genes

MDPI

Review

Automatic Gene Function Prediction in the 2020's

Stavros Makrodimitris [1,2,*], Roeland C. H. J. van Ham [1,2] and Marcel J. T. Reinders [1,3]

[1] Delft Bioinformatics Lab, Delft University of Technology, 2628XE Delft, The Netherlands; r.c.h.j.vanham@tudelft.nl (R.C.H.J.v.H.); M.J.T.Reinders@tudelft.nl (M.J.T.R.)
[2] Keygene N.V., 6708PW Wageningen, The Netherlands
[3] Leiden Computational Biology Center, Leiden University Medical Center, 2333ZC Leiden, The Netherlands
* Correspondence: s.makrodimitris@tudelft.nl

Received: 28 September 2020; Accepted: 21 October 2020; Published: 27 October 2020

Abstract: The current rate at which new DNA and protein sequences are being generated is too fast to experimentally discover the functions of those sequences, emphasizing the need for accurate Automatic Function Prediction (AFP) methods. AFP has been an active and growing research field for decades and has made considerable progress in that time. However, it is certainly not solved. In this paper, we describe challenges that the AFP field still has to overcome in the future to increase its applicability. The challenges we consider are how to: (1) include condition-specific functional annotation, (2) predict functions for non-model species, (3) include new informative data sources, (4) deal with the biases of Gene Ontology (GO) annotations, and (5) maximally exploit the GO to obtain performance gains. We also provide recommendations for addressing those challenges, by adapting (1) the way we represent proteins and genes, (2) the way we represent gene functions, and (3) the algorithms that perform the prediction from gene to function. Together, we show that AFP is still a vibrant research area that can benefit from continuing advances in machine learning with which AFP in the 2020s can again take a large step forward reinforcing the power of computational biology.

Keywords: automatic function prediction; Gene Ontology; protein representation; machine learning

1. Introduction

Automatic function prediction (AFP) deals with the algorithmic assignment of functional annotations—usually Gene Ontology (GO) terms—to proteins/genes of unknown function from proteins/genes whose function has already been determined experimentally. In the past two decades, the amount of new protein sequences has been growing at such a fast pace [1] that no experimental screen can keep up, making AFP a necessity for modern biology. In addition to generating fundamental biological knowledge about what proteins do, AFP is crucial for other aspects of research, such as linking genotype to phenotype, by enabling gene set enrichment analyses or facilitating the interpretation of GWAS hits (e.g., [2]). A far from exhaustive list of some of the recent successful AFP models is given in Table 1.

Table 1. Some of the most important AFP models proposed in the past years.

Name	Reference	Input Data	Method
GOLabeler	[3]	Amino acid sequence, GO term frequencies	Learning to rank
FunFams	[4]	Amino acid sequence	Hidden Markov Model
INGA	[5]	Amino acid sequence	Homology search, enrichment analysis
PFP	[6]	Amino acid sequence	Phylogenetics
COFACTOR	[7]	Amino acid sequence, protein structure, protein interactions	Homology search, structural similarity
NetGO	[8]	Amino acid sequence, GO term frequencies, protein interactions	Learning to rank
DeepGOPlus	[9]	Amino acid sequence	Convolutional neural network, homology search

AFP: Automatic Function Prediction. GO: Gene Ontology.

The Critical Assessment of Functional Annotation (CAFA) challenges provide an objective evaluation of modern AFP algorithms on a set of proteins with newly-acquired GO annotations [10–12]. The main finding of these challenges is that methods significantly improved between CAFA1 and CAFA2, but remained rather stagnant in CAFA3, with the exception of one novel method, GOLabeler [3], that outperformed all others, especially in the Molecular Function Ontology (MFO) [12]. Several methods performed rather similarly in the Biological Process Ontology (BPO) and all participating methods failed to outperform a simple co-expression-based baseline method at predicting cell motility and biofilm formation in *Pseudomonas aeruginosa* [12]. Together, these results show that the problem of AFP is far from solved and that perhaps one or several leaps are required to advance the field.

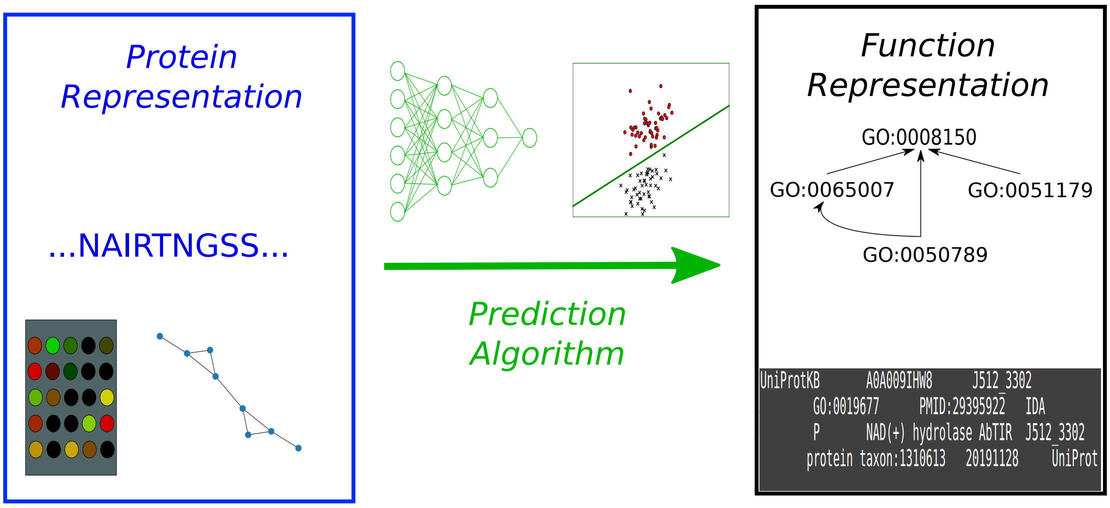
More than a decade ago, in a paper that set the foundation for the CAFA benchmarks, Godzik et al. defined three main challenges for AFP research [13]:

- How to extend AFP beyond homology transfer.
- How to define protein function in a standardized way.
- How to properly evaluate AFP methods.

Since then, all three of these questions have been addressed to varying extents. Several algorithms have been proposed which make use of different data sources, such as sequence features [14], gene expression, and protein-protein interactions [15]. The GO has become the standard vocabulary for describing protein function in the vast majority of AFP models, and the CAFA is a widely-accepted platform to objectively evaluate these models. To inspire the AFP field in realizing new breakthroughs, we have attempted to identify some (new) challenges that we feel are important for advancing the field and we will address them in detail in the next sections. These challenges are:

1. How can we deal with biological function being tissue, cell-type, or condition-specific?
2. How do we predict functions in non-model species?
3. What data sources should be used for predicting function?
4. How does missingness or bias in GO annotations affect the training of AFP models?
5. How can we better exploit the Gene Ontology structure to improve functional annotation?

We will discuss these challenges from three different perspectives of an AFP pipeline (Figure 1): (a) how proteins are represented, (b) how function is encoded, and (c) what kind of prediction algorithms one should use. Protein representations refer to the input data types that feed an AFP model, e.g., sequence similarities or raw amino acid sequences as well as any feature extraction steps. Generally, functions are described using GO terms, but these have limitations or may require adaptations to cope with the challenges identified. Finally, prediction algorithms deal with the mapping between from input protein representation to the target function predictions.



```
UniProtKB      A0A009IHW8      J512_3302
     GO:0019677      PMID:29395922   IDA
     P       NAD(+) hydrolase AbTIR  J512_3302
    protein taxon:1310613   20191128      UniProt
```

Figure 1. AFP algorithms typically consist of a protein representation (protein sequence, expression data, biological networks) (left, in blue), a function representation (often a vector of one-hot encoded GO terms) (right, in black) and a prediction algorithm that connects both (neural networks, support vector machines, Guilt-By-Association methods etc.) (middle, in green).

2. Tissue/Condition-Specificity

Simply assigning a set of GO terms to a protein is often not enough. For instance, it does not necessarily provide information about whether the protein performs this function in specific tissues or under specific conditions, which is especially important for the Biological Process Ontology (BPO). For example, from gene expression experiments, it is known that genes can change co-expression partners [16] and regulators [17] from tissue to tissue [16], when under stress [18] or at different developmental stages [19]. Also, tissue-specific protein-protein interactions are known [20]. In other words, although a protein can be involved in multiple biological processes, it doesn't have to execute all these functions at all times, which has two implications. On one hand, it will be difficult to validate predicted annotations when it is not known for which tissues or cell types the protein performs this function. On the other hand, we need to have tissue or cell-type specific information of the activity of the protein (such as mRNA expression levels) to be able to discover these functions. Greene et al. have demonstrated the importance of this issue, by constructing tissue-specific co-functional networks from existing GO annotations and tissue-specific gene expression information, leading to more accurate predictions of response to perturbation and discovery of gene-disease associations [21]. All of this is made even harder, as, at this point, there is not even a clear definition of a cell type. The Cell Ontology (CO) is a good step towards standardizing cell type definitions, but it is only restricted to animal cells [22]. Despite this, we believe that AFP researchers and curators of the GO need to start preparing for a possible transition to this more specialized AFP phase.

2.1. Protein Representation

To make protein representation tissue-specific, Zitnik and Leskovec adapted the node2vec method to extract tissue-specific protein embeddings from tissue-specific Protein-Protein Interaction (PP) networks which were then fed to a linear classifier to predict function [23]. It would be interesting to extend this approach to co-expression networks, which are probably a lot more variable. For cases where co-expression or interaction evolves over time, e.g., for developmental processes or stress responses, it might be helpful to look at dynamic network embedding algorithms, such as dynnode2vec [24] that can learn condition-specific node embeddings without prior knowledge and more efficiently than the approach of [23]. Advances in single-cell sequencing [25] and the generation of cell atlases [26,27] are expected to elucidate even more subtleties of protein function and thus provide a valuable resource for more fine-grained functional annotation. Being able to represent genes by their expression and/or methylation pattern [25] in millions of cells and not by the average of those quantities, as is done with bulk sequencing, can help us find rare but also specific gene functions. On the other hand, this creates computational challenges, as one single-cell experiment can nowadays generate data for millions of cells. Integrating data from multiple such experiments will require the use of techniques specialized for processing 'big data'.

2.2. Function Representation

Unfortunately, predicting cell-type-specific and/or condition-specific function is a lot harder, as the number of possible outcomes increases in a combinatorial fashion. The number of GO terms is already very large, so creating a separate target variable for each combination of GO term, cell-type and condition would be intractable. Also, this would make the set of existing annotations even sparser posing severe problems for learning algorithms [28]. The Gene Ontology Consortium [29] uses so-called annotation extensions [30] to specify details about an annotation (including that the function occurs at a specific cell type) instead of creating a new GO term for each combination of function and cell type. However, the number of such extensions is also very large, so this does not solve the issue. Perhaps we are in need of a more fundamental representation of BPO functions. For example, for Molecular Function Ontology (MFO) terms, protein domains are often used as clear representatives of function, as specific domains correspond to specific 3D folding patterns that enable specific chemical reactions and therefore can be accurately associated with a molecular function. Certainly, domains can also be associated with BPO terms, but there rarely is a clear causal link. For instance, a DNA-binding domain might indicate that a protein could be a transcription factor, but that does not provide insights into the genes that this transcription factor regulates or the biological process(es) these genes are involved in. The introduction of Causal Activity Modelling (GO-CAM) [31] tries to address this issue and to unify GO terms by using causal graphs to model their interrelations. Alternatively, markers, such as DNA methylation, chromatin accessibility, and transcription factor binding can be used as a tissue-specific function representation, as they describe a gene's regulation, which might provide information about its functions. The different data types can be integrated with appropriate approaches (e.g., [32]) to identify their common and independent components.

2.3. Prediction Methods

When the number of labels increases dramatically, model learning should also be adapted to handle this increase. A promising option would be to move from a discrete to a continuous representation of the conditions. Way and Greene have demonstrated this as a proof of principle by training a Variational Auto-Encoder (VAE) on gene expression data from different cancer types [33]. They then showed that the latent space learned during the unsupervised training contained directions that encoded important information such as gender, tissue of origin and presence or absence of a metastasis [33]. Further work needs to be done on the interpretation of such models, so that we can make use of the latent encoding of different conditions for predicting function. Since VAE's are generative models, they could perhaps also generate predicted gene expression data for combinations of tissues and stresses that are not in the training set. Other generative models, such as Generative Adversarial Networks (GANs) [34] could also be used and specifically conditional GANs [35] are designed to generate data for a specific condition given a (discrete or continuous) numerical representation of that condition. We believe that this line of research will become popular in the near future. GANs have already been used in AFP to generate artificial data to counter class imbalance, thereby performing data augmentation [36] and leading to increased performance [37]. Incorporating the 'grammar' of GO-CAM relations into prediction models will also be an interesting challenge. For example, genes or proteins can also be viewed as part of the GO-CAM causal graph, thereby transforming the AFP problem into a semi-supervised learning problem of predicting new edges in that graph, specifically edges connecting the genes with the terms.

3. Going beyond Model Species

The need for accurate AFP is especially pressing for non-model species. Plants are interesting in that regard, as experimentally-derived functions in most plants are either very sparse or non-existent and the huge number of genes per species (e.g., more than 100,000 in wheat [38]) means that genome-wide experimental annotation would require vast amounts of time and resources. On the other hand, a lot of labeled data are required to train, as well as to test an AFP algorithm. Currently, labeled data come mostly from model species, as proteins from ten species account for 86–88% of the experimental GO annotations in UniProtKB, depending on the ontology (Figure 2).

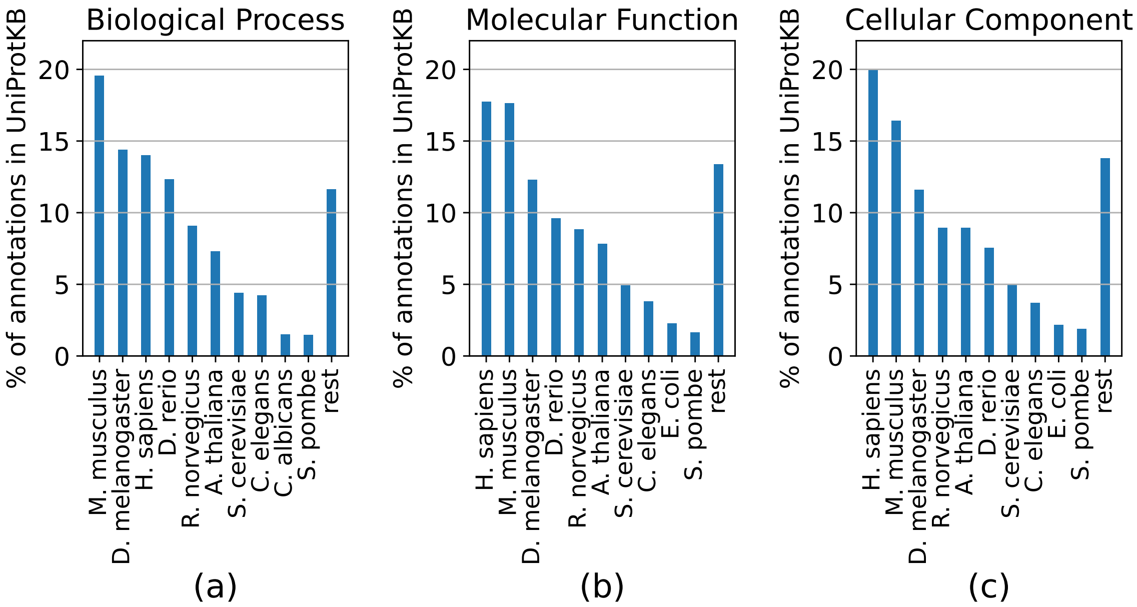


Figure 2. Distribution of experimental annotations from the Biological Process (**a**), Molecular Function (**b**) and Cellular Component (**c**) ontologies per species for proteins in UniProtKB. The ten species with the most annotations are shown for each ontology and annotations for all other species are shown in the 'rest' group.

One of the findings of the CAFA challenges [10–12] is that ensemble methods that combine predictions from many data sources tend to perform very well (e.g., MS-kNN [15] in CAFA2; and GOLabeler [3] and INGA [5] in CAFA3). Although CAFA is extremely useful, the evaluations rely on recent experimental annotations and these, by definition, are in their vast majority from 10–15 model species (Figure 3), because most experimental biologists work on those species. Besides plants, bacteria and archaea are also largely underrepresented in CAFA benchmarks (Figure 3). This focus on model species might hide the fact that perhaps some of the algorithms that are successful in CAFA might not be directly or fully applicable in non-model species, as for newly-sequenced species, typically, only DNA and protein sequences are available. We do by no means attempt to diminish the usefulness and impact of researching multi-omics ensemble methods, but it is important to realize that models that rely on protein-protein interactions or co-expression across different conditions are thus not applicable in the vast majority of species across all domains of life. This is especially important when predicting BPO terms, as these models have been shown to rely more on non-sequence-based data sources [12].

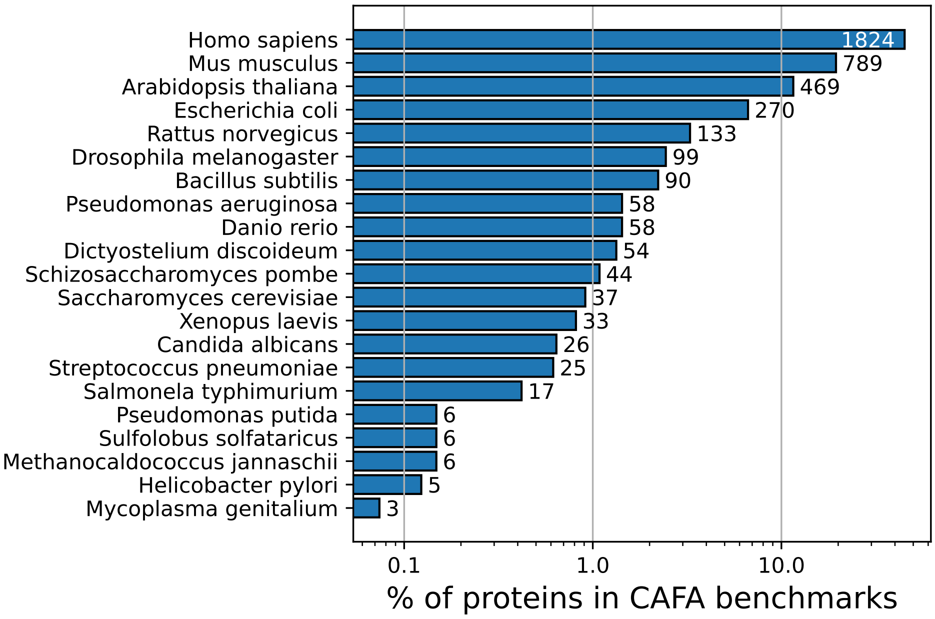


Figure 3. Percentage of proteins used in the three CAFA benchmarks [10–12] (*x*-axis, in log scale) per species. The absolute number of proteins per species is also given next to the bars. Only newly-annotated proteins are included, i.e., proteins that had no GO annotations before the benchmark (referred to as No-Knowledge benchmarks in CAFA). CAFA: Critical Assessment of Functional Annotation.

3.1. Protein Representation

A big question for AFP is whether sequence-only methods can achieve as high performance as methods that use multi-omic data. Recent work has shown that it is possible to accurately predict PPI's from amino-acid sequence [39] or gene expression from DNA sequence [40], which implies that a big part of the multi-omic data is encoded in the sequence data, albeit in a more complicated manner. In other words, one could improve existing techniques that predict multi-omic data and use those predictions for AFP. But, this might also imply that sequence-only models might have the potential to perform equally well to the ensemble methods. Alternatively, as more and more species are being sequenced, finding orthologs or constructing high-quality multiple sequence alignments of proteins will get easier and easier. That, in turn, would imply more robust and accurate old-school, sequence-based annotation transfers with ever-increasing coverage. Essential wet-lab experiments could in that scenario be focused on the ever-decreasing set of proteins without close homologues.

3.2. Function Representation

Another big challenge of predicting protein functions in other species is that certain functions might be more prevalent or even unique in certain lineages. As an extreme example: it is not trivial to predict photosynthesis-related functions when training only on animal data. It would therefore be beneficial to have a function representation that allows extrapolating to new functions. To do so, we need our representation to be agnostic of the training annotations and capture more general functional aspects. This can be done by learning an embedding for each GO term so that terms that describe related functions have high similarity in the embedding space. That would enable us to extrapolate the meaning of terms beyond the training set. Several such representations have been proposed, one of the first ones being clusDCA [41], which used random walks to learn features that reflect the GO graph topology. More recent approaches make use of advances in Natural Language Processing (NLP) to learn embeddings that reflect semantic meanings based on the term names and/or descriptions [42]. Theoretical work has shown the utility of embedding graph-structured data (as GO terms are) in hyperbolic rather than Euclidean spaces [43].

3.3. Prediction Methods

Prediction models across species should mainly deal with two issues: (1) the presence of novel functions, as described above, and (2) the potential differences in distribution of the input data between species. There is a vast machine learning literature on few-shot and zero-shot learning, which deals with classification models that can make predictions for classes for which only very few or even no examples have been seen, respectively [44,45]. Such methods often tend to use class embeddings [46] and try to leverage prior knowledge on similarities between the classes. A similar approach has been applied in predicting novel cell types from gene expression data using Cell Ontology [22] embeddings [47]. Such approaches can additionally be useful for describing new terms that are occasionally added to the ontology, even before they accumulate many annotations. As for the difference in distributions, also known as domain shift [48], it can lead to large performance loss if not taken into account. As an example, methods that use the frequency of amino acids or amino acid n-grams in a sequence can be sensitive to amino acid frequency differences across lineages (e.g., [49]). Given certain assumptions about the type of domain shift, there exist different approaches for correcting it [48]. Even if theoretical assumptions are not met, however, domain adaptation can still give a performance boost (e.g., [50]), so even then it might still be worth attempting to detect and correct domain shifts.

4. Overlooked Data Sources

The amino acid sequence is the most widely-used data source for function prediction, followed by sequence-derived features such as domains as well as other omics data, such as gene expression or protein-protein interactions. There is, however, a wealth of other omics data that has the potential to predict function accurately, but that is currently hardly used. Leveraging these data could also boost performance of AFP algorithms. For the sake of brevity, we focus on three such omics data sources: proteomics, genome proximity, and epigenetic data. In addition, we also address the utility of literature mining.

4.1. Protein Representation

The power of gene expression data in function prediction has been well-documented in the CAFA challenges, especially for predicting BPO terms [12]. Interestingly, a study from 2017 found that co-expression networks constructed from proteomics rather than transcriptomics were more efficient at predicting both GO terms and pathways from the Kyoto Encyclopedia of Genes and Genomes (KEGG) in *Homo sapiens* [51]. Furthermore, mRNA and proteomics have been shown to contain complementary information [52–54] (Figure 4), caused by–amongst others–differences in degradation rates and translation speeds. Hence, protein abundances as measured, for example, by mass spectrometry give a more representative picture of the function of a protein than their mRNA proxies. Integrating these two omics data types could boost prediction accuracy. In addition, the relationship between genes and proteins is not 1-to-1, as RNA splicing [55] as well as post-translational modifications (PTM's) [56] cause the encoding of multiple protein isoforms that have been shown to have different roles and functions [57]. This introduces an additional layer of complexity when trying to predict the function of genes, that can be partly addressed by measuring the expression of isoforms using RNAseq. However, different PTM's can only be measured using proteomics techniques. As the quality, reproducibility, scalability and accessibility of mass spectrometry methods keep increasing [58], we expect proteomics to start playing a more and more prominent role in AFP.

Proximity between genes has also been shown to be indicative of co-functionality [59]. This is particularly true for bacteria, where genes from the same pathway are often organized in operons, but it was recently shown to be applicable in eukaryotes as well [60]. Except for linear proximity, the same holds for genes that are close in three-dimensional space, as these tend to be co-regulated [61] (Figure 4). Chromosome conformation data, generated using e.g., 4C [62] or Hi-C [63] techniques, have been used in a human protein function prediction pipeline [64], but they are certainly not exploited enough.

Epigenetic markers, such as DNA methylation or chromosome accessibility, affect gene regulation, implying that genes that have similar epigenetic patterns across tissues might be regulated jointly. This makes epigenetic data potentially a rich resource for predicting Biological Process terms, although it is not known to what extent this signal is complementary to gene expression. Human and mouse are the two species where this hypothesis can be easily tested as many of their genes have well-documented functions and the ENCODE project has generated large amounts of epigenetic data in both species [65].

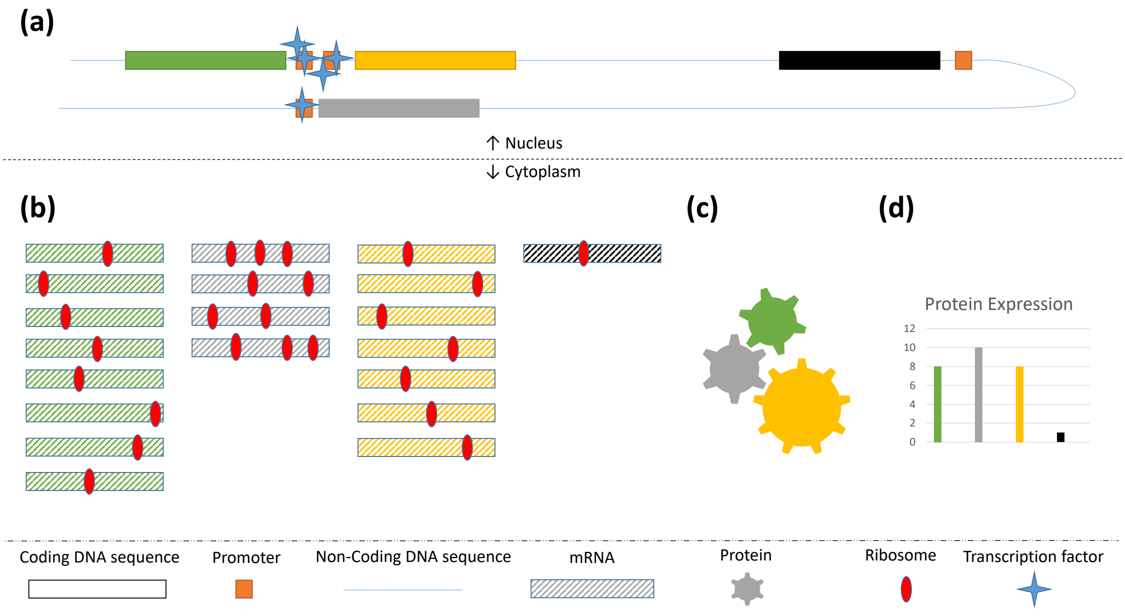


Figure 4. Genes that are close to each other in space are likely to be co-expressed, because transcription factors that bind on a promoter are likely to also bind to nearby promoters (**a**). Differences in expression are indicated by the amount of transcripts in the cytoplasm (**b**). In this case the green, yellow and gray genes code for proteins (shown as gears of the same color) that form a protein complex and perform a function together (**c**). The ribosome occupancy of the mRNAs is used to indicate differences in translation efficiency which result in similar abundances for proteins (**d**) despite different mRNA abundancies.

Finally, mining of scientific literature has been under-used in the past years, but a recent study showed that it can have competitive performance [66]. The idea behind the usage of such text mining methods is that if two genes are often mentioned together in publications, then they are likely to be involved in the same function. In addition, co-occurrence of gene names with other words, such as disease or pathway names [67], can be informative. As successful NLP models are starting to be applied in biomedical literature data [67,68], we expect the role of text mining in AFP to increase in the immediate future.

4.2. Function Representation

Using new data sources in both computational and experimental protein annotation might call for new evidence codes. For instance, if indeed protein co-expression is much more relevant for co-functionality than mRNA co-expression, it might make sense to differentiate between functions discovered using the two technologies by splitting the HEP (High-throughput Expression Pattern) evidence code. For similar epigenetic profiles and co-regulation in general, there is also no appropriate evidence code. The closest one is perhaps IGI (Inferred from Genetic Interaction), but this mainly refers to one gene influencing another gene (e.g., by changes in expression or mutations), while co-regulation implies that both genes are jointly regulated by the same mechanism.

4.3. Prediction Methods

Arguably one of the most important recent methodological leaps is representation learning. Advances in machine learning have been quickly adopted by AFP researchers, causing a shift of the research efforts from the guilt-by-association (GBA) paradigm to automatic representation learning. A lot of recent works, often inspired by natural language processing, use convolutional or recurrent neural networks to automatically learn sequence features useful for predicting function. Such methods have been shown to work better than simple homology search (e.g., [69]). Also, several neural embedding methods have been recently proposed which can learn complex features for nodes of networks. Such methods have been applied to both PPI and co-expression networks and shown to be useful for reconstructing functional relationships [70,71]. However, there is still not enough evidence of whether such methods can outperform simple GBA methods, such as gene co-expression based on Pearson correlation, which is very effective for BPO predictions [12].

5. Biased and Missing Annotations

Another unresolved question that is important to address is how biases in GO annotations influence AFP. One extreme form of such a bias is missing annotations. Missing annotations in the test set do not have a dramatic effect on the ranking of AFP methods [72], but the effect of missing annotations in the training set has not yet been systematically quantified. We suspect that this effect is larger, especially for machine-learning-based methods that try to learn characteristics of proteins with the same function.

5.1. Protein Representation

What is often overlooked is that biases in the generation of annotations might also affect the protein representation. For example, in plant research, scientists are very often interested in stress responses and flowering, so a lot of the available gene expression data come from such conditions, meaning that it might be harder to infer other functions for which very little data are available. The same holds for other species. For example, *Drosophila melanogaster* is typically used to study genetics [73], *Caenorhabditis elegans* to study development [74] and so on. The detection of condition-specific or tissue-specific protein-protein interaction suffers from the same issue. Protein sequence data are also not completely 'safe' from this bias, as it is possible that a functional isoform of a protein used in rare, poorly-studied conditions is not known. However, for sequence data, the effect of that bias is arguably smaller than for other data types.

5.2. Function Representation

Many researchers tend to only include *experimental* evidence codes (EC's) when training and testing AFP methods to avoid biases. This has the downside that a great amount of knowledge about protein function is potentially ignored. Moreover, experimental EC's have also been shown to be highly biased [75], leading to the recent split between "experimental" and "high-throughput experimental" EC's. Also, experimental EC's include co-expression and protein interaction experiments which might introduce circular reasoning for algorithms that use such data types, in the same way that sequence similarity EC's introduce bias for sequence-based algorithms. On the other hand, annotations that are automatically generated by a curated set of rules are labeled "IEA" and are often ignored, although they are rather reliable, given that the rules are made by expert curators [76]. We cannot convince experimentalists to start randomly selecting a protein and testing it for random functions to obtain an unbiased ground-truth with independent, identically distributed observations. Nor can we ever obtain a reliable, experimentally-derived set of negative annotations, as a protein might have a particular function only under certain conditions. Therefore, it might be interesting to try to further quantify the biases introduced by including certain EC's. Previous work quantified the quality of automatic annotations by measuring to what extent they were later experimentally confirmed [76].

Additionally, one could look for possible changes in the performance of the naïve classifier when adding or removing annotations with a specific EC. Also, one could compare the performance of AFP methods for each evidence code separately to assess these biases better.

Some studies have attempted to tackle the missing annotations problem by generating negative examples [77,78]. Somewhat surprisingly, these datasets of negative annotations have not gained popularity among AFP researchers, as they are most often not used during the training of new methods. To our knowledge, there is no study providing evidence against using these data, so we believe that this topic deserves further investigation. A recent study showed that ignoring negative annotations at test time can lead to misleading evaluations [79].

5.3. Prediction Methods

Alternatively to generating negative annotations, we could also change the loss functions used to train AFP models. Most recent machine learning models are trained by minimizing a cross-entropy loss, which assumes no missing labels. That could be replaced by a loss for Positive-Unlabeled (PU) learning [80]. Again, these loss functions are well-known in AFP and related fields [77,81], but they seem not to be very popular. It is possible, that despite their obvious theoretical benefits, such losses do not work in practice for AFP and publication bias has prevented us from seeing these results. A different approach would be to introduce probabilistic labels, where annotations from EC's that are considered reliable are attributed with high certainty, but unreliable ones (e.g., derived automatically by another AFP method and never verified by a curator) are used in the training set but with a low-probability. One could additionally always assign a non-zero probability to all other terms to account for missing annotations. Such a probabilistic ground-truth means that probabilistic learning algorithms will be needed, such as the graphical model proposed in [82].

6. Gene Ontology

GO is a very useful resource that describes biological function in a standardized manner that is human- and computer-readable. This has led to its nearly catholic acceptance as the go-to functional representation, to the point that function prediction is almost a synonym of GO term prediction.

6.1. Protein Representation

As stated above, automatic representation learning is a very promising direction. It can be further enhanced by unsupervised pre-training, where a generic protein representation is learned [83–86] from all available sequences. This representation can then be used to predict GO terms [87]. But end-to-end training is also possible, where the weights of the unsupervised feature extractor are also fine-tuned to create an ontology-specific feature representation designed for predicting GO terms from that ontology. This can lead to better performance, especially when only few labeled examples are available. Such approaches are currently the state-of-the-art in Natural Language Processing [88].

6.2. Function Representation

We previously touched upon embedding GO terms in a vector space and its potential in creating a new functional representation that is simpler but reflects the same semantic relationships as the GO graph. Figure 5 shows an example of such an embedding in two dimensions that reflects term co-occurence patterns. The biggest downside of those approaches is the loss of the interpretability and human-readability that GO terms have. It is still not trivial to always provide a biological interpretation for a given set of GO terms (even though visualization techniques are helping in that regard [89]), but it is still relatively easy to understand the meaning of an individual term by its name, description, and connections to ancestors and descendants. On the other hand, representing terms as high-dimensional vectors makes us lose the intuition, which implies that we would like this representation to be invertible, i.e., also provide us with a rule to convert a given vector in this "functional space" back to a term or a set of terms, ideally in a unique way. This is possible for linear mappings and we and others have worked on such approaches [90,91]. Linear approaches can capture simple relationships between terms such as co-occurrence or mutual exclusivity of a pair of terms [90], but might struggle to find more complicated relationships "hidden" either in the graph or in the semantics of terms. We therefore suspect that non-linear (e.g., neural) term embeddings are required to capture the whole structure. Nevertheless, we think that substantial emphasis and attention should be put on maintaining the interpretability of these models. This is nowadays also a hot topic in machine learning and computer vision [92–94], which has been dominated by neural networks in the past decade.

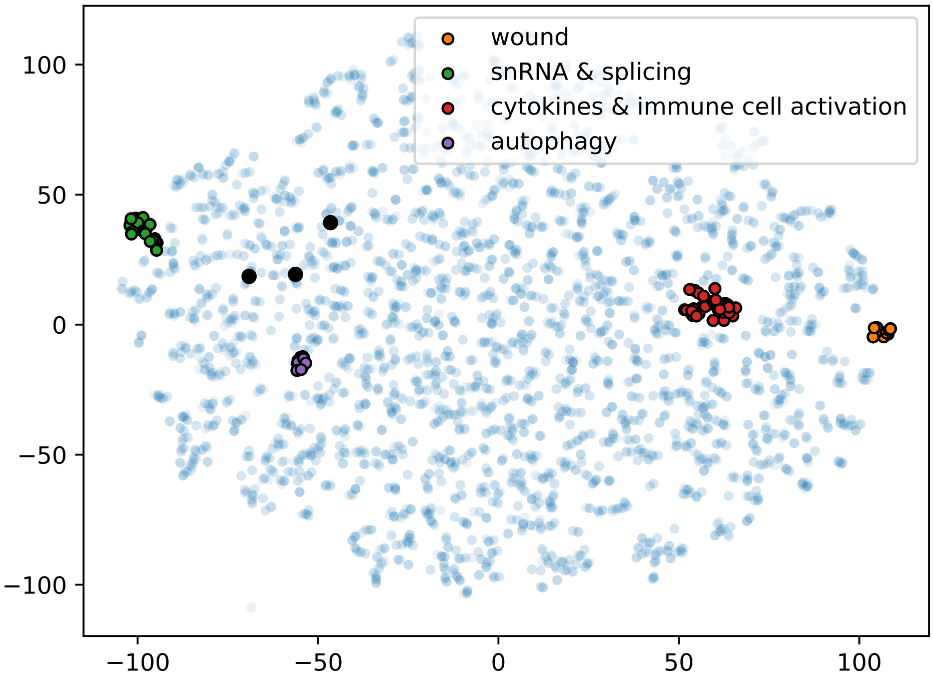


Figure 5. Two-dimensional tSNE embedding of GO terms from all three ontologies that annotate at least 0.1% of SwissProt entries. One minus the Pearson correlation of the occurence patterns of terms across SwissProt proteins was used as a distance measure for calculating the embeddings. Inspecting the terms in some of the clusters observed in this 2D space revealed that terms from the same cluster have similar meanings. Examples of clusters with terms that refer to wound healing, small nuclear RNAs and mRNA splicing complexes, cytokines and immune activation, and autophagy are shown in orange, green, red, and purple respectively. Terms "DNA binding", "nucleus" and "DNA-templated regulation of transcription" are shown as larger black dots.

6.3. Prediction Methods

The three different ontologies contain correlated information. For example, "DNA binding" (MFO) co-occurs with "DNA-templated regulation of transcription" (BPO, $\rho = 0.54$) and "nucleus" (CCO, $\rho = 0.35$) (Figure 5). GO curators are well aware of these co-occurrences and have hand-crafted rules to automatically transfer such annotations. However, little computational work has been done to improve upon this rule-based system, although this seems like a very promising direction, especially for the Limited-Knowledge category of CAFA, where the goal is to predict the functions of proteins in one ontology using its functions in others. A possible approach would be to embed the terms of each ontology to a vector space and then try to align the three ontologies, with the aim of discovering new cross-ontology similarities that are not obvious to the curators and could be exploited by function prediction algorithms. In Onto2vec, the authors learned a joint embedding for terms from all three ontologies and proteins [95]. They used it to calculate protein similarities for the downstream task of protein interaction prediction, but it would also be very interesting to examine similarities in the embedding space between pairs of GO terms to identify potentially unknown correlations.

Over time, unannotated or partly annotated genes obtain new GO annotations, i.e., the ground-truth data that can be used to train AFP models changes. Therefore, it would be interesting to have models which can incorporate such new knowledge without having to re-train from scratch. This could be achieved by using online machine learning algorithms [96,97].

7. Evaluation of AFP Algorithms

Next to the challenges described above, the problem of properly evaluating AFP models, as already initially identified as one of the three main challenges by Godzik et al. [13], is still lingering. It has been shown that temporal hold-out evaluation strategies, like the CAFA challenges, give more realistic estimates of the performance on new unseen data than cross-validation [98]. Given the success of omics data in CAFA—and specifically in CAFA-π [12]—we believe that a GBA-based baseline should be included in future CAFA editions, for example a GBA based on PPI's from e.g., the STRING database [99]. This is to be preferred over a co-expression-based baseline as for many species this data is not readily available.

By comparing the rankings of methods in CAFA3 [12] across five evaluation metrics, we found that most widely-used metrics are highly correlated, with the Semantic Distance [100] being slightly different from the rest (Figure 6). However, a more recent simulation study questioned the validity of these commonly used evaluation measures. The authors also proposed novel measures that more accurately reflect the quality of the predictions [101]. The complicated and biased nature of GO annotations can indeed lead to misleading evaluations, so having appropriate evaluation measures is essential to ensure that the field keeps going forward.

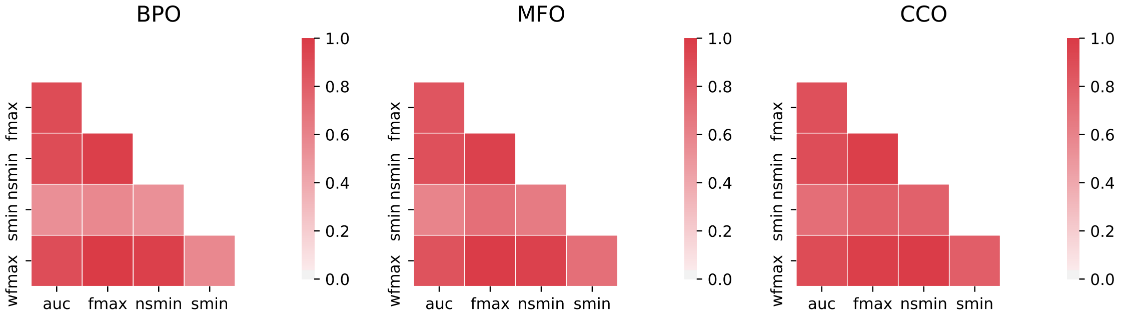


Figure 6. Pairwise absolute rank correlations between five different protein-centric evaluation metrics: area under the precision-recall curve (auc), maximum F1 score (fmax), maximum F1 score weighted by information content (wfmax), minimum semantic distance (smin), and minimum normalized semantic distance (nsmin). Correlation was calculated from the 146 methods that participated in CAFA3 [12] for the biological process (**left**), molecular function (**middle**) and cellular component (**right**) ontologies. More intense red color denotes larger absolute correlation. All pairwise correlations are statistically significant with uncorrected p-values $< 10^{-11}$.

8. Conclusions

AFP remains one of the most challenging bioinformatics tasks and despite its growing interest and recent progress, it is far from being completely solved. Here, we addressed some of the current and future challenges of the field. In our view, significant breakthroughs are expected mainly from the use of neural embeddings (for describing both proteins and GO terms) and the use of new technologies, such as proteomics. The problem of predicting function for non-model species is possibly the most challenging, as it may still require generation of experimental data. However, as the community of AFP researchers is growing [12], we are optimistic that these challenges will soon be tackled.

Author Contributions: Conceptualization, R.C.H.J.v.H.; investigation, S.M.; writing—original draft preparation, S.M.; writing—review and editing, M.J.T.R., R.C.H.J.v.H.; supervision, M.J.T.R., R.C.H.J.v.H.; funding acquisition, R.C.H.J.v.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded—in part—by Keygene N.V., an AgBiotech company in The Netherlands.

Acknowledgments: S.M. would like to thank Soufiane Mourragui, Jesse Krijthe, and Tamim Abdelaal for the useful discussions about the manuscript.

Conflicts of Interest: The funder provided support in the form of salaries for S.M. and R.C.H.J.v.H., but had no role in the design of the study, in the writing of the manuscript, or in the decision to publish.

References

1. Bateman, A. UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res.* **2019**, *47*, D506–D515. [CrossRef]
2. González-Castro, T.B.; Tovilla-Zárate, C.A.; Genis-Mendoza, A.D.; Juárez-Rojop, I.E.; Nicolini, H.; López-Narváez, M.L.; Martínez-Magaña, J.J. Identification of gene ontology and pathways implicated in suicide behavior: Systematic review and enrichment analysis of GWAS studies. *Am. J. Med. Genet. Part B Neuropsychiatr. Genet.* **2019**, *180*, 320–329. [CrossRef] [PubMed]
3. You, R.; Zhang, Z.; Xiong, Y.; Sun, F.; Mamitsuka, H.; Zhu, S. GOLabeler: Improving sequence-based large-scale protein function prediction by learning to rank. *Bioinformatics* **2018**, *34*, 2465–2473. [CrossRef] [PubMed]
4. Das, S.; Lee, D.; Sillitoe, I.; Dawson, N.L.; Lees, J.G.; Orengo, C.A. Functional classification of CATH superfamilies: A domain-based approach for protein function annotation. *Bioinformatics* **2015**, *31*, 3460–3467. [CrossRef] [PubMed]
5. Piovesan, D.; Tosatto, S.C.E. INGA 2.0: Improving protein function prediction for the dark proteome. *Nucleic Acids Res.* **2019**, *47*, W373–W378. [CrossRef]
6. Jain, A.; Kihara, D. Phylo-PFP: Improved automated protein function prediction using phylogenetic distance of distantly related sequences. *Bioinformatics* **2018**, *35*, 753–759. [CrossRef]
7. Zhang, C.; Freddolino, P.L.; Zhang, Y. COFACTOR: Improved protein function prediction by combining structure, sequence and protein–protein interaction information. *Nucleic Acids Res.* **2017**, *45*, W291–W299. [CrossRef]
8. You, R.; Yao, S.; Xiong, Y.; Huang, X.; Sun, F.; Mamitsuka, H.; Zhu, S. NetGO: Improving large-scale protein function prediction with massive network information. *Nucleic Acids Res.* **2019**, 47, W379–W387. [CrossRef]
9. Kulmanov, M.; Hoehndorf, R. DeepGOPlus: Improved protein function prediction from sequence. *Bioinformatics* **2019**, *36*, 422–429. [CrossRef]
10. Radivojac, P.; Clark, W.T.; Oron, T.R.; Schnoes, A.M.; Wittkop, T.; Sokolov, A.; Graim, K.; Funk, C.; Verspoor, K.; Ben-Hur, A.; et al. A large-scale evaluation of computational protein function prediction. *Nat. Methods* **2013**, *10*, 221–227. [CrossRef]
11. Jiang, Y.; Oron, T.R.; Clark, W.T.; Bankapur, A.R.; D'Andrea, D.; Lepore, R.; Funk, C.S.; Kahanda, I.; Verspoor, K.M.; Ben-Hur, A.; et al. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome Biol.* **2016**, *17*, 184. [CrossRef]
12. Zhou, N.; Jiang, Y.; Bergquist, T.R.; Lee, A.J.; Kacsoh, B.Z.; Crocker, A.W.; Lewis, K.A.; Georghiou, G.; Nguyen, H.N.; Hamid, M.N.; et al. The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome Biol.* **2019**, *20*, 1–23. [CrossRef] [PubMed]
13. Godzik, A.; Jambon, M.; Friedberg, I. Computational protein function prediction: Are we making progress? *Cell. Mol. Life Sci.* **2007**, *64*, 2505. [CrossRef]
14. Cozzetto, D.; Buchan, D.W.; Bryson, K.; Jones, D.T. Protein function prediction by massive integration of evolutionary analyses and multiple data sources. *BMC Bioinform.* **2013**, *14*, S1. [CrossRef]
15. Lan, L.; Djuric, N.; Guo, Y.; Vucetic, S. MS-kNN: Protein function prediction by integrating multiple data sources. *BMC Bioinform.* **2013**, *14* (Suppl. 3), S8. [CrossRef] [PubMed]

16. Farahbod, M.; Pavlidis, P. Differential coexpression in human tissues and the confounding effect of mean expression levels. *Bioinformatics* **2019**, *35*, 55–61. [CrossRef]
17. Sonawane, A.R.; Platig, J.; Fagny, M.; Chen, C.Y.; Paulson, J.N.; Lopes-Ramos, C.M.; DeMeo, D.L.; Quackenbush, J.; Glass, K.; Kuijjer, M.L. Understanding Tissue-Specific Gene Regulation. *Cell Rep.* **2017**, *21*, 1077–1088. [CrossRef]
18. Jiang, Z.; Dong, X.; Li, Z.G.; He, F.; Zhang, Z. Differential coexpression analysis reveals extensive rewiring of arabidopsis gene coexpression in response to pseudomonas syringae infection. *Sci. Rep.* **2016**, *6*, 35064. [CrossRef]
19. Singh, A.J.; Ramsey, S.A.; Filtz, T.M.; Kioussi, C. Differential gene regulatory networks in development and disease. *Cell. Mol. Life Sci.* **2018**, *75*, 1013–1025. [CrossRef] [PubMed]
20. Basha, O.; Shpringer, R.; Argov, C.M.; Yeger-Lotem, E. The DifferentialNet database of differential protein-protein interactions in human tissues. *Nucleic Acids Res.* **2018**, *46*, D522–D526. [CrossRef]
21. Greene, C.S.; Krishnan, A.; Wong, A.K.; Ricciotti, E.; Zelaya, R.A.; Himmelstein, D.S.; Zhang, R.; Hartmann, B.M.; Zaslavsky, E.; Sealfon, S.C.; et al. Understanding multicellular function and disease with human tissue-specific networks. *Nat. Genet.* **2015**, *47*, 569–576. [CrossRef] [PubMed]
22. Diehl, A.D.; Meehan, T.F.; Bradford, Y.M.; Brush, M.H.; Dahdul, W.M.; Dougall, D.S.; He, Y.; Osumi-Sutherland, D.; Ruttenberg, A.; Sarntivijai, S.; et al. The Cell Ontology 2016: Enhanced content, modularization, and ontology interoperability. *J. Biomed. Semant.* **2016**, *7*, 44. [CrossRef]
23. Zitnik, M.; Leskovec, J. Predicting multicellular function through multi-layer tissue networks. *Bioinformatics* **2017**, *33*, i190–i198. [CrossRef]
24. Mahdavi, S.; Khoshraftar, S.; An, A. Dynnode2vec: Scalable Dynamic Network Embedding. In Proceedings of the 2018 IEEE International Conference on Big Data, Big Data 2018, Seattle, WA, USA, 10–13 December 2018; pp. 3762–3765. [CrossRef]
25. Jaitin, D.A.; Kenigsberg, E.; Keren-Shaul, H.; Elefant, N.; Paul, F.; Zaretsky, I.; Mildner, A.; Cohen, N.; Jung, S.; Tanay, A.; et al. Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types. *Science* **2014**, *343*, 776–779. [CrossRef]
26. Papatheodorou, I.; Moreno, P.; Manning, J.; Fuentes, A.M.P.; George, N.; Fexova, S.; Fonseca, N.A.; Füllgrabe, A.; Green, M.; Huang, N.; et al. Expression Atlas update: From tissues to single cells. *Nucleic Acids Res.* **2019**, *48*, D77–D83. [CrossRef]
27. Thul, P.J.; Lindskog, C. The human protein atlas: A spatial map of the human proteome. *Protein Sci.* **2018**, *27*, 233–244. [CrossRef]
28. Japkowicz, N.; Stephen, S. The class imbalance problem: A systematic study. *Intell. Data Anal.* **2002**, *6*, 429–449. [CrossRef]
29. GO Consortium. Guide to GO Evidence Codes. 2016. Available online: http://geneontology.org/page/guide-go-evidence-codes (accessed on 30 July 2020).
30. Annotation Extension. Available online: http://wiki.geneontology.org/index.php/Annotation_Extension (accessed on 12 September 2020).
31. Thomas, P.D.; Hill, D.P.; Mi, H.; Osumi-Sutherland, D.; Van Auken, K.; Carbon, S.; Balhoff, J.P.; Albou, L.P.; Good, B.; Gaudet, P.; et al. Gene Ontology Causal Activity Modeling (GO-CAM) moves beyond GO annotations to structured descriptions of biological functions and systems. *Nat. Genet.* **2019**, *51*, 1429–1433. [CrossRef] [PubMed]
32. Lock, E.F.; Hoadley, K.A.; Marron, J.S.; Nobel, A.B. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *Ann. Appl. Stat.* **2013**, *7*, 523–542. [CrossRef] [PubMed]
33. Way, G.P.; Greene, C.S. Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. *Pac. Symp. Biocomput. Pac. Symp. Biocomput.* **2018**, *23*, 80–91.
34. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27*; Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2014; pp. 2672–2680.
35. Mirza, M.; Osindero, S. Conditional Generative Adversarial Nets. *arXiv* **2014**, arXiv: 1411.1784.

36. Perez, L.; Wang, J. The Effectiveness of Data Augmentation in Image Classification using Deep Learning. *arXiv* **2017**, arXiv:1712.04621.
37. Wan, C.; Jones, D.T. Protein function prediction is improved by creating synthetic feature samples with generative adversarial networks. *Nat. Mach. Intell.* **2020**, *2*, 540–550. [CrossRef]
38. Appels, R.; Eversole, K.; Stein, N.; Feuillet, C.; Keller, B.; Rogers, J.; Pozniak, C.J.; Choulet, F.; Distelfeld, A.; Poland, J.; et al. Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* **2018**, *361*, 661. [CrossRef]
39. Richoux, F.; Servantie, C.; Borès, C.; Téletchéa, S. Comparing two deep learning sequence-based models for protein-protein interaction prediction. *arXiv* **2019**, arXiv:1901.06268.
40. Sigalova, O.M.; Shaeiri, A.; Forneris, M.; Furlong, E.E.; Zaugg, J.B. Predictive features of gene expression variation reveal a mechanistic link between expression variation and differential expression. *bioRxiv* **2020**. [CrossRef]
41. Wang, S.; Cho, H.; Zhai, C.; Berger, B.; Peng, J. Exploiting ontology graph for predicting sparsely annotated gene function. *Bioinformatics* **2015**, *31*, i357–i364. [CrossRef] [PubMed]
42. Duong, D.; Uppunda, A.; Gai, L.; Ju, C.; Zhang, J.; Chen, M.; Eskin, E.; Li, J.J.; Chang, K.W. Evaluating Representations for Gene Ontology Terms. *bioRxiv* **2020**. [CrossRef]
43. Chamberlain, B.P.; Clough, J.; Deisenroth, M.P. Neural Embeddings of Graphs in Hyperbolic Space. *arXiv* **2017**, arXiv:1705.10359.
44. Li, X.; Sun, Z.; Xue, J.H.; Ma, Z. A Concise Review of Recent Few-shot Meta-learning Methods. *arXiv* **2020**, arXiv:2005.10953.
45. Xian, Y.; Schiele, B.; Akata, Z. Zero-Shot Learning—The Good, the Bad and the Ugly. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
46. Huynh, D.; Elhamifar, E. Fine-Grained Generalized Zero-Shot Learning via Dense Attribute-Based Attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020.
47. Wang, S.; Pisco, A.O.; McGeever, A.; Brbic, M.; Zitnik, M.; Darmanis, S.; Leskovec, J.; Karkanias, J.; Altman, R.B. Unifying single-cell annotations based on the Cell Ontology. *bioRxiv* **2020**. [CrossRef]
48. Kouw, W.M.; Loog, M. An introduction to domain adaptation and transfer learning. *arXiv* **2018**, arXiv:1812.11806.
49. Kumar, V.; Sharma, A.; Kaur, R.; Thukral, A.K.; Bhardwaj, R.; Ahmad, P. Differential distribution of amino acids in plants. *Amino Acids* **2017**, *49*, 821–869. [CrossRef] [PubMed]
50. Munro, J.; Damen, D. Multi-Modal Domain Adaptation for Fine-Grained Action Recognition. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Korea, 27–28 October 2019; pp. 3723–3726.
51. Wang, J.; Ma, Z.; Carr, S.A.; Mertins, P.; Zhang, H.; Zhang, Z.; Chan, D.W.; Ellis, M.J.; Townsend, R.R.; Smith, R.D.; et al. Proteome profiling outperforms transcriptome profiling for coexpression based gene function prediction. *Mol. Cell. Proteom.* **2017**, *16*, 121–134. [CrossRef] [PubMed]
52. Griffin, T.J.; Gygi, S.P.; Ideker, T.; Rist, B.; Eng, J.; Hood, L.; Aebersold, R. Complementary Profiling of Gene Expression at the Transcriptome and Proteome Levels in Saccharomyces cerevisiae. *Mol. Cell. Proteom.* **2002**, *1*, 323–333. [CrossRef]
53. Wang, X.; Liu, Q.; Zhang, B. Leveraging the complementary nature of RNA-Seq and shotgun proteomics data. *Proteomics* **2014**, *14*, 2676–2687. [CrossRef] [PubMed]
54. Grabowski, P.; Kustatscher, G.; Rappsilber, J. Epigenetic Variability Confounds Transcriptome but Not Proteome Profiling for Coexpression-based Gene Function Prediction. *Mol. Cell. Proteom.* **2018**, *17*, 2082–2090. [CrossRef]
55. Wang, D.; Zou, X.; Fai Au, K. A network-based computational framework to predict and differentiate functions for gene isoforms using exon-level expression data. *Methods* **2020**. [CrossRef]
56. Perchey, R.T.; Tonini, L.; Tosolini, M.; Fournié, J.J.; Lopez, F.; Besson, A.; Pont, F. PTMselect: Optimization of protein modifications discovery by mass spectrometry. *Sci. Rep.* **2019**, *9*, 4181. [CrossRef]
57. Csizmok, V.; Forman-Kay, J.D. Complex regulatory mechanisms mediated by the interplay of multiple post-translational modifications. *Curr. Opin. Struct. Biol.* **2018**, *48*, 58–67. [CrossRef]

58. Müller, J.B.; Geyer, P.E.; Colaço, A.R.; Treit, P.V.; Strauss, M.T.; Oroshi, M.; Doll, S.; Virreira Winter, S.; Bader, J.M.; Köhler, N.; et al. The proteome landscape of the kingdoms of life. *Nature* **2020**, *582*, 592–596. [CrossRef]
59. Huynen, M.; Snel, B.; Lathe, W.; Bork, P. Predicting protein function by genomic context: Quantitative evaluation and qualitative inferences. *Genome Res.* **2000**, *10*, 1204–1210. [CrossRef] [PubMed]
60. Foflonker, F.; Blaby-Haas, C.E. Co-locality to co-functionality: Eukaryotic gene neighborhoods as a resource for function discovery. *Mol. Biol. Evol.* **2020**, msaa221. [CrossRef] [PubMed]
61. Schoenfelder, S.; Sexton, T.; Chakalova, L.; Cope, N.F.; Horton, A.; Andrews, S.; Kurukuti, S.; Mitchell, J.A.; Umlauf, D.; Dimitrova, D.S.; et al. Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. *Nat. Genet.* **2010**, *42*, 53–61. [CrossRef]
62. Zhao, Z.; Tavoosidana, G.; Sjölinder, M.; Göndör, A.; Mariano, P.; Wang, S.; Kanduri, C.; Lezcano, M.; Singh Sandhu, K.; Singh, U.; et al. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat. Genet.* **2006**, *38*, 1341–1347. [CrossRef]
63. van Berkum, N.L.; Lieberman-Aiden, E.; Williams, L.; Imakaev, M.; Gnirke, A.; Mirny, L.A.; Dekker, J.; Lander, E.S. Hi-C: A method to study the three-dimensional architecture of genomes. *J. Vis. Exp. JoVE* **2010**, 1869. [CrossRef]
64. Cao, R.; Cheng, J. Integrated protein function prediction by mining function associations, sequences, and protein-protein and gene-gene interaction networks. *Methods* **2016**, *93*, 84–91. [CrossRef]
65. Moore, J.E.; Purcaro, M.J.; Pratt, H.E.; Epstein, C.B.; Shoresh, N.; Adrian, J.; Kawli, T.; Davis, C.A.; Dobin, A.; Kaul, R.; et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **2020**, *583*, 699–710. [CrossRef] [PubMed]
66. You, R.; Huang, X.; Zhu, S. DeepText2GO: Improving large-scale protein function prediction with deep semantic text representation. *Methods* **2018**, *145*, 82–90. [CrossRef]
67. Chen, Q.; Lee, K.; Yan, S.; Kim, S.; Wei, C.H.; Lu, Z. BioConceptVec: Creating and evaluating literature-based biomedical concept embeddings on a large scale. *PLoS Comput. Biol.* **2020**, *16*, e1007617. [CrossRef] [PubMed]
68. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C.H.; Kang, J. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **2019**, *36*, 1234–1240. [CrossRef]
69. Rifaioglu, A.S.; Doğan, T.; Martin, M.J.; Cetin-Atalay, R.; Atalay, M.V. Multi-task Deep Neural Networks in Automated Protein Function Prediction. *arXiv* **2017**, arXiv:1705.04802.
70. Grover, A.; Leskovec, J. node2vec: Scalable Feature Learning for Networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining—KDD '16*; Association for Computing Machinery: New York, NY, USA, 2016; pp. 855–864. [CrossRef]
71. Du, J.; Jia, P.; Dai, Y.; Tao, C.; Zhao, Z.; Zhi, D. Gene2vec: Distributed representation of genes based on co-expression. *BMC Genom.* **2019**, *20*, 82. [CrossRef]
72. Jiang, Y.; Clark, W.T.; Friedberg, I.; Radivojac, P. The impact of incomplete knowledge on the evaluation of protein function prediction: A structured-output learning perspective. *Bioinformatics* **2014**, *30*, 609–616. [CrossRef]
73. Hales, K.G.; Korey, C.A.; Larracuente, A.M.; Roberts, D.M. Genetics on the Fly: A Primer on the Drosophila Model System. *Genetics* **2015**, *201*, 815–842. [CrossRef] [PubMed]
74. Kuwabara, P.E.; O'Neil, N. The use of functional genomics in C. elegans for studying human development and disease. *J. Inherit. Metab. Dis.* **2001**, *24*, 127–138. [CrossRef] [PubMed]
75. Schnoes, A.M.; Ream, D.C.; Thorman, A.W.; Babbitt, P.C.; Friedberg, I. Biases in the Experimental Annotations of Protein Function and Their Effect on Our Understanding of Protein Function Space. *PLoS Comput. Biol.* **2013**, *9*, e1003063. [CrossRef] [PubMed]
76. Škunca, N.; Altenhoff, A.; Dessimoz, C. Quality of computationally inferred gene ontology annotations. *PLoS Comput. Biol.* **2012**, *8*, e1002533. [CrossRef]
77. Youngs, N.; Penfold-Brown, D.; Bonneau, R.; Shasha, D. Negative Example Selection for Protein Function Prediction: The NoGO Database. *PLoS Comput. Biol.* **2014**, *10*, e1003644. [CrossRef]
78. Fu, G.; Wang, J.; Yang, B.; Yu, G. NegGOA: Negative GO annotations selection using ontology structure. *Bioinformatics* **2016**, *32*, 2996–3004. [CrossRef]
79. Warwick Vesztrocy, A.; Dessimoz, C. Benchmarking gene ontology function predictions using negative annotations. *Bioinformatics* **2020**, *36*, i210–i218. [CrossRef] [PubMed]

80. Kiryo, R.; Niu, G.; du Plessis, M.C.; Sugiyama, M. Positive-Unlabeled Learning with Non-Negative Risk Estimator. In *Advances in Neural Information Processing Systems 30*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; pp. 1675–1685.
81. Yang, P.; Li, X.L.; Mei, J.P.; Kwoh, C.K.; Ng, S.K. Positive-unlabeled learning for disease gene identification. *Bioinformatics* **2012**, *28*, 2640–2647. [CrossRef] [PubMed]
82. Akbarnejad, A.; Baghshah, M.S. A probabilistic multi-label classifier with missing and noisy labels handling capability. *Pattern Recognit. Lett.* **2017**, *89*, 18–24. [CrossRef]
83. Rao, R.; Bhattacharya, N.; Thomas, N.; Duan, Y.; Chen, P.; Canny, J.; Abbeel, P.; Song, Y. Evaluating Protein Transfer Learning with TAPE. In *Advances in Neural Information Processing Systems 32*; Wallach, H., Larochelle, H., Beygelzimer, A., d' Alché-Buc, F., Fox, E., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2019; pp. 9689–9701.
84. Heinzinger, M.; Elnaggar, A.; Wang, Y.; Dallago, C.; Nechaev, D.; Matthes, F.; Rost, B. Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinform.* **2019**, 723. [CrossRef]
85. Alley, E.C.; Khimulya, G.; Biswas, S.; AlQuraishi, M.; Church, G.M. Unified rational protein engineering with sequence-based deep representation learning. *Nat. Methods* **2019**, *16*, 1315–1322. [CrossRef]
86. Rives, A.; Meier, J.; Sercu, T.; Goyal, S.; Lin, Z.; Guo, D.; Ott, M.; Zitnick, C.L.; Ma, J.; Fergus, R. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *bioRxiv* **2020**, 622803. [CrossRef]
87. Villegas-Morcillo, A.; Makrodimitris, S.; van Ham, R.C.H.J.; Gomez, A.M.; Sanchez, V.; Reinders, M.J.T. Unsupervised protein embeddings outperform hand-crafted sequence and structure features at predicting molecular function. *Bioinformatics* **2020**, btaa701. [CrossRef]
88. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2018**, arXiv:1810.04805.
89. Wei, Q.; Khan, I.K.; Ding, Z.; Yerneni, S.; Kihara, D. NaviGO: Interactive tool for visualization and functional similarity and coherence analysis with gene ontology. *BMC Bioinform.* **2017**, *18*, 177. [CrossRef]
90. Makrodimitris, S.; Van Ham, R.C.H.J.; Reinders, M.J.T. Improving Protein Function Prediction in Ara-bidopsis Using Protein Sequence and GO-term Similarities. *Bioinformatics* **2019**, under review. [CrossRef]
91. Bi, W.; Kwok, J. *Multi-Label Classification on Tree-and DAG-Structured Hierarchies*; ICML: New York, NY, USA, 2011; pp. 17–24.
92. Adadi, A.; Berrada, M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* **2018**, *6*, 52138–52160. [CrossRef]
93. Longo, L.; Goebel, R.; Lecue, F.; Kieseberg, P.; Holzinger, A. Explainable Artificial Intelligence: Concepts, Applications, Research Challenges and Visions. In *Machine Learning and Knowledge Extraction*; Holzinger, A., Kieseberg, P., Tjoa, A.M., Weippl, E., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 1–16.
94. Escalante, H.J.; Escalera, S.; Guyon, I.; Baró, X.; Güçlütürk, Y.; Güçlü, U.; van Gerven, M. (Eds.) *Explainable and Interpretable Models in Computer Vision and Machine Learning*; Springer International Publishing: Cham, Switzerland, 2018. [CrossRef]
95. Smaili, F.Z.; Gao, X.; Hoehndorf, R. Onto2Vec: Joint vector-based representation of biological entities and their ontology-based annotations. *Bioinformatics* **2018**, *34*, i52–i60. [CrossRef] [PubMed]
96. Venkatesan, R.; Er, M.J.; Dave, M.; Pratama, M.; Wu, S. A novel online multi-label classifier for high-speed streaming data applications. *Evol. Syst.* **2017**, *8*, 303–315. [CrossRef]
97. Ahmadi, Z.; Kramer, S. Online Multi-Label Classification: A Label Compression Method. *arXiv* **2018**, arXiv:1804.01491.
98. Kahanda, I.; Funk, C.S.; Ullah, F.; Verspoor, K.M.; Ben-Hur, A. A close look at protein function prediction evaluation protocols. *GigaScience* **2015**, *4*, 41. [CrossRef]
99. Szklarczyk, D.; Gable, A.L.; Lyon, D.; Junge, A.; Wyder, S.; Huerta-Cepas, J.; Simonovic, M.; Doncheva, N.T.; Morris, J.H.; Bork, P.; et al. STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **2019**, *47*, D607–D613. [CrossRef]

100. Clark, W.T.; Radivojac, P. Information-theoretic evaluation of predicted ontological annotations. *Bioinformatics* **2013**, *29*, i53–i61. [CrossRef]
101. Plyusnin, I.; Holm, L.; Törönen, P. Novel comparison of evaluation metrics for gene ontology classifiers reveals drastic performance differences. *PLoS Comput. Biol.* **2019**, *15*, e1007419. [CrossRef]

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).