iii)* cellular component – CC (i.e., subcellular location, where the protein
89 carries out its intended activity; e.g., GO:0016020 - membrane). Similar to the other ontological
90 systems, the basic way of annotating a gene or protein with a GO term is the manual curation by
91 reviewing the relevant literature. GO also employs the concept of "evidence codes", where all
92 annotations are labelled with descriptions indicating the quality of the source information used for
93 the annotation (e.g., ECO:0000006 - experimental evidence, ECO:0000501 – IEA: evidence used
94 in automatic assertion). UniProt-GOA (Gene Ontology Annotation) database (Huntley *et al.*, 2015)
95 houses an extensive collection of GO annotations for UniProt protein sequence and annotation
96 knowledgebase records. In the UniProtKB/Swiss-Prot database (i.e., housing manually reviewed
97 protein entries with highly reliable annotation) version 2018_02, there are a total of 2,850,015 GO
98 term annotations for 529,941 protein records; whereas in UniProtKB/TrEMBL database (i.e.,
99 housing mostly electronically translated uncharacterized protein entries) version 2018_02, there
100 are a total of 189,560,296 GO term annotations for 67,760,658 protein records. Most of the

101 annotations for the UniProtKB/TrEMBL database entries are produced by automated predictions
102 (UniProt Consortium, 2017).

103 Due to the high volume of experimental research that *(i)* discover new associations between
104 biomolecules and ontological terms, and *(ii)* produce completely new and uncharacterized
105 gene/protein sequences; curation efforts are having hard time in keeping up with the annotation
106 process. To aid manual curation efforts, automated computational methods come into play. These
107 computational methods exploit the approaches and techniques widely used in the fields of data
108 mining, machine learning and statistics, to produce probabilistic associations between biomedical
109 entities. Critical Assessment of Functional Annotation (CAFA) challenge (Radivojac *et al.*, 2013;
110 Jiang *et al.*, 2016) aims to evaluate the automated methods that produce GO and HPO term
111 association predictions for protein entries, on standard temporal hold-out benchmarking datasets.
112 Now after its third instalment, CAFA organization has already brought together a research
113 community, dedicated to elevate the capabilities of automated function prediction approaches
114 closer to the level of expert review.

115 Protein function prediction using GO terms is a highly active area of research, where various types
116 of approaches utilizing: amino acid sequence similarities (Hawkins *et al.*, 2009), 3D structure
117 analysis (Roy, Yang & Zhang, 2012), semantic similarities between the ontological terms (Falda
118 *et al.*, 2012), gene expression profiles (Lan *et al.*, 2013), protein-protein interactions - PPIs (Wass,
119 Barton & Sternberg, 2012), shared functional domains and their arrangements (Fang & Gough,
120 2012; Finn et al, 2016; Doğan *et al.*, 2016) and ensemble approaches that exploit multiple feature
121 types (Wass, Barton & Sternberg, 2012; Cozzetto *et al.*, 2013; Lan *et al.*, 2013; Rifaioglu *et al.*,
122 2017); are employed to model the proteins and to transfer the functional annotations from
123 characterized proteins (i.e., the ones that have reliable annotation), to the uncharacterized ones
124 with highly similar features. Known GO associations of genes and proteins are also used in
125 different contexts in the literature. For example, the method "MedSim" uses the semantic
126 similarities between GO terms for the prioritization of disease genes (Schlicker, Lengauer &
127 Albrecht, 2010). The method "spgk" uses a shortest-path graph kernel to compute functional
128 similarities between gene products using their GO annotations and the term relations on the GO
129 DAG (Alvarez, Qi & Yan, 2011).

130   The automated prediction of the associations between human genes/proteins and
131   phenotype/disease defining ontological terms is also a non-trivial task. The resulting predictions
132   can then be utilized to identify large-scale novel disease-gene-pathway/system relations. The
133   identification of direct disease-gene relations is a widely studied topic (Moreau & Tranchevent,
134   2012). A considerable amount of the existing literature about disease-gene associations involve
135   the calculation of semantic similarities between gene products, based on the already existing
136   ontological term annotations (Washington *et al.*, 2009; Smedley *et al.*, 2013; Deng *et al.*, 2015;
137   Rodríguez-García *et al.*, 2017). For example, the method "PhenomeNET" was employed to
138   generate mappings between the highly related terms across similar ontological systems
139   (Rodríguez-García *et al.*, 2017) such as the HPO, Mammalian Phenotype Ontology – MP (Smith,
140   Goldsmith & Eppig, 2005), Human Disease Ontology – DO (Kibbe *et al.*, 2014) and Orphanet
141   Rare Disease Ontology – ORDO (Vasant *et al.*, 2014); for discovering novel gene-disease
142   associations. However, semantic similarity based approaches sometimes suffers from the low
143   coverage of the HPO annotations on the protein space. The authors of two recent studies have
144   investigated this issue (Kulmanov & Hoehndorf, 2017; Peng *et al.*, 2017). In this context,
145   increasing the coverage of HPO annotations by predicting gene/protein-HPO term associations
146   may help semantic similarity based association studies.

147   There are only a few examples of HPO term-protein association prediction methods in the
148   literature. In the "dcGO" method, the authors mapped ontological terms (including HPO) to protein
149   domains, which are the functional units, and transferred the ontology mapping to proteins
150   according to known domain annotations (Fang & Gough, 2012). The objective in the
151   "PHENOstruct" method is the prediction of gene-HPO term associations using heterogeneous
152   biological data consisting of protein-protein interactions (PPIs), GO annotations, literature
153   relations, variants and known HPO annotations, together with a structured SVM classifier
154   (Kahanda *et al.*, 2015). In this sense, it is interesting see another method that utilize already existing
155   GO annotations to predict HPO terms. Nevertheless, according to the authors, the contribution of
156   the GO features to the overall predictive performance of PHENOstruct was minimal. A text mining
157   based method "EVEX", was employed for protein-HPO term association prediction. Originally,
158   EVEX utilizes text mining approaches for large-scale integration of heterogeneous biological data
159   and event extraction to generate a structured resource of relations, to be used in pathway curation
160   (Van Landeghem *et al.*, 2013). In the context of HPO term prediction, EVEX scans the literature

161   to detect proteins and phenotypic terms that co-occur on the same text corpus, and associates them
162   with each other based on certain criteria. According to the CAFA2 challenge results (Jiang *et al.*,
163   2016), the participating method EVEX was the top performer. A network based HPO prediction
164   method was the "RANKS", in which the authors developed a flexible algorithmic scheme for
165   heterogeneous biological network analysis, and used previously generated functional Interaction
166   and functional human gene networks for gene-HPO term association prediction (Valentini *et al.*,
167   2016). In a recent study, the authors proposed two hierarchical ensemble methods: *(i)* the
168   Hierarchical Top-Down, and *(ii)* the True Path Rule, for gene-HPO term associations; in which
169   the hierarchical graph structure of HPO has been utilized together with the RANKS algorithm and
170   the SVM classifier (Notaro, *et al.*, 2017).

171   The text mining approach is highly effective for predicting gene-disease relations in disease gene
172   prioritization studies (Krallinger, Valencia & Hirschman, 2008). However, this approach suffers
173   from low coverage in some cases, due to knowledge limitation in the literature. If a certain
174   abnormality and a gene/protein has not been studied together in the same framework yet, it is often
175   not possible to identify the relation. Network based methods are proposed on top of either text-
176   mining results, protein-protein interactions and/or pathway data (Bromberg, 2013; Guney & Oliva,
177   2014; Guala & Sonnhammer, 2017) to detect indirect relations, which greatly increased the
178   coverage; nevertheless, they still moderately rely on the previously reported relations. It is also
179   important to note that, any predictive approach is limited by the quality and the coverage of its
180   source information. However, the predictive output of different approaches often complement each
181   other, contributing to fill different portions of the missing information in the knowledge space.
182   Due to this reason, developing novel approaches to complement the conventional methods is
183   crucial for the automated ontological association prediction. The observed low performance of
184   even the best methods in the HPO term prediction track of the CAFA2 challenge displayed the
185   necessity of novel approaches for the biomedical entity relation prediction.

186   In this study, a new approach "HPO2GO" is proposed to produce phenotypic abnormality HPO
187   term associations to both GO terms and human genes/proteins, with the analysis of co-annotation
188   fractions between the HPO and GO term combinations. For this, HPO and GO terms that are
189   continually co-occurring on different genes/proteins as annotations, are linked to each other (i.e.,
190   the system training step), entitled as the "HPO2GO mappings". After that, proteins with a linked
191   GO term annotation receives the corresponding HPO term as the phenotypic term prediction (i.e.,

192 the application step), entitled as the "HPO2protein predictions". The idea here is to associate a

193 HPO term Y with a GO term X in the sense that: "if a protein loses its function defined by the GO

194 term X (or at least a reduction in the defined functionality) as a result of a genetic mutation, the

195 loss of function may cause the disease, which is defined by the phenotype term Y". This idea is

196 based on the nature of annotating genes/proteins with HPO terms; for example, only the

197 functionally perturbed variants of these genes/proteins (e.g., disease causing variants) are

198 associated with the relevant genetic diseases and their defining phenotypic abnormality terms.

199 Mutations often lead to diseases by causing either a loss of the existing functionality or a gain of

200 new functionality in the gene products. As a result, if the HPO term Y and the GO term X are

201 observed to be frequently co-occurring on different proteins, then the lost function, which gave

202 way to the corresponding disease may be the one defined by the GO term X. This logic would

203 make biological sense especially when the corresponding function is a large-scale biological

204 process. This approach exploits the significantly higher coverage of GO term annotations for

205 genes/proteins, compared to the HPO term annotations; to produce novel gene/protein - HPO term

206 associations.

207 In order to test the biological relevance of this approach, selected HPO2GO mappings were

208 manually examined. Additionally, the proposed methodology was employed to predict HPO terms

209 for the human protein target dataset provided in the CAFA2 challenge. Using the benchmark set,

210 the prediction performance was calculated and compared with the state-of-the-art HPO prediction

211 methods. Another set of HPO2GO mappings were generated for this test, using the temporal hold-

212 out training data provided in CAFA2. Finally, the up-to-date HPO2GO mappings were employed

213 to generate HPO term predictions to human protein entries in the UniProtKB/Swiss-Prot database

214 (i.e., HPO2protein predictions). The training and test datasets, along with the source code of the

215 proposed methodology and the analyses are available for download at

216 https://github.com/cansyl/HPO2GO.

217   **2. METHODS**

218   **Dataset Construction**

219   In order to generate the training sets; first, gene to HPO term association file was downloaded from

220   the   HPO   web-site   (January   2017   version   of   the   file   named:

221   "ALL_SOURCES_ALL_FREQUENCIES_ genes_ to_phenotype.txt"). This file contained

222   153,575 annotations between 3,526 human genes and 6,018 HPO terms. This file is shared in the

223   HPO2GO   repository   with   the   filename:

224   "HPO_gene_to_phenotype_annotation_01_2017_ALL_SOURCES_ALL_FREQUENCIES.txt".

225   These gene-phenotype associations are generated as a part of the HPO project by merging two

226   source files: first, the gene-disease associations provided in the incorporated disease centric

227   databases (e.g., OMIM and Orphanet); and second, the disease-phenotype associations generated

228   by HPO. In HPO, "genes_to_phenotype" file only contains the asserted (i.e., specific) annotations

229   to genes; whereas "phenotype_to_genes" file contains all annotations propagated through the root

230   of the HPO DAG, according to the true path rule. As a result, parents of the asserted terms are

231   included as well. In this study, the asserted annotations are used (in terms of HPO) at the input

232   level, in order to avoid the propagation of potential false positive annotations that may be presented

233   in the source dataset. This application usually comes with the cost of increased number of false

234   negatives; however, avoiding the potential false positives is considered more important here.

235   In order to generate the second training dataset, all GO term annotations to the human proteins in

236   UniProtKB with the manually assigned (curator assigned) evidence codes (i.e., EXP, IDA, IPI,

237   IMP, IGI, IEP, IBA, IC, IKR, ISS, NAS, ND and TAS) were downloaded from the UniProt-GOA

238   database   2017_01   version,   using   the   QuickGO   browser   (filename:

239   "GOA_UniProt_human_protein_annotation.tsv"). The reason behind not using electronically

240   assigned annotations (i.e., evidence code: IEA) was that, these annotations have a reduced level of

241   annotation reliability compared to the curator assigned ones; as a result, they may contain

242   erroneous cases (i.e., false positives). After eliminating the repeating (i.e., redundant) annotations,

243   the finalized file contained 179,651 GO annotations between 18,577 unique human genes and

244   14,632   GO   terms   (filename   of   the   finalized   GO   annotation   file:

245   "GO_annot_human_proteins_UniProtGOA_01_2017.txt"). An additional column containing the

246   corresponding HGNC symbols (i.e., gene symbols) of the coding genes was also included in the

247   downloaded GO annotation file. This column was later used to combine the GO annotations with

248   the HPO annotations, since the HPO annotation file includes the gene symbols.

249

Applied Methodology

251   The proposed methodology is divided into 2 steps: *(i)* training of the system (i.e., the generation

252   of the HPO2GO mappings), and *(ii)* the application step (i.e., the prediction of HPO term-protein

253   associations – HPO2protein, using the previously generated HPO2GO mappings).

254   Figure 1 represents the complete HPO2GO mapping (i.e., training) procedure. For the training of

255   the system, first, the HPO and GO annotation datasets were prepared (Figure 1.A and Figure 1.B)

256   and the initial HPO-GO mappings were generated (Figure 1.C) by identifying the genes/proteins

257   shared between individual HPO and GO terms (i.e., the cases where HPO and GO terms were co-

258   annotated to the same genes/proteins). This mapping has generated 1,433,208 unique pairs

259   between 6,005 HPO terms and 9,685 GO terms. At this point, it was observed that some of GO

260   and HPO terms were annotated to high number of proteins, and it was highly probable for them to

261   co-occur on the same protein once or twice just by chance. In order to eliminate these cases, a

262   filtering procedure was required. For each HPO-GO term pair, a co-occurrence similarity measure,

263   inspired from an information-theoretic definition of similarity (Lin, 1998), has been calculated.

264   The co-occurrence similarity formulation is given in Equation 1.

265

$$S_{HPOi,GOj} = \frac{2 * N_{G\,HPOi\&GOj}}{N_{G\,HPOi} + N_{G\,GOj}} \tag{1}$$

267

268   Here, $S_{HPOi,GOj}$ is the co-occurrence similarity between the HPO term "*HPOi*" and the GO term

269   "*GOj*", $N_{G\,HPOi\&GOj}$ is the number of genes/proteins where these terms are annotated together, N_G

270   $_{HPOi}$ is the total number of genes with the annotation "*HPOi*", and $N_{G\,GOi}$ is the total number of

271   genes with the annotation "*GOi*".

272   The mapping process and the co-occurrence similarity calculation are shown in Figure 2 with a

273   toy example. Following the calculation of the co-occurrence similarities between all HPO-GO

274   pairs, a thresholding operation was applied in order to distinguish between relevant mappings and

275   the random ones. Two parameters were used for the thresholding operation: *(i)* the co-occurrence

276   similarities: S, and *(ii)* the number of genes with co-occurring annotations: n. The aim behind

277   employing a second parameter (i.e., n) was to eliminate the potential random pairing cases, where

278   the co-occurrence similarity is still high. These cases are rare; however, it is still possible to

279   observe a few of them especially when n is very small, due to extremely high number of term

280   combinations. In Figure 2, this situation is represented on the toy example, here $S_{HPOD,GO4}$ is equal

281   to $S_{HPOB,GO3}$; however the HPO_D-GO_4 mapping is probably less reliable compared to HPO_B-GO_3

282   since $n_{HPOD,GO4}$ is equal to 1.

283   Statistical resampling was used to determine the optimal parameter values (to be used as

284   thresholds), that separate meaningful mappings from random ones. A permutation (i.e.,

285   randomization) test was constructed for this purpose. At this point, a randomized HPO-GO term

286   mapping table was required, which was generated (Figure 1.D) by first, shuffling the indices of

287   the original "HPO vs. gene" and "GO vs. gene" annotation tables; and second, calculating both the

288   randomized co-occurrence similarities (i.e., S_R) and the number of genes with co-occurring

289   annotations (i.e., n_R) for each random HPO-GO mapping. For each arbitrarily selected S (i.e., $S >$

290   0, $S \geq 0.1$, $S \geq 0.2$, …, $S \geq 0.6$) and n (i.e., $n \geq 1$, $n \geq 2$, …, $n \geq 5$) threshold value combination, the

291   original GO-HPO mappings with lower than the threshold S and n values were deleted and a co-

292   occurrence similarity distribution histogram was plotted using the remaining mappings (i.e.,

293   histograms plots in Figure 1 and in Figure 3). The same procedure was applied for the randomized

294   mapping set as well. Finally, Kolmogorov–Smirnov test -KS test- (Lilliefors, 1967; Hollander,

295   Wolfe & Chicken, 2013) is employed to calculate a test statistic for estimating whether the samples

296   from the random and the original sets (at each S and n selection) are from the same distribution or

297   not. KS is a nonparametric test of 1-dimensional probability distributions that can be used to

298   compare two samples, considering the quantized distance between the samples. The null

299   hypothesis states that the two samples are drawn from the same distribution. Here, the distribution

300   (i.e., histogram) of the S values for the original and the randomized mapping sets represent the two

301   samples. The reason behind using histograms instead of the actual S values was that, both high and

302   low S values were presented in both distributions; as a result, the significance test by checking

303   sample distances approach would not work. However, the frequencies of these high and low S

304   values are different from each other in the original and the random distributions. If the null

305   hypothesis is accepted at a selected threshold value pair (S and n), which means that the

306   distributions are not statistically different from each other, then it is concluded that the selected

307   thresholds failed to eliminate the random pairings in the original mapping (i.e., a higher threshold

308   is required). The lowest threshold values, where the samples from the two distributions became

309   significantly different from each other, were selected as the official thresholds. Excessive threshold

310   values were not considered in order not to eliminate too many GO-HPO mappings. After the

311   determination of the parameter values (i.e., S and n thresholds), the HPO2GO mappings were

312   finalized, which ended the training process.

313   An alternative to the statistical resampling methodology used here would be employing a direct

314   approach such as an exact test (e.g., Fisher's exact test), and to calculate a significance value for

315   each HPO-GO mapping individually. The advantage of this approach would be accepting or

316   rejecting the null hypotheses specific to each mapping (as opposed to finding a global significance

317   using the whole data at once, as applied in this study). The main disadvantage of this approach

318   would be the diminished statistical power to calculate significance values for the mappings where

319   the number of instances are too low. Unfortunately, this is the case for most of the HPO-GO

320   mappings as the mean number of occurrences (for all HPO-GO mappings) in the original or in the

321   random sets were 1.77 and 1.48, respectively. As a result, the direct approach was not considered

322   in this study.

323   HPO2protein prediction step was a simple procedure, where query proteins were annotated with

324   the HPO terms, by taking their already existing GO annotations into account. HPO2GO mappings

325   were employed for this purpose. There were a total of 3 application runs in this study using: *(i)*

326   CAFA2 targets as the query set (for the performance tests and for the comparison with the state-

327   of-the-art), *(ii)* CAFA3 targets as the query set (to officially participate in the CAFA3 challenge),

328   and *(iii)* all human protein entries in the UniProtKB/Swiss-Prot database (to generate the

329   HPO2protein predictions).

330

331   **Performance Evaluation Metrics**

332   In this study, it was not possible to use a standard fold based cross-validation to measure the

333   performance and to determine the parameter values in the training procedure, since in most cases,

334   the number of genes/proteins that have a co-occurring HPO-GO term annotations were extremely

335   low. As a result, it was not impossible to separate the samples into training and validation sets.

336   Instead, the optimal parameter values were determined by using statistical resampling. However,

337   a performance test was still required in order to assess the success of the proposed approach. For

338   this, CAFA2 challenge benchmark set was employed. Due to the fact that CAFA2 challenge was

339   long before the analysis done in this study, HPO2GO mappings were re-generated using both the

340   GO and HPO annotation data from January 2014. This was followed by the production of the

341   HPO-protein association predictions on the CAFA2 target gene set. This analysis both served as a

342   performance test with the temporal hold-out data (one of the hardest and most informative tests

343   for predictive models) and a performance comparison with the state-of-the-art (i.e., other HPO

344   prediction methods participated in CAFA2). The most basic definitions of the evaluation metrics

345   used in this test; *recall, precision, Fmax* and minimum semantic distance (*Smin*) are shown in

346   Equation 2, 3, 4 and 5.

347

348
$$Rc_{\tau i} = \frac{TP_{\tau i}}{TP_{\tau i} + FN_{\tau i}} \qquad (2)$$

349

350
$$Pr_{\tau i} = \frac{TP_{\tau i}}{TP_{\tau i} + FP_{\tau i}} \qquad (3)$$

351

352
$$F_{max} = \max_{i=1...N} \left\{ \frac{2 * Pr_{\tau i} * Rc_{\tau i}}{Pr_{\tau i} + Rc_{\tau i}} \right\} \qquad (4)$$

353

354
$$S_{min} = \min_{i=1...N} \left\{ \sqrt{Ru_{\tau i}^2 + Mi_{\tau i}^2} \right\} \qquad (5)$$

355

356   In equations 2,3 and 4; $TP_{\tau i}$, $FN_{\tau i}$, $FP_{\tau i}$, $Rc_{\tau i}$ and $Pr_{\tau i}$ represent the number of true positives, the

357   number of false negatives, the number of false positives, *recall* and *precision* values, respectively;

358   at the ith probabilistic score threshold (τ_i). *Fmax* correspond to the maximum of the *F-score* values

359   (i.e., harmonic mean of *precision* and *recall*, shown inside the curly brackets in Equation 4)

360   calculated for each arbitrarily selected probabilistic score threshold. i=1...N represents there are N

361   different arbitrarily selected probabilistic score thresholds. Higher *Fmax* values indicate higher

362   performance. $Ru_{\tau i}$ and $Mi_{\tau i}$ in Equation 5 corresponds to remaining uncertainty and normalized

363   misinformation at the ith probabilistic score threshold (τ_i), respectively. *Smin* is the minimum

364   semantic distance. Lower *Smin* indicate higher performance. Also, a weighted version of the *Fmax*

365   measure has been calculated. Weighting procedure was applied using the information content of

366 each ontology term; so that, more informative terms obtained higher weights. Information
367 regarding the calculation of remaining uncertainty, misinformation and term based information
368 contents can be found in Clark and Radivojac, 2013 and in Jiang *et al.*, 2016.

369 In the proposed method, probabilistic scores for each HPO-protein association prediction is
370 calculated using the term co-occurrence similarity scores in Equation 1. If the mapping between
371 the terms *HPOi* and *GOj* received the co-occurrence similarity score $S_{HPOi,GOj}$, then all proteins
372 that receive the *HPOi* prediction due to the presence of *GOj* annotation obtain the probabilistic
373 prediction score: $S_{HPOi,GOj}$. The calculation of the score in Equation 1 is set to range between 0
374 and 1; as a result, it can directly be used as a probabilistic score. Apart from that, probabilistic
375 score thresholds represent values, under which the predictions are discarded. This way, a different
376 set of predictions are given for each arbitrarily selected probabilistic score thresholds, leading to
377 different precision and recall values. It is important to note that, probabilistic score thresholds are
378 different from the thresholds we used to filter out unreliable HPO2GO mappings during the
379 training process. The probabilistic score thresholds are used here (i.e., after the production of
380 HPO2protein predictions) to produce binary predictions from continuous prediction scores, to be
381 able to calculate performances. More details regarding the CAFA2 evaluation metrics are given in
382 Jiang *et al.*, 2016.

383

384 **3. RESULTS**

385 **Statistical Analysis of the Mappings**

386 The initial HPO to GO mappings were generated according to the procedure explained in the
387 Methods section (Figure 2). The initial mapping of the original set resulted in 1,433,208 mappings
388 between 6,005 HPO terms and 9,685 GO terms. The same procedure for the randomized set
389 produced 1,543,917 mappings between 5,995 HPO terms and 9,685 GO terms. The initial HPO-
390 GO mappings for both the original and the randomized sets are available for download in the
391 repository of the study (respective filenames: "HPO_GO_Raw_Original_Mapping.txt" and
392 "HPO_GO_Random_Mapping.txt"). It was expected that the mappings generated from the random
393 set would have lower co-occurrence similarity values on average compared to the original set
394 mappings; in other words, they would contain less number of mappings for a particular co-
395 occurrence similarity value. Table 1 displays the comparison of the number of mappings for
396 different co-occurrence similarity values, between the original and the randomized sets. As
397 observed from Table 1, when $S > 0$ there is no difference between the mappings; however as S is
398 increased, the difference between the mappings becomes clear. Also, when S is increased, the
399 number of mapped HPO and GO terms were decreased since many terms did not have any
400 mappings that satisfied the stringent S values. The parameter n was not taken into account while
401 calculating the statistics in Table 1 (i.e., $n \geq 1$ for all values in the table).

402 The histograms in Figure 3 display the co-occurrence similarity distributions (i.e., S) for arbitrarily
403 selected n values. As observed from the histograms, when the mappings with low n values are
404 eliminated, the distributions shift to the right (i.e., the mean of S increases), which can be
405 interpreted as the mappings became more reliable. However, excessive values of n thresholds leave
406 only a few mappings to work with, especially at $n=25$ and $n=75$ (please see the number of
407 mappings at the vertical axis of Figure 3.C and D). Histograms in Figure 3 also show that
408 thresholding the mappings using only n (not using S at all) would not be sufficient because there
409 are mappings with very low S values even at very high n thresholds (i.e., 25 and 75). This
410 observation verified the decision to use both of the parameters for the filtering operation. At this
411 point, the statistical resampling (i.e., KS test) was applied since it was not possible to determine
412 the optimal n threshold by just manually checking the histograms.

413   In order to find the minimum S and n values that significantly separate the original mapping from

414   the randomized mapping, 35 different distributions, all combinations of the selected n (i.e., $n \geq 1$,

415   2, ..., 5) and S (i.e., $S > 0$, $S \geq 0.1$, ..., 0.6) values, were prepared and tested individually against

416   the co-occurrence distribution of the random mapping, generated with the same S and n thresholds.

417   This test resulted in 35 different p-value calculations and the minimum parameter values that

418   satisfied the statistical significance (i.e., rejection of the null hypothesis, which states that the two

419   samples are from the same distribution) were selected. Table 2 displays the significance results of

420   all KS tests. The cells with "NaN" indicate the cases, where the test could not be completed due

421   insufficient number of samples to calculate the statistic. However, incomplete tests were not a

422   problem since the aim here was observing the minimum threshold values, where the distributions

423   significantly diverge from each other (NaNs are located far away from this point). In Table 2, the

424   cell with the p-value written in bold font (i.e., 0.0057) signifies the point, where the corresponding

425   thresholds $n \geq 2$ and $S \geq 0.1$ yielded the required significance (p-value < 0.01); and thus, these

426   values were selected as the finalized thresholds. This means that, all of the mappings with $n < 2$

427   and $S < 0.1$ were considered unreliable and eliminated from the initial HPO-GO mappings.

428   Figure 4 displays the total number of unique mappings (vertical axis) with co-occurrence similarity

429   values greater than the corresponding threshold value (horizontal axis), for the original and the

430   randomized distributions on the blue and red coloured curves, respectively. Figure 4.A shows the

431   plot for the combination with greater than or equal to one co-annotated gene (i.e., $n \geq 1$), Figure

432   4.B displays the same value for $n \geq 2$, Figure 4.C and D for $n \geq 3$ and 4; respectively. The

433   differences between Figure 3 and Figure 4 is that, *(i)* in Figure 4 cumulative number of mappings

434   are given (i.e., all mappings left after thresholding with $S \geq 0.1$, 0.2, ...), whereas in Figure 3, the

435   number of mappings that fall into each S bin is given; and *(ii)* in Figure 4, plots are given for $n \geq$

436   1, 2, 3 and 4 since the aim was to display the curves around the selected threshold n value; whereas

437   in Figure 3, there are plots for $n \geq 1$, 5, 25 and 75 to visually indicate the distribution shifts

438   especially at high n values (i.e., $n=25$ and $n=75$). Figure 4 was drawn as a visual representation of

439   the likeness between the original and the randomized distributions at different parameter

440   selections. As observed from Figure 4, the distributions diverged from each other at $n \geq 2$, which

441   also is consistent with the KS test results. Considering the co-occurrence similarity parameter, $S \geq$

442   0.1 produced a clear separation between the original and the randomized distributions as long as n

443   is greater than 1. Following the HPO-GO mapping elimination according to the selected

444   thresholds, finalized HPO2GO mappings contained 45,805 associations between 3,693 HPO terms

445   and 2,801 GO terms. HPO2GO mappings are available for download in the repository of the study

446   (filename: "HPO2GO_Finalized_Mapping.txt").

447   It was only possible to use a small portion of the input GO annotations for the generation of the

448   HPO2GO mappings because the number of HPO annotated genes were only 3,526; whereas, the

449   number of GO annotated human genes were 18,577. Since mappings can be done over the

450   genes/proteins with co-occurring GO and HPO annotations, only 3,526 genes/proteins were used

451   in the process. The remaining 15,051 human genes with GO annotations were only used in the

452   application step (i.e., HPO2protein), to predict HPO term associations.

453

454   **Comparison of HPO2GO with the manual HPO-GO associations**

455   As a part of the main HPO project, a sub-set of the HPO terms has been mapped to the relevant

456   terms from different ontologies (e.g., anatomy, Gene Ontology process or cell type) to yield

457   semantic interoperability with these systems. Finalized HPO2GO mappings were compared with

458   these manual GO associations in order to observe the correspondence. The manual mappings are

459   available in the "hp.owl" ontology file in the HPO repository. From the most up to date version

460   (03_2018) of the owl file, the corresponding associations have been extracted and stored in the

461   HPO2GO repository with the filename: "HPO_manual_GO_associations_03_2018.txt". There

462   were 489 manual associations between 488 HPO terms and 239 GO terms, which was significantly

463   lower compared to the HPO2GO mappings (i.e., 45,805 mappings between 3,693 HPO terms and

464   2,801 GO terms). Considering the most frequent GO terms in these manual associations, 71 HPO

465   terms were associated with the GO term "ossification" (GO:0001503), 51 HPO terms were

466   associated with the GO term "inflammatory response" (GO:0006954) and 37 HPO terms were

467   associated with the GO term "pigmentation" (GO:0043473). The comparison of manual

468   associations with HPO2GO mappings have revealed that nearly 40% of the HPO terms in the

469   manual associations were also mapped to at least one GO term in HPO2GO. When the same

470   calculation was done to reveal how many of the HPO terms in HPO2GO were also in the manual

471   associations, the result was only 5%. The same correspondence results for GO terms were 35%

472   and 3%, respectively. Finally, the correspondence between the actual HPO-GO mappings has been

473   calculated. Only 23 out of 489 manual associations were retrieved by HPO2GO. An inspection

474   was done to reveal the possible reasons behind the low correspondence, and it was found that, the

475   low number of gene annotations of either the corresponding HPO term or the GO term was the

476   main reason. However, in many cases, an ancestor of the corresponding HPO or GO term was able

477   to be mapped. For example, there was a manual association between "epididymitis: the presence

478   of inflammation of the epididymis" (HP:0000031) and "inflammatory response" (GO:0006954).

479   HPO2GO managed to retrieve an association between "epididymitis" (HP:0000031) and "defense

480   response" (GO:0006952), which is the direct parent term (with "is_a" relationship) of

481   "inflammatory response" (GO:0006954). The "inflammatory response" GO term could not be

482   mapped to any HPO term by HPO2GO since this term was not directly annotated to any

483   gene/protein in UniProt-GOA.

484

485   **Performance Comparison with the State-of-the-art**

486   The test for the comparison with the state-of-the-art had two objectives: *(i)* measuring the

487   performance of the method on a temporal hold-out dataset to observe the relevance of the proposed

488   approach, and *(ii)* investigating how the proposed method competes with the best performing

489   methods in the literature. For this, we have re-generated the HPO2GO mappings using the CAFA2

490   training set, which contained 133,175 annotations between 5,586 HPO terms and 4,418 proteins,

491   from January 2014. Whereas, CAFA2 evaluation set (i.e., the benchmarking set) contained 19,743

492   annotations between 1,845 HPO terms and 238 proteins (considering only the no-knowledge

493   benchmark samples). The reason behind the presence of low number of annotations (and proteins)

494   in the evaluation set was that, only the HPO annotations produced between the time of the

495   challenge participation deadline and the end of the annotation collection period (a total duration of

496   nearly 8 months) were used to generate the temporal hold-out evaluation set. One important

497   observation about the benchmark set is that, a few HPO terms dominates the benchmark set (i.e.,

498   the most frequently annotated 185 terms own 55% of all annotations). Also, most of these HPO

499   terms were generic (e.g., HP:0000707 - Abnormality of the nervous system). The small size of the

500   benchmark dataset, together with the uneven distribution of the term frequencies limits the

501   evaluative capacity of this set. All of the datasets, the source code and the supplementary files used

502   in the CAFA2 challenge, and thus in this benchmarking experiment, is available through the CAFA

503   project   repositories   (URLs: https://github.com/yuxjiang/CAFA2   and

504   https://ndownloader.figshare.com/files/3658395). The CAFA performance evaluation scripts

505   published in these addresses were directly used to calculate the predictive performance of

506   HPO2GO.

507   HPO2GO mappings generated using the CAFA2 training set contained 27,424 mappings between

508   2,640 HPO terms and 2,488 GO terms. Considering the whole CAFA2 human target protein set,

509   this mapping produced 1,922,333 HPO predictions for 16,256 proteins and 2,640 HPO terms. The

510   calculated performance of this prediction set was low (*Fmax* = 0.30), mainly due to high number

511   of false positive (FP) hits. However, it is probable that many of these false positives were actually

512   non-documented HPO annotations of the corresponding protein, as the benchmark annotation set

513   is incomplete. Increasing the thresholds with the aim of reducing the number of false positives

514   resulted in a matching increase in the number of false negatives (FN), with a similar *Fmax* value.

515   With the aim of enriching the mappings (to be able to reduce FPs without a significant increase in

516   FNs), HPO annotations of genes from January 2014 (i.e., the CAFA2 training set) were propagated

517   to the root of HPO DAG according to the true path rule. The propagated training set contained

518   379,513 annotations between 4,418 human proteins and 6,576 HPO terms; as opposed to 133,175

519   annotations between 4,418 human proteins and 5,586 HPO terms in the asserted CAFA2 set. As

520   observed from the dataset statistics, propagating the annotations have only added about one

521   thousand new terms to the set; however, the number of annotations were significantly increased.

522   Repeating the CAFA2 benchmark analysis using propagated HPO annotations and the same GO

523   annotations set resulted in the same performance (*Fmax* = 0.30). Next, automated GO annotations

524   (i.e., evidence code: IEA) have been included in the source GO annotation set, which increased

525   the number of unique GO annotations from 128,947 to 214,235 (a 66% increase). Using the

526   propagated HPO annotations together with enlarged GO annotation set, the new HPO-GO

527   mappings, namely "HPOprop2GOall", were generated. The finalized HPOprop2GOall contained

528   198,928 mappings between 4,780 HPO terms and 5,196 GO terms; as opposed to 27,424 mappings

529   between 2,640 HPO terms and 2,488 GO terms in the original mappings. The drastic difference

530   between the numbers have indicated the enrichment provided by annotation propagation and GO

531   set enlargement. Subsequently, HPOprop2GOall mappings were used to predict HPO associations

532   for all CAFA2 targets, producing 13,022,574 predictions (as opposed to 1,922,333 predictions

533   with the asserted set). Considering only the CAFA2 benchmark proteins, the predictions generated

534   by using the optimized parameters (i.e., $n = 170$ and $S = 0.11$) resulted in 34,486 HPO predictions

535   for 221 benchmark proteins and 235 HPO terms, with a performance of *Fmax* = 0.35 (no-
536   knowledge benchmark sequences in the full evaluation mode), which is among the top
537   performances considering all of the models from 38 participating groups in the CAFA2 HPO
538   prediction track. The *Fmax* performance of the top model in the challenge was 0.36 (Jiang *et al.*,
539   2016), and the performance of the naïve baseline classifier was also the same. In Figure 5, each
540   bar displays the overall performance (*Fmax*) of the CAFA2 participators, baseline classifiers and
541   HPO2GO. Additionally, weighted precision-recall curves were plotted, to assess the performance
542   of the method at different threshold selections (Figure 6). The weighting procedure was done
543   according to the information content of each HPO term; as a result, informative terms received
544   higher weights. The term weights were officially calculated and published by CAFA2 challenge
545   evaluators and these weights were directly used in this study. As shown by the black curve in
546   Figure 6, HPO2GO performed the same as the best methods on the optimal point (*wFmax* = 0.29),
547   indicated by the circular marking over the curve. Furthermore, the minimum semantic distance
548   (*Smin*) was calculated for HPO2GO and compared again with the CAFA2 participators (Figure 7).
549   Here, lower *Smin* values indicate higher performance. As shown in Figure 7, HPO2GO performed
550   slightly worse (*Smin* = 57.2) compared to the top performing methods and the naïve classifier.
551   Finally, term-centric predictive performance of HPO2GO was measured. In the term-centric
552   evaluation, the performance of a predictor is measured independently for each ontology term.
553   Figure 8 displays the average HPO term-centric area under the ROC curve (AUROC) measures
554   for HPO2GO and the CAFA2 participating methods. In this evaluation mode, HPO2GO came
555   second (AUROC = 0.59), with a significantly better result compared to the baseline classifiers.
556   HPO2GO CAFA2 benchmark predictions are available in the repository of the study (filename:
557   "HPO_CAFA2_benchmark_predictions.txt").

558   The coverage of HPO2GO on the CAFA2 benchmark protein set was nearly 97%. The high
559   coverage indicates that HPO2GO managed to annotate a wide range of proteins. According to the
560   analysis of 6 genes, which codes for the proteins in the CAFA2 benchmark set that HPO2GO could
561   not annotate (gene symbols: *ATOH7*, *DMP4*, *GNT2C*, *CE126*, *PGAP3*, *SERAC1*), only two of
562   these proteins have GO annotations (gene symbols: *ATOH7*, *PGAP3*), and the others had no GO
563   annotations at all with experimental evidence codes; as a result, it was impossible for HPO2GO to
564   assign HPO terms for those proteins. Considering *ATOH7* and *PGAP3*, their annotated GO terms
565   (GO:0003407 - neural retina development and GO:0021554 - optic nerve development for *ATOH7*,

566   GO:0016788 - hydrolase activity, acting on ester bonds, GO:0006505 - GPI anchor metabolic

567   process and GO:0031227 - intrinsic component of endoplasmic reticulum membrane for *PGAP3*)

568   have been associated with HPO terms in the raw HPO-GO term mappings; however, all of these

569   mappings were eliminated at the resampling step due to low *n* and *S* values.

570   At this point in the study, CAFA2 benchmark performance test was also repeated using the raw

571   HPO-GO mappings (without statistical resampling) in order to observe the impact of eliminating

572   unreliable term mappings using *n* and *S* thresholds. The raw mappings were composed of 879,873

573   HPO-GO term associations, which led to 66,522,438 predictions between 18,155 proteins and

574   5,559 HPO terms. The statistics indicate that nearly 65% all possible combinations between the

575   target proteins and HPO terms were produced as predictions. The performance analysis of this

576   prediction set resulted in a very low *Fmax* value (i.e., 0.001), as expected, indicating the

577   effectiveness of the statistical resampling procedure applied in this study. At all stages of the

578   performance analysis, different HPO2GO mapping sets were generated using various resampling

579   parameters (i.e., different *n* and *S* threshold selections), and tested on the CAFA2 benchmark;

580   however, these mappings produced performances slightly inferior to the ones reported above. The

581   most probable reason behind observing reduced performance with a lowered threshold was the

582   inclusion of high number of mappings in the finalized set, most of which were false positives. This

583   in turn provided a reduced precision and a reduced *Fmax* measure. On the other hand, when

584   excessively high thresholds were selected, many of the reliable mappings were probably discarded,

585   leading to high number of false negatives, thus a reduced recall and a reduced *Fmax*, as well.

586

587   **Generation of the Finalized HPO2protein Predictions**

588   Up-to-date HPO2GO mappings were employed to predict HPO terms for the human protein entries

589   in the UniProtKB/Swiss-Prot database (i.e., 20,258 protein records), and the resulting prediction

590   set was marked as the finalized HPO2protein predictions. This set contained 3,468,582 HPO

591   predictions for 18,101 proteins and 3,693 HPO terms. HPO2protein predictions are available in

592   the repository of the study (filename: "HPO2protein_Predictions.txt").

593   Finally, up-to-date HPO2GO model was run on the CAFA3 human protein targets, which produced

594   3,453,130 predictions on 16,609 human proteins with 3,719 HPO terms. A more stringent subset

595   of this prediction set (i.e., predictions produced from mappings with $S \geq 0.2$) has been officially

596 submitted to the CAFA3 challenge. HPO2GO CAFA3 target predictions are available in the
597 repository of the study (filename: "HPO_CAFA3_target_predictions.txt"). There was a small
598 difference between the number of query proteins in HPO2protein and the CAFA3 target sets
599 (20,258 as opposed to 20,197, respectively). At the time of writing this manuscript, the CAFA3
600 challenge results have not been announced yet.

601

602 **The Biological Relevance of the Selected HPO2GO Mappings – A Case Study**

603 In order to discuss the biological relevance of HPO2GO mappings, selected HPO-GO term
604 mappings were examined. For this purpose, 3 confidence bins (high-level, mid-level and low-level
605 reliabilities) were determined, considering their respective S and n values (high S and n values
606 together indicate elevated reliability). Six example mappings (2 from each confidence bin) were
607 randomly selected for the case study. The first case was a highly reliable mapping between the
608 phenotypic abnormality HPO term "absence of bactericidal oxidative respiratory burst in
609 phagocytes" (HP:0002723) and the GO term "respiratory burst" (GO:0045730), which is in the BP
610 category. The exact definition of this GO term in the UniProt-GOA database is: "*A phase of
611 elevated metabolic activity, during which oxygen consumption increases; this leads to the
612 production, by an NADH dependent system, of hydrogen peroxide (H2O2), superoxide anions and
613 hydroxyl radicals.*" (URL: https://www.ebi.ac.uk/QuickGO/term/GO:0045730). These two terms
614 were mapped to each other in HPO2GO with high confidence (i.e., $S = 0.89$ and $n = 4$). The
615 symbols of the co-annotated genes were *CYBA, CYBB, NCF2* and *NCF1*. As observed from the
616 names of both terms and from the description of the GO term, the HPO term defines an abnormal
617 condition that corresponds to the absence of the biological process portrayed by the mapped GO
618 term. This is in accordance with the logic behind mapping HPO terms with GO terms, which stated
619 the occurrence of an abnormality (i.e., the HPO term) due to the loss of the biomolecular function
620 defined by the mapped GO term. In addition, there is a GO term named "respiratory burst after
621 phagocytosis" (GO:0045728), which is related (i.e., is_a relationship) to GO:0045730 as its child
622 (descendant) term. These 2 terms are two-step away from each other on the GO DAG. This term
623 (GO:0045728) defines a more specific function that is the exact opposite of the mapped HPO term
624 (HP:0002723), semantically. Also, there is an evidence for the relation between HP:0002723 and
625 GO:0045728 in both the OBO and OWL formatted ontology files of HPO (URL:

626   http://purl.obolibrary.org/obo/hp.obo and http://purl.obolibrary.org/obo/hp.owl). However, in

627   HPO2GO, GO:0045728 could not be mapped to HP:0002723 due to low coverage in the source

628   GO annotation set. GO:0045728 was only annotated to one gene (symbol: *HCK*), which was not

629   annotated to HP:0002723, as a result, the mapping could not be generated. Nevertheless, the

630   mapped GO term (GO:0045730) still defined a sufficiently related function.

631   The second selected case with high-level reliability was the mapping between the HPO term

632   "cerebellar hemisphere hypoplasia" (HP:0100307) and the MF category GO term "tRNA-intron

633   endonuclease activity" (GO:0000213). The exact definition of this specific GO term in the

634   UniProt-GOA database is: "*Catalysis of the endonucleolytic cleavage of pre-tRNA, producing 5'-*

635   *hydroxyl and 2',3'-cyclic phosphate termini, and specifically removing the intron*" (URL:

636   https://www.ebi.ac.uk/QuickGO/term/GO:0000213). These two terms were mapped to each other

637   in HPO2GO with high confidence (i.e., $S = 0.86$ and $n = 3$). The symbols of the co-annotated genes

638   were *TSEN2*, *TSEN34* and *TSEN54*. The HPO term HP:0100307 is associated with the disease

639   entry "Pontocerebellar Hypoplasia, Type 2C (PCH2C)" (OMIM:612390) in the OMIM database.

640   According to the disease definition, pontocerebellar hypoplasia is a heterogeneous group of

641   neurodegenerative disorders associated with abnormally small cerebellum and brainstem, and the

642   type 2C is characterized by a progressive microcephaly from child birth (Barth, 1993). The

643   occurrence of the disease is associated with missense mutations in either *TSEN2*, *TSEN34* or

644   *TSEN54* genes, which are parts of the tRNA splicing endonuclease complex (Budde *et al.*, 2008).

645   It was reported that, due to the abovementioned mutations, there was a partial loss in the function

646   of cleaving the pre-tRNAs by the endonuclease complex (Budde *et al.*, 2008). This is another clear

647   example for a HPO term defining an abnormal condition, that is caused by the perturbation in the

648   function defined by the mapped GO term.

649   An example with mid-level reliability (with parameter values: $0.7 > S > 0.5$ and $10 > n > 2$), where

650   HPO2GO produced a mapping with reduced biologically relevance is the case where "protein

651   binding" (GO:0005515) MF GO term was associated with "cognitive impairment" (HP:0100543)

652   phenotype term. The parameter values for this mapping were $S = 0.38$ and $n = 571$. It might be

653   possible to find an indirect connection between some of the diseases that cause cognitive

654   impairment and the loss of protein binding function; however, it is impossible to conclude that all

655   protein binding function losses would result in cognitive impairment. Both of these terms are quite

656 generic; as a result, the mapping is not informative. The interesting observation here is the

657 extremely high n value, which is resulted from the fact that protein binding GO term is a generic

658 and frequent term that is annotated to many different protein entries. This result also indicated that

659 using the n value alone to judge the reliability of a mapping would not be sufficient, instead both

660 S and n should be taken into account.

661 Another example with mid-level reliability, that was randomly selected from the mappings, was

662 the association between the HPO term "anemic pallor" (HP:0001017) and the CC category GO

663 term "Fanconi anaemia nuclear complex" (GO:0043240). The exact definition of this GO term in

664 the UniProt-GOA database is: "*A protein complex composed of the Fanconi anaemia (FA) proteins*

665 *including A, C, E, G and F (FANCA-F). Functions in the activation of the downstream protein*

666 *FANCD2 by monoubiquitylation, and is essential for protection against chromosome breakage.*"

667 (URL: https://www.ebi.ac.uk/QuickGO/term/GO:0043240). The textual definition of the

668 corresponding HPO term (HP:0001017) is "*A type of pallor that is secondary to the presence of*

669 *anemia*". These two terms were mapped to each other in HPO2GO with mid-level confidence (i.e.,

670 $S = 0.64$ and $n = 8$). The symbols of the co-annotated genes were *FANCA, FANCB, FANCC,*

671 *FANCE, FANCF, FANCG, FANCL* and *FANCM*. These genes are parts of the Fanconi anaemia

672 (FA) protein complex, which protects the cell against chromosomal breakage (Pace *et al.*, 2002).

673 The corresponding CC GO term (GO:0043240) directly describes the FA complex; whereas, the

674 mapped HPO term (HP:0001017) is associated with the Fanconi anemia disease sub-types in the

675 HPO database (e.g., OMIM:600901, OMIM:227650, OMIM:227645, OMIM:227646). According

676 to the disease definition in OMIM, Fanconi anemia is a heterogeneous disorder associated with

677 genomic instability, mainly characterized by developmental abnormalities in major organ systems

678 (Deakyne and Mazin, 2011). Each of the different sub-types listed in OMIM is associated with a

679 mutation in a different FA complex gene. This is a clear example for a HPO term defining an

680 abnormal condition, that is caused by the disease rooted from the disfunction in the biomolecular

681 complex defined by the mapped GO term.

682 An example with low-level reliability (parameter values: $0.2 > S > 0.1$ and $n = 2$) is the case where

683 "abnormality of reproductive system physiology" (HP:0000080) HPO term was associated with

684 "collagen catabolic process" (GO:0030574) BP GO term. The parameter values for this mapping

685 were $S = 0.11$ and $n = 2$. The symbols of the co-annotated genes were *COL7A1* and *MMP1*. The

686 definition of this GO term in the UniProt-GOA database is: "*The proteolytic chemical reactions*

687 *and pathways resulting in the breakdown of collagen in the extracellular matrix, usually carried*

688 *out by proteases secreted by nearby cells.*" (URL:

689 https://www.ebi.ac.uk/QuickGO/term/GO:0030574). On the other hand, HP:0000080 is a generic

690 phenotype term that has 45 descendent terms, and it is possible to reach the root of the sub-ontology

691 from this term in just 3 term-to-term jumps. Naturally, HP:0000080 has been associated with high

692 number of diseases (i.e., 465) in the HPO database, one of which is the Myotonic dystrophy

693 (OMIM:160900). Myotonic dystrophy is an autosomal disorder characterized by muscular

694 dystrophy, myotonia, hypogonadism (i.e., functional activity related issues in the testes/ovaries),

695 and etc. (Musova *et al.*, 2009). In this sense, the relationship between muscular dystrophy and the

696 collagen breakdown is evident; as a result, it can be stated that HPO2GO identified a relevant

697 mapping. However, HP:0000080 is quite generic and associated with many other processes besides

698 muscular dystrophy; consequently, the target mapping is not very specific.

699 The last randomly selected case study example is another one from the low-level reliability bin.

700 The HPO term "polycythemia" (HP:0001901) and the BP category GO term "bicarbonate

701 transport" (GO:0015701) were mapped to each other with $S = 0.15$ and $n = 2$. The symbols of the

702 co-annotated genes were *HBA1* and *HBB* (i.e., haemoglobin sub-units). The definition of

703 GO:0015701 in the UniProt-GOA database is "*The directed movement of bicarbonate into, out of*

704 *or within a cell, or between cells, by means of some agent such as a transporter or pore*" (URL:

705 https://www.ebi.ac.uk/QuickGO/term/GO:0015701). One of the ways the carbon dioxide is

706 removed from tissues is first the generation of carbonic acid, and then the decomposition of

707 carbonic acid into bicarbonate to be transported to the lungs, via the red blood cells (i.e.,

708 erythrocytes). This process also drives the transport of oxygen molecules from blood to the tissues

709 via an allosteric mechanism. Thus, bicarbonate transport process have high importance for

710 oxygenation. The definition of HP:0001901 in the HPO database is "*Polycythemia is diagnosed if*

711 *the red blood cell count, the haemoglobin level, and the red blood cell volume all exceed the upper*

712 *limits of normal*". HP:0001901 is also associated with 23 diseases most of which are blood related.

713 A few of these diseases are polycythemia vera – PV (OMIM:263300) and various sub-types of

714 familial erythrocytosis (OMIM:13310, OMIM:263400, OMIM:609820, OMIM:611783).

715 Erythrocytosis is an autosomal disorder characterized by increased haemoglobin concentration,

716 increased mass of serum red blood cells, and etc. (Kralovics *et al.*, 1998). It is probable that certain

717 mutations in the proteins that take part in the bicarbonate transport process would cause a

718   disruption in this function, which would in turn lead to decreased efficiency in both the
719   oxygenation of tissues, and the removal of carbon dioxide from them, causing the polycythemia
720   phenotype related diseases such as polycythemia vera or erythrocytosis. In this example, HPO2GO
721   managed to return an indirect but relevant association.

722

4. DISCUSSION

HPO project's manual HPO-GO term associations have been generated only for a sub-set of HPO and GO terms, by comparing the term definitions. As a result, the coverage of these associations is limited (i.e., a total of 489 mappings). In HPO2GO approach, all GO-HPO term combinations that satisfy the co-occurrence similarity test conditions were linked. This way, the non-documented relations were also identified. The results of the mapping comparison analysis has indicated that the direct correspondence between the manual associations and HPO2GO was low, mainly due to the limitations in the source annotation sets. Nearly all of the manual HPO-GO associations had highly similar definitions. Whereas, HPO2GO retrieved associations with mostly dissimilar term definitions, which often made biological sense considering the underlying molecular mechanisms. The manual identification of this type of associations may require a comprehensive curation process. In this context, it is expected that the HPO2GO mappings will be valuable for the research community.

In this study, individual terms from both ontologies were mapped to each other considering the co-annotated genes/proteins. However, the initial design of the experiment considered the mapping of an HPO term to a trio of GO terms, one from each GO category (i.e., biological process – BP, molecular function – MF and cellular component – CC). This way, the corresponding phenotypic abnormality would be associated with a problem in a specific molecular event (defined by the MF term), as a part of a defined large-scale process (the BP term), occurring at a particular sub-cellular location (the CC term). This approach would have been biologically more relevant compared to the current design; however, the initial design failed due to the scarcity of both HPO annotations and MF, BP and CC GO term containing triple annotations. At this point in the study, a second option was considered, where HPO terms were tried to be mapped to MF and BP term pairs; nevertheless, the same problem was encountered again. Reliable annotation sets with higher coverage, which may become available in the future with more curation efforts, may solve this problem and make the abovementioned mapping approach practical. However today, even for the currently applied one to one term mapping approach, the main challenge is the low coverage of the predicted associations due to the small size of the source annotation sets. There can be a few alternative solutions to this problem. First of all, the training sets with enriched GO annotation may be obtained by including the annotations with evidence codes of reduced reliability (e.g., IEA – electronically generated), as this approach has been shown to work well with the CAFA2

754   benchmark set. Another option for enlarging the GO annotations would be utilizing the

755   gene/protein similarity information (i.e., protein function prediction). Scaling up the coverage of

756   both the HPO and the GO sets can be provided by propagating the annotations to the parent terms

757   according to the true path rule. "HPO2GO asserted vs. propagated mappings" analysis have

758   indicated that this approach significantly increases the coverage. Another option here would be

759   assuming a more elaborate approach in the mapping procedure by taking the graph-based

760   hierarchical term relationships into account while generating the HPO2GO mappings (i.e., the

761   parent and child terms of the target HPO-GO term pair, that are co-annotated to different

762   genes/proteins, will also contribute to the calculation of the co-occurrence similarity of the target

763   HPO-GO pair). This approach have been widely accepted in the area of sematic similarity based

764   functional analysis of biomolecules.

765   The official CAFA2 challenge results have indicated that, the methods based on sequence

766   similarities (e.g., the baseline classifier BLAST and a few models from the participating groups)

767   can achieve a good predictive performance considering the GO terms in the molecular function

768   (MF) category. This was expected since it is possible to detect most of the signatures related to the

769   molecular functions by analysing the amino acid sequence. However, most of the sequence-

770   similarity based methods failed in predicting the cellular component (CC) GO term and HPO term

771   associations. This can be explained for CC terms as either by the cleavage of the signals from the

772   sequence post-translationally or the difficulties in detecting weak signals used for directing

773   proteins to different compartments. Considering the HPO prediction, the case may completely be

774   different. As opposed to GO terms, which define the attributes the proteins contain, HPO terms

775   define phenotypic abnormalities caused by the protein when it loses one (or more) of its functions,

776   usually due to certain mutations in the gene that codes the protein. Due to this reason, transferring

777   a HPO annotation from one protein to another based on sequence similarity does not have a

778   biological relevance, which explains the poor performance of the BLAST classifier.

779   It was interesting to see that the HPO2GO CAFA2 benchmark set predictions produced using the

780   training set of asserted annotations resulted in the same overall performance as the predictions

781   produced using the training set of propagated annotations (without the inclusion of electronically

782   made GO annotations). In theory, employing the true path rule propagation would enrich the

783   training set (i.e., less false negatives) without any sacrifice in terms of the type I error (i.e., the

784   same number of false positives), which should have produced an elevated *Fmax* value. Instead, the

785 performance remained the same. There could be two possible reasons for this. First, there were
786 already false positive instances in the training set beforehand, and the propagation process just
787 made the situation worse by increasing the number of false positive instances in a magnitude equal
788 to the number of ancestor terms of the false positive instances (i.e., equal to the number of
789 propagation operations). This in turn compensated for the increase in the coverage obtained by
790 propagating the annotations. The second reason could be that, there were no significant errors
791 regarding the propagated HPO2GO predictions; however, the benchmark annotations were
792 incomplete. As a result, a portion of the real true positive predictions were counted as false
793 positives. Nevertheless, employing the propagated HPO annotations together with the enlarged
794 GO annotation set (including the electronically made annotations) have increased the predictive
795 performance by nearly 17%. However, it is not possible to be sure about the accuracy of these
796 results due to the small size of the CAFA2 test dataset. A larger benchmark annotation set that is
797 guaranteed to be complete would be required in order to discuss the performance further.

798 An important observation regarding the CAFA tests done in this study is that, there was a large
799 difference between the number of HPO predictions for CAFA2 and CAFA3 targets, using
800 HPO2GO with default parameters (i.e., 1,922,333 in CAFA2 as opposed to 3,453,130 in CAFA3).
801 There was also an increase in the number of predicted HPO terms (i.e., 2,640 in CAFA2 as opposed
802 to 3,719 in CAFA3), and there were no significant increase in the number of targets. The increase
803 in the number of predictions and the predicted HPO terms can be attributed to the training set
804 getting larger and more informative in time. The training set used for CAFA2 contained 133,175
805 annotations; whereas, it was 153,575 for CAFA3. The comparison of the predictive performances
806 of HPO2GO trained by the CAFA2 and the CAFA3 training sets may reveal more about the
807 situation.

808 Considering the HPO2GO biological relevance case studies, it was possible to find a biological
809 connection between the mapped HPO and GO terms in most cases; however, the connection
810 became indirect and more abstract, when the S and n values were low (i.e., reduced reliability).
811 Frequently, highly generic/shallow HPO and GO terms were mapped to multiple terms from the
812 other ontology, with generally low parameter values. These mappings were still observed to be
813 relevant but less informative. On the other hand, most of the mappings with high-level reliability
814 were between specific HPO and GO terms, and the selected cases from these mappings were
815 observed to be highly informative. Similar to the other automated methods that produce predictions

816   on the biological data, HPO2GO has limitations. This was also reflected in the results of the

817   predictive performance test on the CAFA2 benchmark set (*Fmax* = 0.35), which can be considered

818   low for real-life applications. For HPO2GO to be employed in biological data analysis pipelines

819   in the future, HPO-GO term mappings should also be manually curated.

820

821 **5. CONCLUSION**

822 In this study, a simple and effective strategy, HPO2GO, was proposed to semantically map

823 phenotypic abnormality defining HPO terms with biomolecular function defining GO terms,

824 considering the cross-ontology annotation co-occurrences on different genes/proteins. This

825 approach can easily be translated into novel HPO term predictions for genes/proteins. A literature

826 based case study was carried to discuss the biological relevance of the selected HPO2GO

827 mappings. This work also presents an application of the cross-ontology term mapping approach

828 by generating HPO-protein associations. HPO2GO was benchmarked on CAFA2 challenge

829 protein targets and it was revealed that the method was among the best performers of the HPO

830 term prediction track participators (i.e., the state-of-the-art methods). Also, the up-to-date trained

831 system was employed to predict HPO associations for all human proteins in the UniProtKB/Swiss-

832 Prot database (i.e., HPO2protein predictions). The methodology proposed here may also support

833 the already established approaches (e.g., text mining), as it is possible for different techniques with

834 different data sources and perspectives to produce results that complement distinct missing pieces

835 of the knowledge space. This property is often utilized in ensemble based classification

836 approaches. In this sense, It would also be interesting to analyse the complementarity between the

837 predictions of the proposed method and the predictions of the state-of-the-art approaches

838 participated in CAFA2 challenge; however, this was not possible since the actual prediction results

839 of the participant groups are not publicly available.

840 As for the future work, it is first planned to map the HPO terms to GO term trios (i.e., MF, BP and

841 CC terms at the same time) using enriched annotation datasets, as explained at the discussion

842 section. Another future task is the integration of HPO2GO mappings to our freely available GO

843 based automated protein function prediction tool/server UniGOPred (Rifaioglu *et al.*, 2018); so

844 that, query proteins that receive a GO term prediction will be automatically associated with the

845 HPO term(s) that are mapped to the corresponding GO term. It is expected that this approach

846 would produce large-scale HPO predictions for uncharacterized proteins without any curated

847 annotation, where the only available information is the amino acid sequence. The knowledge

848 extraction methodology proposed here can easily be combined with various types of protein

849 features employed in other predictive methods (e.g., variant information, PPIs, gene expression

850 profiles, etc.) to generate an ensemble HPO term prediction tool that identifies novel HPO-

851 gene/protein-disease associations.

6. REFERENCES

Alvarez MA, Qi X, Yan C. 2011. A shortest-path graph kernel for estimating gene product semantic similarity. *Journal of biomedical semantics*, 2(1), 3.

Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. 2014. OMIM. org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic acids research*, 43(D1), D789-D798.

Barth PG. 2014. Pontocerebellar hypoplasias: an overview of a group of inherited neurodegenerative disorders with fetal onset. *Brain Dev.*, 15: 411-422.

Bromberg Y. 2013. Disease gene prioritization. *PLoS computational biology*, 9(4), e1002902.

Budde BS, Namavar Y, Barth PG, Poll-The BT, Nürnberg G, Becker C, van Ruissen F, Weterman MA, Fluiter K, te Beek ET, Aronica E, van der Knaap MS, Höhne W, Toliat MR, Crow YJ, Steinling M, Voit T, Roelenso F, Brussel W, Brockmann K, Kyllerman M, Boltshauser E, Hammersen G, Willemsen M, Basel-Vanagaite L, Krägeloh-Mann I, de Vries LS, Sztriha L, Muntoni F, Ferrie CD, Battini R, Hennekam RC, Grillo E, Beemer FA, Stoets LM, Wollnik B, Nürnberg P, Baas F. 2008. tRNA splicing endonuclease mutations cause pontocerebellar hypoplasia. *Nature genetics*, 40(9), 1113.

Clark, WT, Radivojac, P. 2013. Information-theoretic evaluation of predicted ontological annotations. *Bioinformatics*, 29(13), i53-i61.

Cozzetto D, Buchan DW, Bryson K, Jones DT. 2013. Protein function prediction by massive integration of evolutionary analyses and multiple data sources. *BMC bioinformatics*, 14(3), p.S1.

Deakyne JS, Mazin AV. (2011). Fanconi anemia: at the crossroads of DNA repair. *Biochemistry (Moscow)*, 76(1), 36-48.

Deng Y, Gao L, Wang B, Guo X. 2015. HPOSim: an R package for phenotypic similarity measure and enrichment analysis based on the human phenotype ontology. *PloS one*, 10(2), e0115692.

Doğan T, MacDougall A, Saidi R, Poggioli D, Bateman A, O'Donovan C, Martin MJ. 2016. UniProt-DAAC: domain architecture alignment and classification, a new method for automatic functional annotation in UniProtKB. *Bioinformatics*, 32(15), 2264-2271.

Falda M, Toppo S, Pescarolo A, Lavezzo E, Di Camillo B, Facchinetti A, Cilia E, Velasco R,

880   Fontana P. 2012. Argot2: a large scale function prediction tool relying on semantic similarity of
881   weighted Gene Ontology terms. *BMC bioinformatics*, 13(4), S14.

882   Fang H, Gough J. 2012. DcGO: database of domain-centric ontologies on functions, phenotypes,
883   diseases and more. *Nucleic acids research*, 41(D1), D536-D544.

884   Finn RD, Attwood TK, Babbitt PC, Bateman A, Bork P, Bridge AJ, Chang HY, Dosztányi Z, El-
885   Gebali S, Fraser M, Gough J, Haft D, Holliday GL, Huang H, Huang X, Letunic I, Lopez R, Lu S,
886   Marchler-Bauer A, Mi H, Mistry J, Natale DA, Necci M, Nuka G, Orengo CA, Park Y, Pesseat S,
887   Piovesan D, Potter SC, Rawlings ND, Redaschi N, Richardson L, Rivoire C, Sangrador-Vegas A,
888   Sigrist C, Sillitoe I, Smithers B, Squizzato S, Sutton G, Thanki N, Thomas PD, Tosatto SC,, Wu
889   CH, Xenarios I, Yeh LS, Young SY, Mitchell AL. 2016. InterPro in 2017—beyond protein family
890   and domain annotations. *Nucleic acids research*, 45(D1), D190-D199.

891   Firth HV, Richards SM, Bevan AP, Clayton S, Corpas M, Rajan D, Van Vooren S, Moreau Y,
892   Pettett RM, Carter NP. 2009. DECIPHER: database of chromosomal imbalance and phenotype in
893   humans using ensembl resources. *The American Journal of Human Genetics*, 84(4), 524-533.

894   Gene Ontology Consortium. 2014. Gene ontology consortium: going forward. *Nucleic acids
895   research*, 43(D1), D1049-D1056.

896   Groza T, Köhler S, Moldenhauer D, Vasilevsky N, Baynam G, Zemojtel T, Schriml LM, Kibbe
897   WA, Schofield PN, Beck T, Vasant D, Brookes AJ, Zankl A, Washington NL, Mungall CJ, Lewis
898   SE, Haendel MA, Parkinson H, Robinson PN. 2015. The human phenotype ontology: semantic
899   unification of common and rare disease. *The American Journal of Human Genetics*, 97(1), 111-
900   124.

901   Guala D, Sonnhammer EL. 2017. A large-scale benchmark of gene prioritization methods.
902   *Scientific reports*, 7, 46598.

903   Guney E, Oliva B. 2014. Analysis of the robustness of network-based disease-gene prioritization
904   methods reveals redundancy in the human interactome and functional diversity of disease-genes.
905   *PLoS one*, 9(4), e94686.

906   Hawkins T, Chitale M, Luban S, Kihara D. 2009. PFP: Automated prediction of gene ontology
907   functional annotations with confidence scores using protein sequence data. *Proteins: Structure,
908   Function, and Bioinformatics*, 74(3), 566-582.

909 Hollander M, Wolfe DA, Chicken E. 2013. *Nonparametric statistical methods*. John Wiley &
910 Sons.

911 Huntley RP, Sawford T, Mutowo-Meullenet P, Shypitsyna A, Bonilla C, Martin MJ, O'Donovan
912 C. 2015. The GOA database: gene ontology annotation updates for 2015. *Nucleic acids research*,
913 43(D1):1057–63.

914 Jiang Y, Oron TR, Clark WT, Bankapur AR, D'Andrea D, Lepore R, Funk CS, Kahanda I,
915 Verspoor KM,, Ben-Hur A, Koo da CE, Penfold-Brown D,, Shasha D, Youngs N,,, Bonneau R,,,
916 Lin A, Sahraeian SM, Martelli PL, Profiti G, Casadio R, Cao R, Zhong Z, Cheng J, Altenhoff A,,
917 Skunca N,, Dessimoz C,,, Dogan T, Hakala K,, Kaewphan S,,, Mehryary F,, Salakoski T,, Ginter
918 F, Fang H, Smithers B, Oates M, Gough J, Törönen P, Koskinen P, Holm L,, Chen CT, Hsu WL,
919 Bryson K, Cozzetto D, Minneci F, Jones DT, Chapman S, Bkc D, Khan IK, Kihara D,, Ofer D,
920 Rappoport N,, Stern A,, Cibrian-Uhalte E, Denny P, Foulger RE, Hieta R, Legge D, Lovering RC,
921 Magrane M, Melidoni AN, Mutowo-Meullenet P, Pichler K, Shypitsyna A, Li B, Zakeri P,, ElShal
922 S,, Tranchevent LC,,, Das S, Dawson NL, Lee D, Lees JG, Sillitoe I, Bhat P, Nepusz T, Romero
923 AE, Sasidharan R, Yang H, Paccanaro A, Gillis J, Sedeño-Cortés AE, Pavlidis P, Feng S, Cejuela
924 JM, Goldberg T, Hamp T, Richter L, Salamov A, Gabaldon T,,, Marcet-Houben M,, Supek F,,,
925 Gong Q,, Ning W,, Zhou Y,, Tian W,, Falda M, Fontana P, Lavezzo E, Toppo S, Ferrari C, Giollo
926 M,, Piovesan D, Tosatto SC, Del Pozo A, Fernández JM, Maietta P, Valencia A, Tress ML, Benso
927 A, Di Carlo S, Politano G, Savino A, Rehman HU, Re M, Mesiti M, Valentini G, Bargsten JW,
928 van Dijk AD,, Gemovic B, Glisic S, Perovic V, Veljkovic V, Veljkovic N, Almeida-E-Silva DC,
929 Vencio RZ, Sharan M, Vogel J, Kansakar L, Zhang S, Vucetic S, Wang Z, Sternberg MJ, Wass
930 MN, Huntley RP, Martin MJ, O'Donovan C, Robinson PN, Moreau Y, Tramontano A, Babbitt PC,
931 Brenner SE, Linial M, Orengo CA, Rost B, Greene CS, Mooney SD, Friedberg I,, Radivojac P.
932 2016. An expanded evaluation of protein function prediction methods shows an improvement in
933 accuracy. *Genome biology*, 17(1), 184.

934 Kahanda I, Funk C, Verspoor K, Ben-Hur A. 2015. PHENOstruct: Prediction of human phenotype
935 ontology terms using heterogeneous data sources. *F1000Research*, 4.

936 Kibbe WA, Arze C, Felix V, Mitraka E, Bolton E, Fu G, Mungall CJ, Binder JX, Malone J, Vasant
937 D, Parkinson H, Schriml LM . 2014. Disease ontology 2015 update: an expanded and updated
938 database of human diseases for linking biomedical knowledge through disease data. *Nucleic acids*

939   *research*, 43:1071–8.

940   Köhler S, Schulz MH, Krawitz P, Bauer S, Dölken S, Ott CE, Mundlos C, Horn D, Mundlos S,

941   Robinson PN. 2009. Clinical diagnostics in human genetics with semantic similarity searches in

942   ontologies. *The American Journal of Human Genetics*, 85(4), 457-464.

943   Köhler S, Vasilevsky NA, Engelstad M, Foster E, McMurry J, Aymé S, Baynam G,, Bello SM,

944   Boerkoel CF, Boycott KM, Brudno M, Buske OJ, Chinnery PF,, Cipriani V,, Connell LE, Dawkins

945   HJ, DeMare LE, Devereau AD, de Vries BB, Firth HV, Freson K, Greene D,, Hamosh A, Helbig

946   I,, Hum C, Jähn JA, James R,, Krause R, F Laulederkind SJ, Lochmüller H, Lyon GJ, Ogishima

947   S, Olry A, Ouwehand WH, Pontikos N,, Rath A, Schaefer F, Scott RH, Segal M, Sergouniotis PI,

948   Sever R, Smith CL, Straub V, Thompson R, Turner C, Turro E,, Veltman MW, Vulliamy T, Yu J,

949   von Ziegenweidt J, Zankl A,, Züchner S, Zemojtel T, Jacobsen JO, Groza T,, Smedley D, Mungall

950   CJ, Haendel M, Robinson PN. 2016. The human phenotype ontology in 2017. *Nucleic acids*

951   *research*, 45(D1), D865-D876.

952   Krallinger M, Valencia A, Hirschman L. 2008. Linking genes to literature: text mining,

953   information extraction, and retrieval applications for biology. *Genome biology*, 9(2), S8.

954   Kralovics R, Sokol L, Prchal, JT. 1998. Absence of polycythemia in a child with a unique

955   erythropoietin receptor mutation in a family with autosomal dominant primary polycythemia. *The*

956   *Journal of clinical investigation*, 102(1), 124-129.

957   Kulmanov M, Hoehndorf R. 2017. Evaluating the effect of annotation size on measures of

958   semantic similarity. *Journal of biomedical semantics*, 8(1), 7.

959   Lan L, Djuric N, Guo Y, Vucetic S. 2013. MS-k NN: protein function prediction by integrating

960   multiple data sources. *BMC bioinformatics*, 14(3), p.S8.

961   Lilliefors HW. 1967. On the Kolmogorov-Smirnov test for normality with mean and variance

962   unknown. *Journal of the American statistical Association*, 62(318), 399-402.

963   Lin D. 1998. An information-theoretic definition of similarity. *Proc. of the 15th International*

964   *Conference on Machine Learning*. San Francisco, CA: Morgan Kaufmann. pp. 296–304.

965   Moreau Y, Tranchevent LC. 2012. Computational tools for prioritizing candidate genes: boosting

966   disease gene discovery. *Nature reviews genetics*, 13(8), 523.

967   Musova Z, Mazanec R, Krepelova A, Ehler E, Vales J, Jaklova R, Prochazka T, Koukal P,
968   Marikova T, Kraus J, Havlovicova M, Sedlacek Z. 2009. Highly unstable sequence interruptions
969   of the CTG repeat in the myotonic dystrophy gene. *American journal of medical genetics Part A*,
970   149(7), 1365-1374.

971   Notaro M, Schubach M, Robinson PN, Valentini G. 2017. Prediction of Human Phenotype
972   Ontology terms by means of hierarchical ensemble methods. *BMC bioinformatics*, 18(1), 449.

973   Pace P, Johnson M, Tan WM, Mosedale G, Sng C, Hoatlin M, de Winter J, Joenje H, Gergely F,
974   Patel KJ. 2002. FANCE: the link between Fanconi anaemia complex assembly and activity. The
975   *EMBO journal*, 21(13), 3414-3423.

976   Peng J, Li Q, Shang X. 2017. Investigations on factors influencing HPO-based semantic similarity
977   calculation. *Journal of biomedical semantics*, 8(1), 34.

978   Radivojac P, Clark WT, Oron TR, Schnoes AM, Wittkop T, Sokolov A, Graim K, Funk C,
979   Verspoor K, Ben-Hur A, Pandey G, Yunes JM, Talwalkar AS, Repo S, Souza ML, Piovesan D,
980   Casadio R, Wang Z, Cheng J, Fang H, Gough J, Koskinen P, Törönen P, Nokso-Koivisto J, Holm
981   L, Cozzetto D, Buchan DW, Bryson K, Jones DT, Limaye B, Inamdar H, Datta A, Manjari SK,
982   Joshi R, Chitale M, Kihara D, Lisewski AM, Erdin S, Venner E, Lichtarge O, Rentzsch R, Yang
983   H, Romero AE, Bhat P, Paccanaro A, Hamp T, Kaßner R, Seemayer S, Vicedo E, Schaefer C,
984   Achten D, Auer F, Boehm A, Braun T, Hecht M, Heron M, Hönigschmid P, Hopf TA, Kaufmann
985   S, Kiening M, Krompass D, Landerer C, Mahlich Y, Roos M, Björne J, Salakoski T, Wong A,
986   Shatkay H, Gatzmann F, Sommer I, Wass MN, Sternberg MJ, Škunca N, Supek F, Bošnjak M,
987   Panov P, Džeroski S, Šmuc T, Kourmpetis YA, van Dijk AD, ter Braak CJ, Zhou Y, Gong Q,
988   Dong X, Tian W, Falda M, Fontana P, Lavezzo E, Di Camillo B, Toppo S, Lan L, Djuric N, Guo
989   Y, Vucetic S, Bairoch A, Linial M, Babbitt PC, Brenner SE, Orengo C, Rost B, Mooney SD,
990   Friedberg I. 2013. A large-scale evaluation of computational protein function prediction. *Nature*
991   *methods*, 10(3), 221.

992   Rath A, Olry A, Dhombres F, Brandt MMC, Urbero B, Ayme S. 2012. Representation of rare
993   diseases in health information systems: the Orphanet approach to serve a wide range of end users.
994   *Hum. Mutat.*, 33:803–808.

995   Rifaioglu AS, Doğan T, Saraç ÖS, Ersahin T, Saidi R, Atalay MV, Martin MJ, Cetin-Atalay R.

996 2018. Large-scale automated function prediction of protein sequences and an experimental case
997 study validation on PTEN transcript variants. *Proteins: Structure, Function, and Bioinformatics*,
998 86(2), 135-151.

999 Robinson PN, Köhler S, Bauer S, Seelow D, Horn D, Mundlos S. 2008. The Human Phenotype
1000 Ontology: a tool for annotating and analyzing human hereditary disease. *The American Journal of*
1001 *Human Genetics*, *83*(5), 610-615.

1002 Rodríguez-García MÁ, Gkoutos GV, Schofield PN, Hoehndorf R. 2017. Integrating phenotype
1003 ontologies with PhenomeNET. *Journal of biomedical semantics*, 8(1), 58.

1004 Roy A, Yang J, Zhang Y. 2012. COFACTOR: an accurate comparative algorithm for structure-
1005 based protein function annotation. *Nucleic acids research*, 40(W1), W471-W477.

1006 Schlicker A, Lengauer T, Albrecht M. 2010. Improving disease gene prioritization using the
1007 semantic similarity of Gene Ontology terms. *Bioinformatics*, 26(18), i561-i567.

1008 Smedley D, Oellrich A, Köhler S, Ruef B; Sanger Mouse Genetics Project, Westerfield M,
1009 Robinson P, Lewis S, Mungall C. 2013. PhenoDigm: analyzing curated annotations to associate
1010 animal models with human diseases. *Database*, bat025.

1011 Smith CL, Goldsmith CAW, Eppig JT. 2005. The Mammalian Phenotype Ontology as a tool for
1012 annotating, analyzing and comparing phenotypic information. *Genome biology*, 6(1), R7.

1013 UniProt Consortium. 2017. UniProt: the universal protein knowledgebase. *Nucleic acids research*,
1014 45(D1), D158-D169.

1015 Valentini G, Armano G, Frasca M, Lin J, Mesiti M, Re M. 2016. RANKS: a flexible tool for node
1016 label ranking and classification in biological networks. *Bioinformatics*, 32(18), 2872-2874.

1017 Van Landeghem S, Björne J, Wei CH, Hakala K, Pyysalo S, Ananiadou S, Kao HY, Lu Z,
1018 Salakoski T, Van de Peer Y, Ginter F. 2013. Large-scale event extraction from literature with
1019 multi-level gene normalization. *PloS one*, 8(4), e55814.

1020 Vasant D, Chanas L, Malone J, Hanauer M, Olry A, Jupp S, Robinson PN, Parkinson H, Rath A.
1021 2014. Ordo: An ontology connecting rare disease, epidemiology and genetic data. *In Proceedings*
1022 *of ISMB*.

1023   Washington NL, Haendel MA, Mungall CJ, Ashburner M, Westerfield M, Lewis SE. 2009.

1024   Linking human diseases to animal models using ontology-based phenotype annotation. *PLoS*

1025   *biology*, 7(11), e1000247.

1026   Wass MN, Barton G, Sternberg MJ. 2012. CombFunc: predicting protein function using

1027   heterogeneous data sources. *Nucleic acids research*, 40(W1), W466-W470.

1028

Statistics of the initial (i.e., raw) original and randomized HPO-GO mappings ($n \geq 1$).

1   **Table 1.** Statistics of the initial (i.e., raw) original and randomized HPO-GO mappings ($n \geq 1$).

S	# of mappings		# of mapped HPO terms		# of mapped GO terms	
	Original mapping	Random mapping	Original mapping	Random mapping	Original mapping	Random mapping
= 1	2 433	1 898	844	877	1 108	1 265
≥ 0.9	2 440	1 898	848	877	1 109	1 265
≥ 0.8	2 658	1 899	962	878	1 179	1 266
≥ 0.7	2 805	1 899	1 028	878	1 212	1 266
≥ 0.6	7 355	5 249	1 941	1 653	2 577	2 844
≥ 0.5	8 075	5 252	2 188	1 655	2 712	2 847
≥ 0.4	15 462	9 724	3 014	2 243	4 053	4 207
≥ 0.3	32 393	21 615	4 082	3 017	6 011	6 081
≥ 0.2	63 439	43 593	5 032	3 662	7 569	7 490
≥ 0.1	181 048	134 038	5 920	5 199	8 884	9 005
> 0.0	1 433 208	1 543 917	6 005	5 995	9 685	9 685

KS test significance values for the comparison of original vs. randomized distributions

KS test significance values for the comparison of original vs. randomized distributions at different co-occurrence similarity (S) and the number of co-annotated genes (n) thresholds.

1 **Table 2.** KS test significance values for the comparison of original vs. randomized distributions at

2 different co-occurrence similarity (*S*) and the number of co-annotated genes (*n*) thresholds.

KS test statistic		Co-occurrence similarity threshold						
		$S > 0$	$S \geq 0.1$	$S \geq 0.2$	$S \geq 0.3$	$S \geq 0.4$	$S \geq 0.5$	$S \geq 0.6$
# of co-annotated genes threshold	$n \geq 1$	0.6882	0.6884	0.4536	0.2366	0.3921	0.3484	0.3113
	$n \geq 2$	0.0423	**0.0057**	0.0005	0.0001	0.0002	0.0038	NaN
	$n \geq 3$	0.2636	0.0045	0.0000	NaN	NaN	NaN	NaN
	$n \geq 4$	0.2830	0.0039	0.0000	NaN	NaN	NaN	NaN
	$n \geq 5$	0.3349	0.0105	0.0000	NaN	NaN	NaN	NaN

Figure 1

Schematic representation of the whole HPO2GO mapping (i.e., training) procedure.
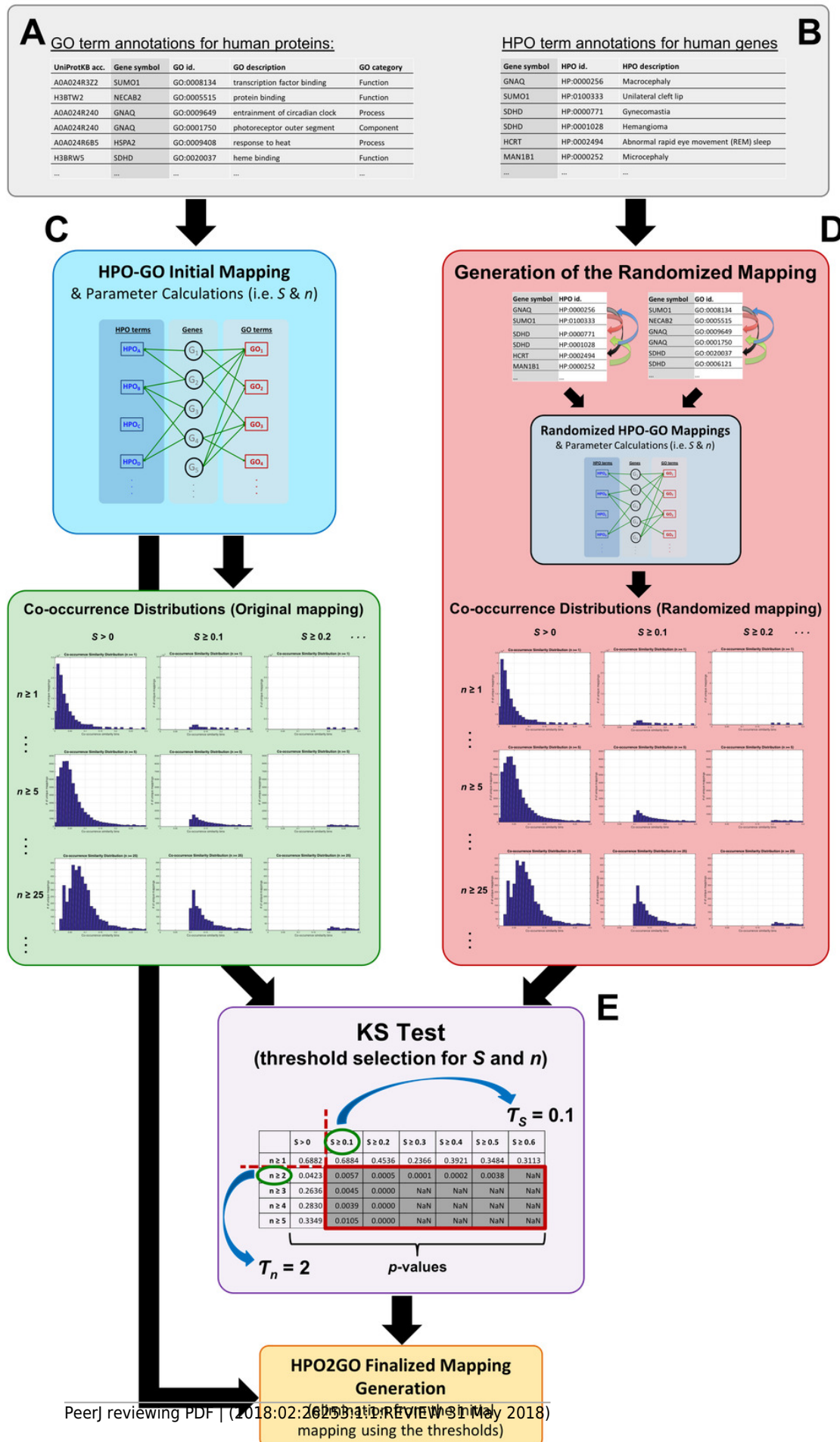
Figure 2

Representation of the initial HPO-GO mapping process

Representation of the initial HPO-GO mapping process together with the calculation of co-occurrence similarities (S) and the number of genes with co-occurring annotations (n), on a toy example.
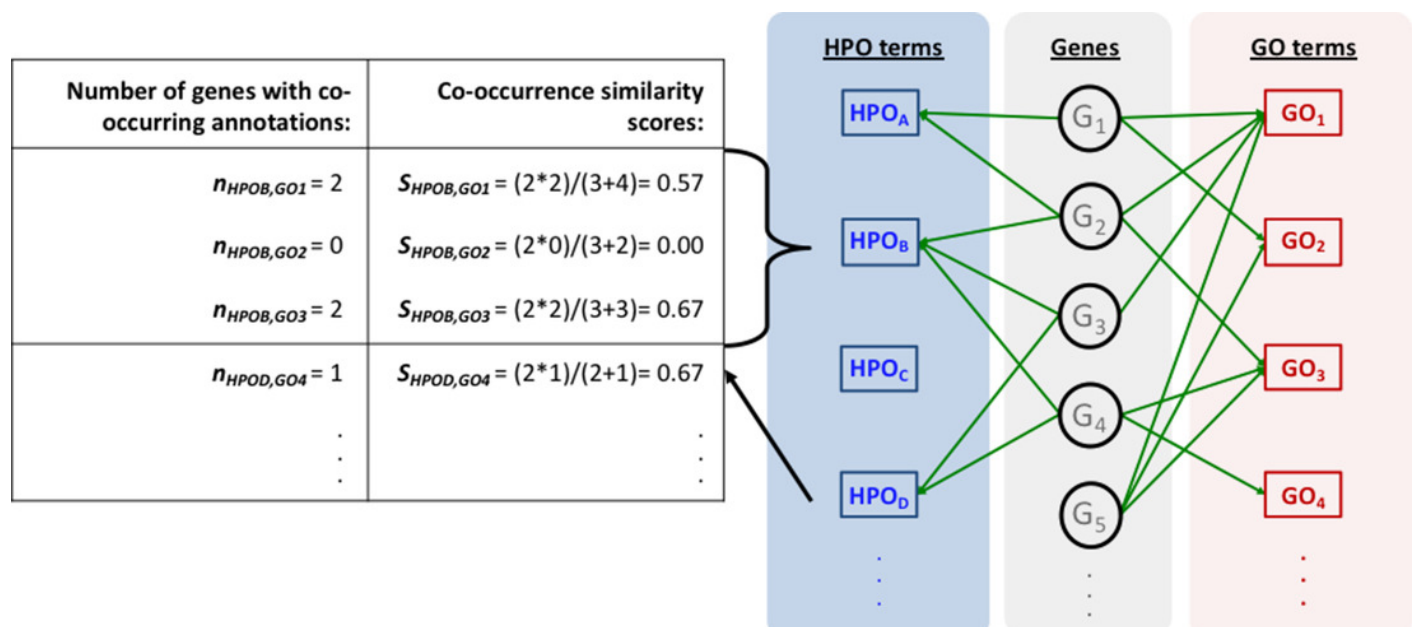
Figure 3

HPO-GO initial mappings co-occurrence similarity distributions

Each plot is drawn for a different value of the number of co-annotated genes (i.e., *n*).
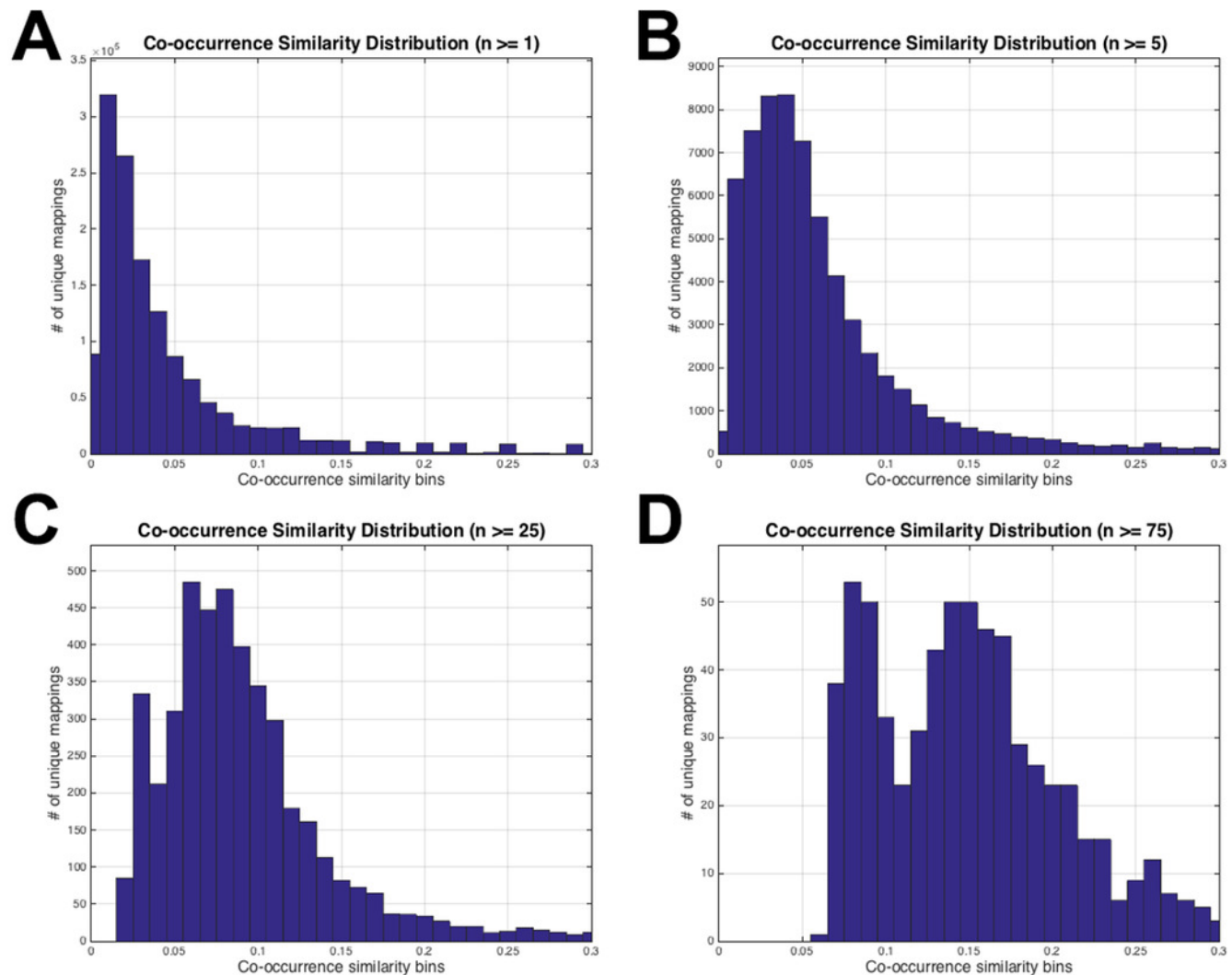
Figure 4

Cumulative plots displaying the number of HPO-GO mappings for the original (blue curve) and the randomized (red curve) distributions.

Horizontal axis displays the arbitrarily selected co-occurrence similarity thresholds (i.e., τ_s), and the vertical axis represents the logarithm of the total number of mappings left after the application of the corresponding threshold. Each plot is drawn for a different value of the number of co-annotated genes (i.e., n). As the threshold (i.e., the minimum required co-occurrence similarity value to keep a mapping in the system) increase, more mappings are eliminated; thus, a monotonic decrease was observed for all plots.
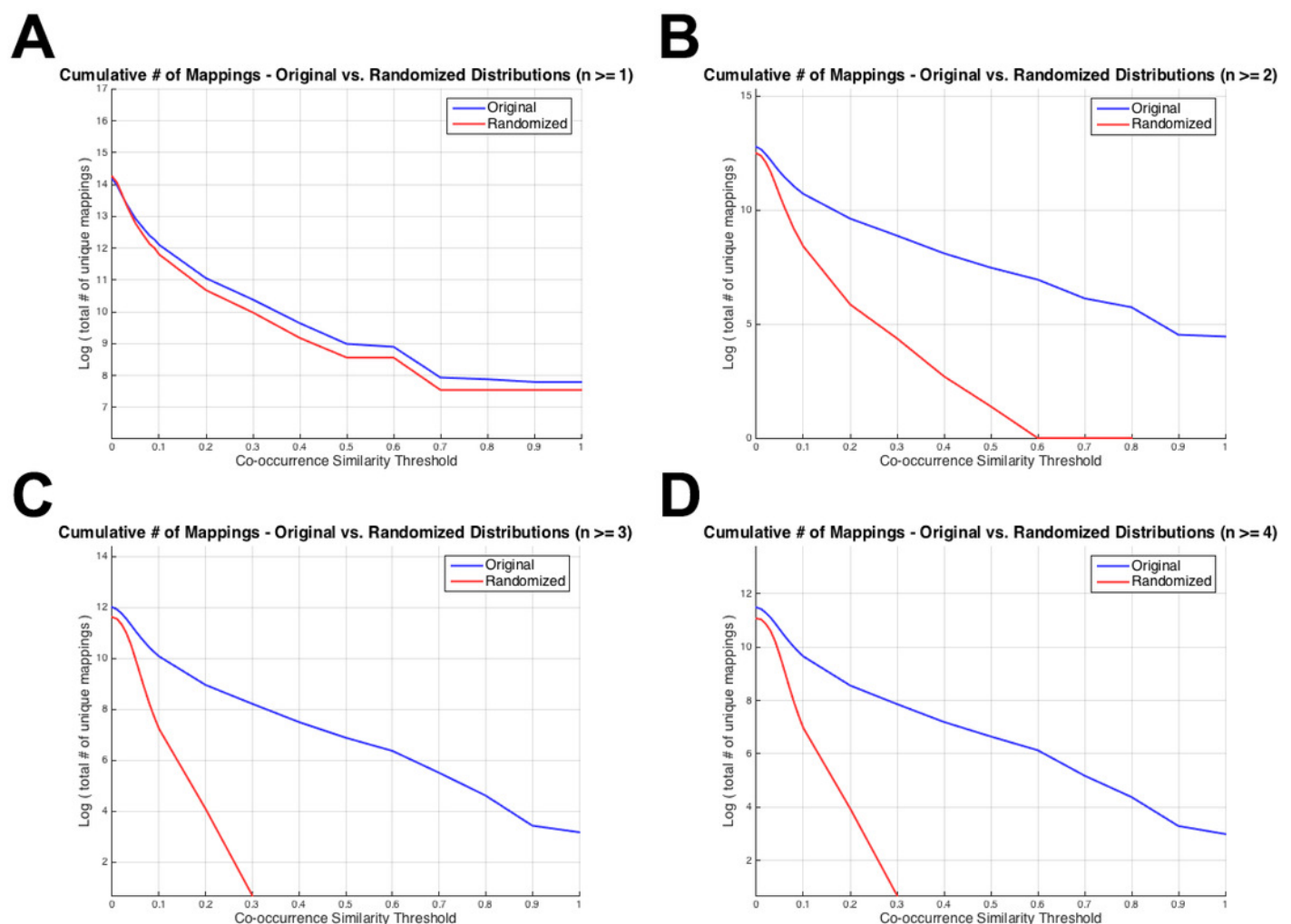
Figure 5

Fmax performance results in CAFA2 HPO prediction benchmark

Performance results (*Fmax*) of the top performing groups (grey bars), baseline classifiers (red and blue bars) and HPO2GO (dark grey bar) in CAFA2 HPO prediction benchmark. C represents the coverage of predictions on the target protein set.
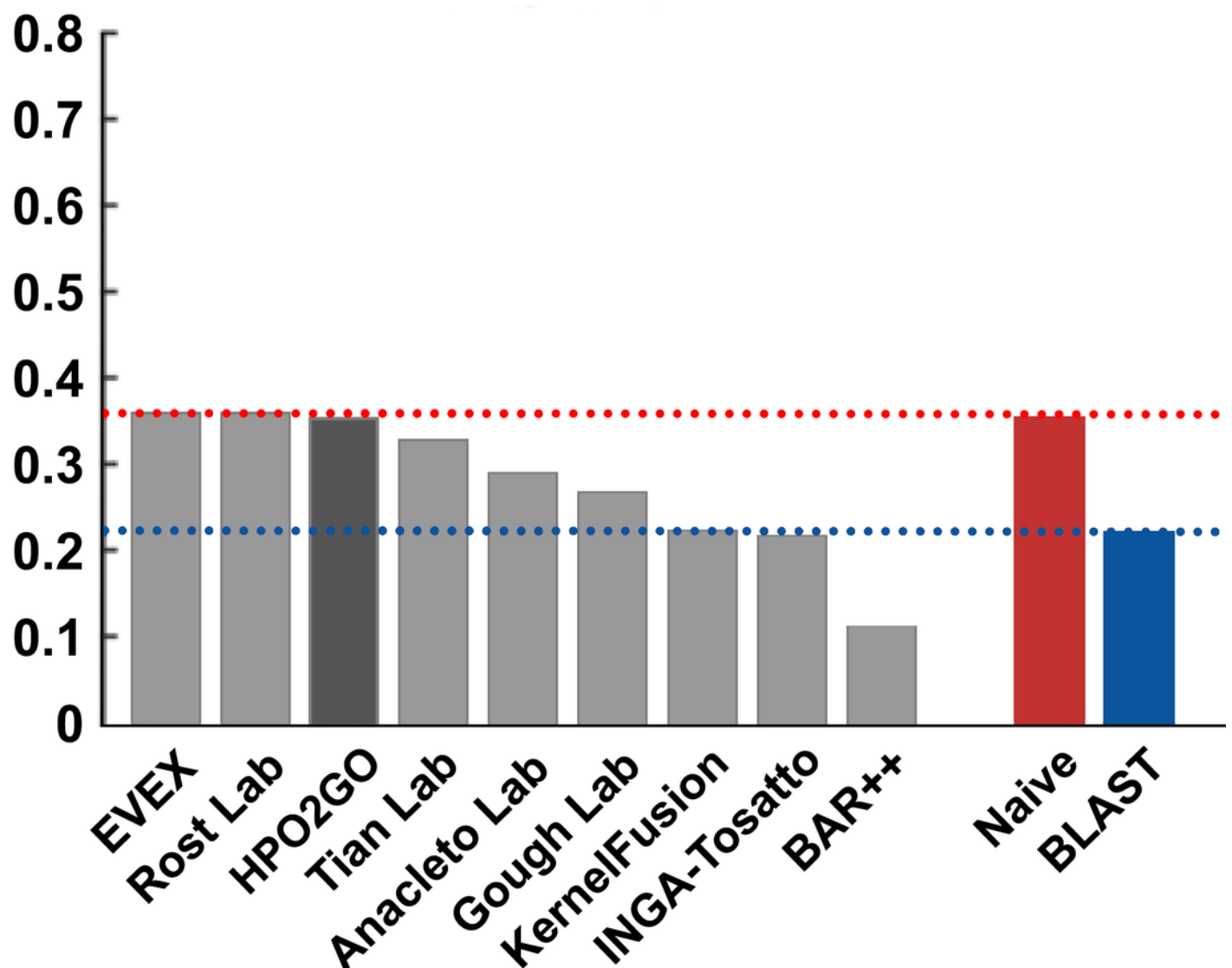
Figure 6

Precision-recall curves for the CAFA2 HPO prediction benchmark

Precision-recall curves of the top performing groups (coloured curves), baseline classifiers (red and blue dashed curves) and HPO2GO (black curve) in CAFA2 HPO prediction benchmark. The circle on each curve represent the performance at the optimal threshold. Dashed grey iso-performance curves project F-scores. F (inside the box, at the right side) display Fmax values.
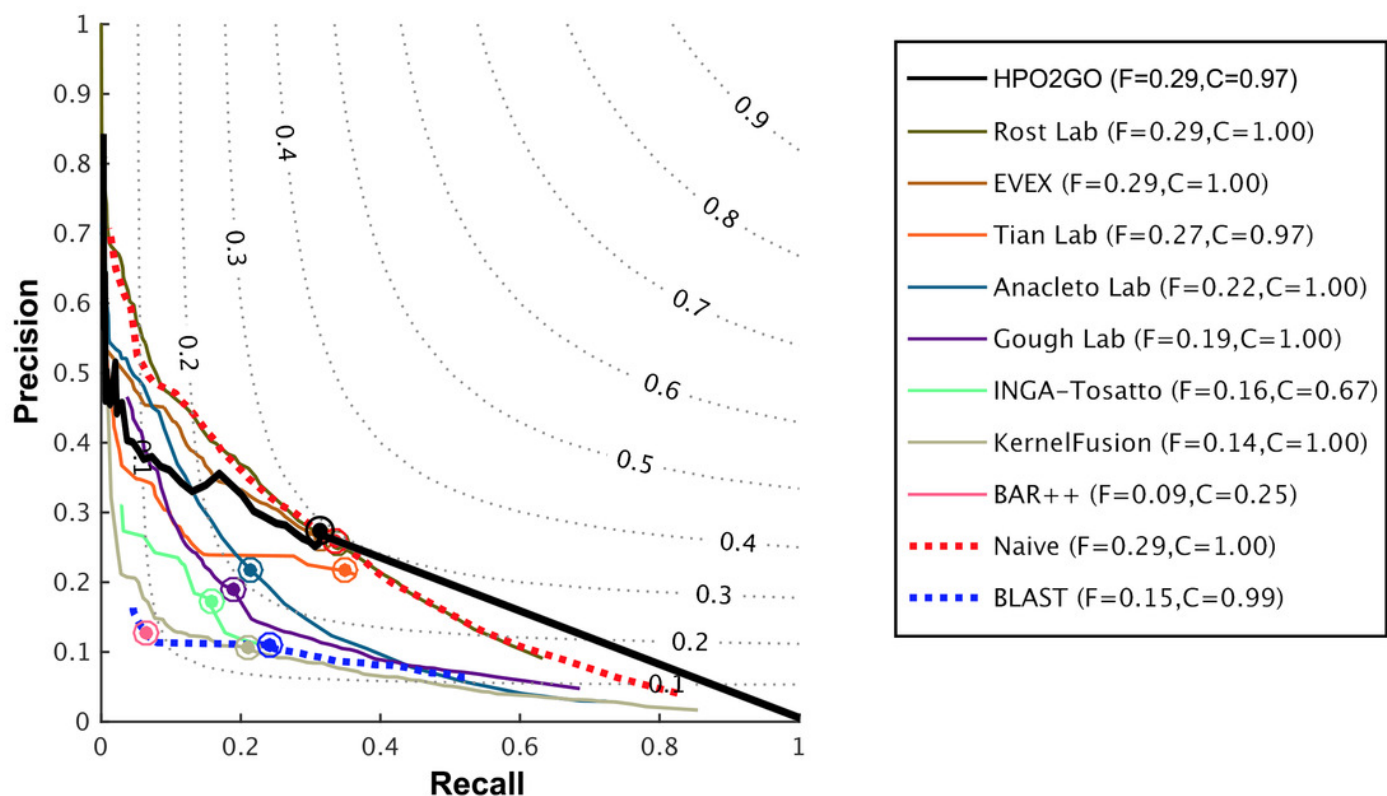
Figure 7

Normalized remaining uncertainty - misinformation curves for the CAFA2 HPO prediction benchmark

Normalised remaining uncertainty - misinformation curves of the top performing groups (coloured curves), baseline classifiers (red and blue dashed curves) and HPO2GO (black curve) in CAFA2 HPO prediction benchmark. The circle on each curve represent the normalized minimum semantic distance measure (nSmin). Dashed grey iso-performance curves project nSmin values. S (inside the box, at the right side) display nSmin values.
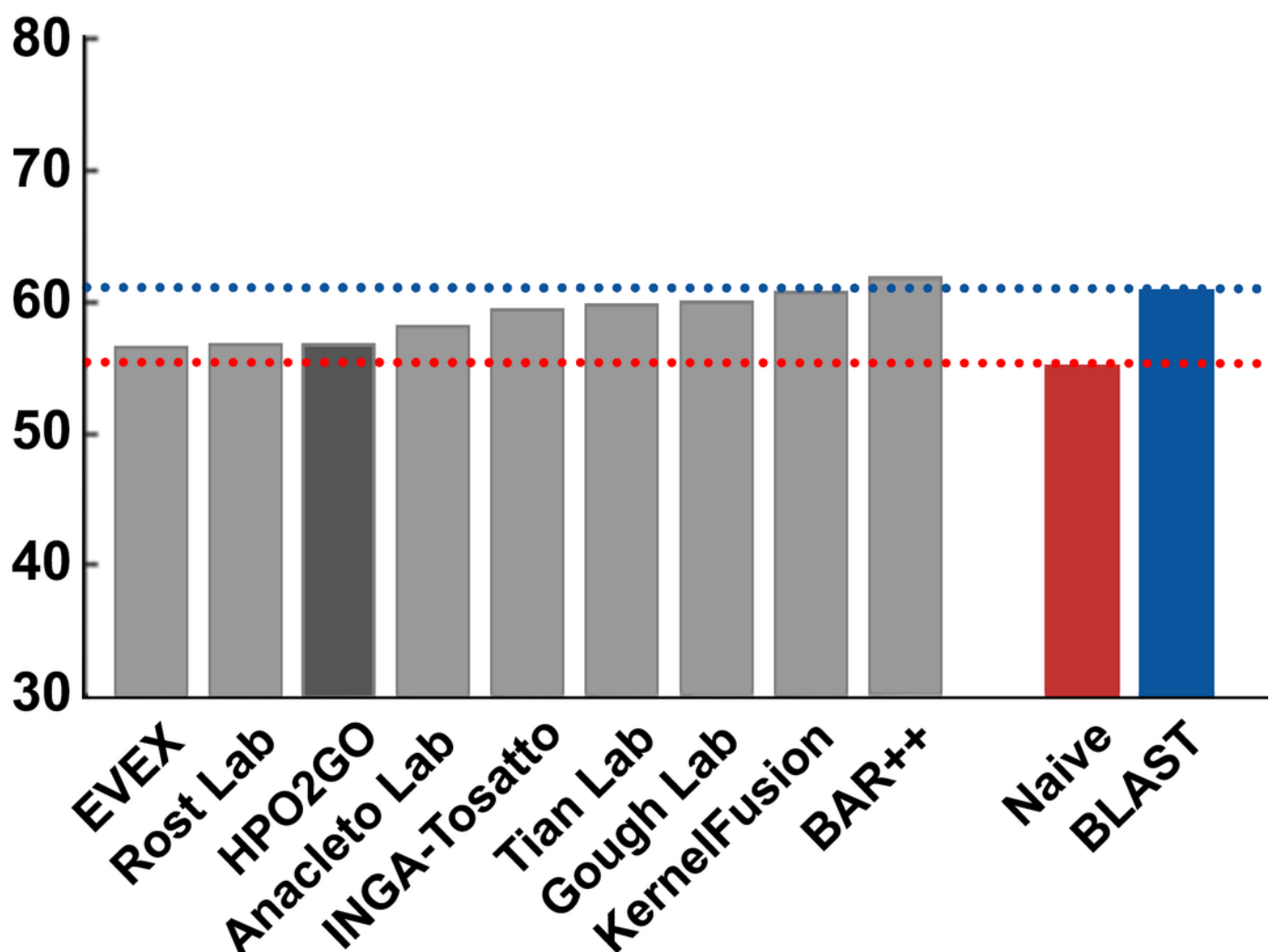
Figure 8

Average term centric AUROC performance results in CAFA2 HPO prediction benchmark

Performance results (average term centric AUROC curve) of the top performing groups (grey bars), baseline classifiers (red and blue bars) and HPO2GO (dark grey bar) in CAFA2 HPO prediction benchmark.