bioRxiv preprint doi: https://doi.org/10.1101/2022.02.07.479382; this version posted February 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

CONSTRUCTION OF *IN SILICO* PROTEIN-PROTEIN INTERACTION NETWORKS ACROSS DIFFERENT TOPOLOGIES USING MACHINE LEARNING

Loïc Lannelongue[1,2,3,*], Michael Inouye[1,2,3,4,5,6,*]

[1]Cambridge Baker Systems Genomics Initiative, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK

[2]British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK

[3]Health Data Research UK Cambridge, Wellcome Genome Campus and University of Cambridge, Cambridge, UK

[4]Cambridge Baker Systems Genomics Initiative, Baker Heart and Diabetes Institute, Melbourne, Victoria, Australia

[5]British Heart Foundation Centre of Research Excellence, University of Cambridge, Cambridge, UK

[6]The Alan Turing Institute, London, UK

* Correspondence: LL (LL582@medschl.cam.ac.uk) and MI (mi336@medschl.cam.ac.uk; minouye@baker.edu.au)

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.07.479382; this version posted February 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

ABSTRACT

Protein-protein interactions (PPIs) are essential to understanding biological pathways as well as their roles in development and disease. Computational tools have been successful at predicting PPIs *in silico*, but the lack of consistent and reliable frameworks for this task has led to network models that are difficult to compare and, overall, a low level of trust in the PPI predictions. To better understand the underlying mechanisms that underpin these models, we designed B4PPI, an open-source framework for benchmarking that accounts for a range of biological and statistical pitfalls while facilitating reproducibility. We use B4PPI to shed light on the impact of network topology and how different algorithms deal with highly connected proteins. By studying functional genomics-based and sequence-based models (the two most popular approaches) on human PPIs, we show their complementarity as the former performs best on lone proteins while the latter specialises in interactions involving hubs. We also show that algorithm design has little impact on performance with functional genomic data. We replicate our results between both human and *S. Cerevisiae* data and demonstrate that models using functional genomics are better suited to PPI prediction across species. With rapidly increasing amounts of sequence and functional genomics data, our study provides a systematic foundation for future construction, comparison and application of PPI networks.

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.07.479382; this version posted February 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

INTRODUCTION

Protein-protein interactions (PPIs) are central to protein function and inform a wide range of biomedical applications, from mechanistic studies [1], [2] to drug development [3], [4]. Better understanding these interactions is critical for successfully mapping biological pathways, but the diversity of PPIs and the scale of the network make this a difficult task. Experimental methods to map PPIs exist, but even when high-throughput tend to focus on proteins of interest.

Computational methods can address the issue of scalability and experimental bias. Given a pair of proteins and some characteristics of each one, machine learning models can learn to predict the likelihood of interaction. Numerous methods have been developed for this, using the full range of machine learning models, from early work on *S. Cerevisiae* [5]–[8] to algorithms dedicated to human PPIs [9]–[13]. Yet, despite a wealth of tools, the mechanics and consequences of the underlying inference are still poorly understood, and it is unclear why models with similar performance make vastly different predictions. Reported performance scores often cannot be compared or replicated due to proprietary data and inconsistent or flawed assessment methods. As a consequence, there are multiple issues for *in silico* PPIs: it is unclear what the state-of-the-art is, analyses are difficult to reconcile, the development of new models is inefficient, follow-up mechanisms studies are likely undermined and, ultimately, there are different versions of the underlying molecular networks that describe protein function.

A unified framework for PPI inference would improve the development and reliable assessment of new models, and would facilitate the overdue widespread adoption of PPI predictions for downstream analysis. Replicable, trustworthy and generalisable high-performing models can capture more causal biology and enhance many aspects of biological research such as experimental designs and drug development.

In this work, we design a robust and standardised approach to *in silico* PPI prediction that accounts for both biological and statistical pitfalls and leverages the strength of large, open-source and professionally curated databases. We make publicly available benchmarking standards for human and yeast PPIs to accelerate future discoveries and lay the foundations for similar datasets for other organisms. Within this framework, we study and compare the main approaches to PPI prediction in humans, based on functional genomic (FG) information or amino acid sequences alone. We highlight why both perspectives are still relevant today and how each adapts to the PPI network's topology. In particular, we show that the presence of highly connected proteins in the networks has a drastic impact on prediction models and is an area where FG and sequence models diverge. We also replicate these results between human and yeast (*S. Cerevisiae*) and show which tools are most suitable to cross-species predictions. This work provides robust foundations for future developments in PPI prediction models, but also gives critical insight into which models can and should be used in different situations.

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.07.479382; this version posted February 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

RESULTS

B4PPI: A robust and open-source benchmark for PPI prediction

The lack of a consistent way to assess PPI prediction algorithms has hindered the development of such algorithms and reduced their impact by making it difficult to reuse models for downstream analysis [14]. Benchmarks are important for replicability, and when combined with carefully curated datasets, they enable fast development through trial and error. Our Benchmarking Pipeline for the Prediction of Protein-Protein Interactions in Humans (B4PPI-Human) includes both carefully selected training and testing sets and a collection of input features to enable such trials. Standard UniProt IDs are used throughout to easily combine these with other data sources. Relevant metrics are selected with guidelines on how to share them. All this, alongside the pre-processing steps and relevant guidelines, is made available online and can be downloaded easily from https://github.com/Llannelongue/B4PPI. An example of a reporting sheet is in **Figure 1**.

The complexity of the underlying biological mechanisms of PPIs introduces pitfalls that need to be considered when evaluating models. First, the way non-interacting proteins are selected for training is important. While some efforts have used proteins known to be localised in different parts of the cell [5], [12], [15], [16], this has been shown to be unreliable and a source of significant bias that overestimates accuracy [17]. An alternative is to use a database of experimentally tested non-interacting proteins, but leading resources such as Negatome have only ~1,300 pairs and thus offer limited coverage [15], [18]. Considering the scarcity of PPIs, randomly sampling pairs of proteins has a very low risk of false negative and limits selection bias (i.e. focusing on known proteins of interest) [17], [19]. However, the impact of the associated imbalance between interacting and non-interacting proteins should be taken into account when training models on balanced datasets [20]. Lastly, each observation is itself a pair of proteins. Even when ensuring that the two sets don't have pairs in common, there can be individual proteins present in both the training and testing sets. This protein-level overlap, often overseen, has been shown to significantly affect the performance of an algorithm and should therefore be properly assessed [21]. Despite being documented in the literature, these pitfalls are still unevenly accounted for in published works. This, alongside inconsistencies in the choice of testing sets and performance metrics explains why, despite the number of algorithms released in recent years, there is still no simple way to compare a new approach to the state-of-the-art, or even know what the state-of-the-art is.

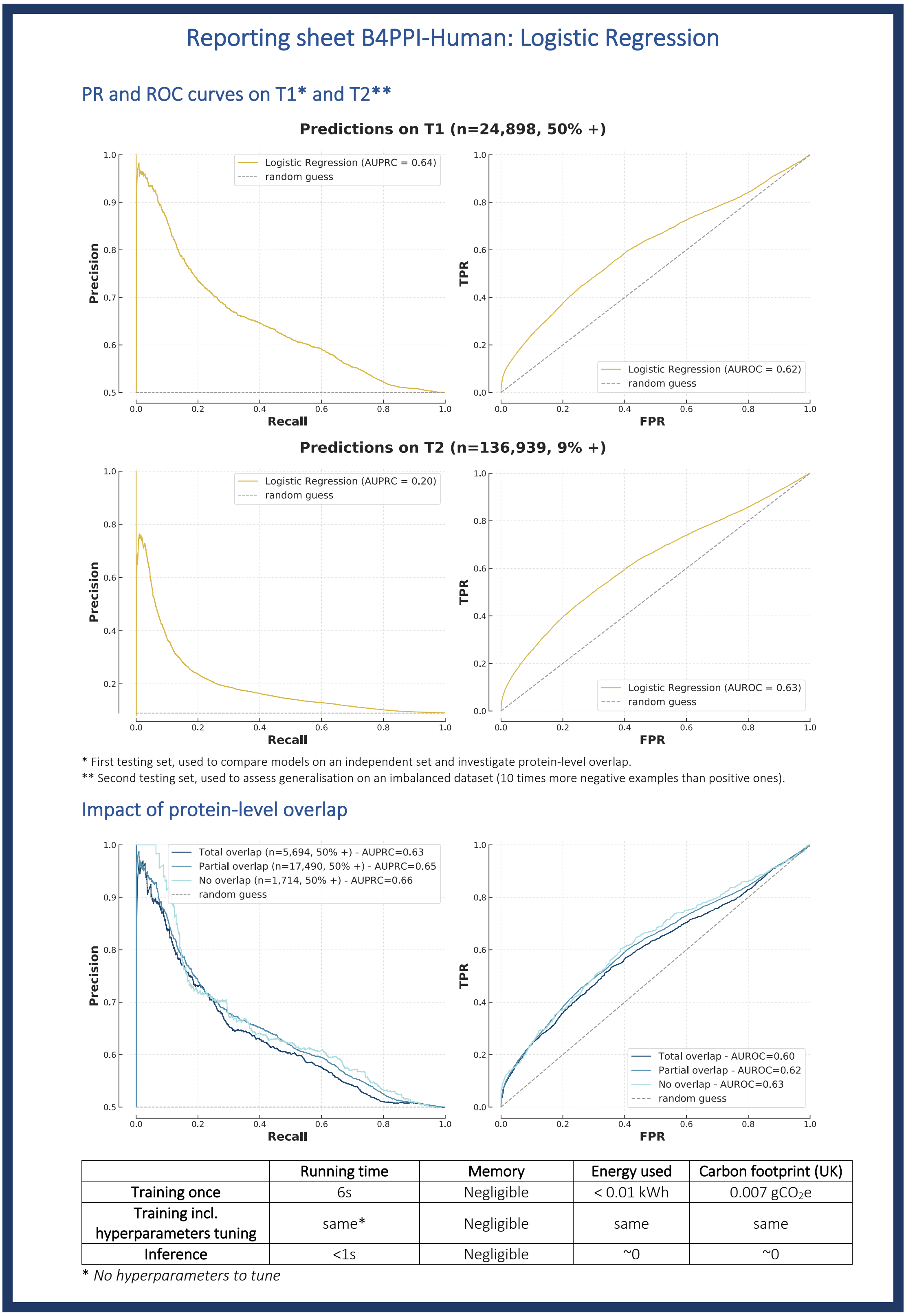
The essential aspects of training and assessment that should be systematically accounted for are (1) the quality of the positive examples (i.e. the interacting proteins), (2) how non-interacting proteins are selected for the gold standard, (3) a suitable split between training and testing sets, in particular regarding individual proteins, and (4) the metrics to evaluate and compare models. B4PPI seeks to address these four aspects of benchmarking.

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.07.479382; this version posted February 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

When building a gold standard for machine learning algorithms, quality and representativity are the most important aspects to consider, which makes IntAct [22] a database of choice for interacting proteins. It aggregates reliable evidence of molecular interactions from over 20,000 publications, which are manually curated, and includes data from other interactions databases such as the IMEx consortium [23]. We further limited the risk of false positives by removing low-quality interactions, for example the ones based on spatial colocalisation only (**Methods**). The final dataset comprised 78,229 interactions, covering 12,026 proteins (out of the 20,386 registered in UniProt).

To select non-interacting proteins to serve as negative examples, randomly sampling protein pairs is the approach with the lowest probability of error considering the scarcity of the PPI network [24]. Non-interacting proteins can be sampled using a uniform distribution, i.e. all proteins have an equal probability of being selected, which leads to an unbiased set, representative of the general population of protein pairs. However, PPI networks are known to be similar to scale-free networks, i.e. composed of a few highly connected nodes, called *hubs*, and numerous *lone* proteins with few interactions [25] (**Supplementary Figure 1**). Consequently, hubs are over-represented in a set of PPIs. For example in our curated set from IntAct, the top 20% of proteins by number of interactions were involved in 94% of PPIs. But when uniformly sampling protein pairs, the same top 20% were only involved in 37% of non-interacting proteins. Although expected, this can be an issue for machine learning algorithms that would identify hubs and systematically predict a positive interaction when hubs are involved. Such a strategy would maximise accuracy on the training set but lead to a majority of false positives when making predictions on new pairs. To mitigate this, a balanced sampling can be used [26], where the probability of sampling a protein for the negative set is proportionate to its frequency in the positive set. It has been shown that each strategy serves a different purpose [19]; balanced sampling is beneficial for training models but shouldn't be used for evaluating them, as the induced bias makes metrics less meaningful. This was the strategy implemented here, where non-interacting proteins were selected with balanced sampling for the training set and uniform sampling for the testing sets.

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.07.479382; this version posted February 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



Reporting sheet B4PPI-Human: Logistic Regression

PR and ROC curves on T1* and T2**

Predictions on T1 (n=24,898, 50% +)

Predictions on T2 (n=136,939, 9% +)

* First testing set, used to compare models on an independent set and investigate protein-level overlap.

** Second testing set, used to assess generalisation on an imbalanced dataset (10 times more negative examples than positive ones).

Impact of protein-level overlap

Precision
Recall
Total overlap (n=5,694, 50% +) - AUPRC=0.63
Partial overlap (n=17,490, 50% +) - AUPRC=0.65
No overlap (n=1,714, 50% +) - AUPRC=0.66
random guess
TPR
FPR
Total overlap - AUROC=0.60
Partial overlap - AUROC=0.62
No overlap - AUROC=0.63
random guess
0.0
0.2
0.4
0.6
0.8
1.0
0.5
0.6
0.7
0.8
0.9

	Running time	Memory	Energy used	Carbon footprint (UK)
Training once	6s	Negligible	< 0.01 kWh	0.007 gCO$_2$e
Training incl. hyperparameters tuning	same*	Negligible	same	same
Inference	<1s	Negligible	~0	~0

** No hyperparameters to tune*

Figure 1: Reported performance sheet of the logistic regression (FG-based) on B4PPI-Human.

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.07.479382; this version posted February 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

In the presence of limited data, the division of the gold standard between training and testing sets is critical to simultaneously optimise learning and obtain meaningful generalisation metrics. Here, the testing set should achieve several objectives, (1) provide performance metrics on a new, independent set, (2) measure the impact, or absence of impact, of protein-level overlap, (3) demonstrate how the model can generalise to real-world data. Since a single set cannot achieve simultaneously (2) and (3), as the careful selection of proteins to measure overlap biases the dataset, we designed two testing sets *T1* and *T2* (**Methods**). *T1* should be used to compare different approaches with an independent set and investigate protein-level overlap, and *T2* should be used to assess generalisation. *T1* was built by purposefully leaving some proteins out of the training set; we demonstrated the importance of this as dividing the training and testing sets conventionally (using, for example, the popular *scikit-learn* library) resulted in almost all pairs (95%) in the testing set sharing at least one protein with the training set (**Figure 2**), which may lead to overestimating performances [21]. *T2*, with ten times more negative examples than positive ones (**Supplementary Table 1**), can then be used to assess how models perform in a more realistic setting where positive interactions are rare compared to non-interacting proteins.

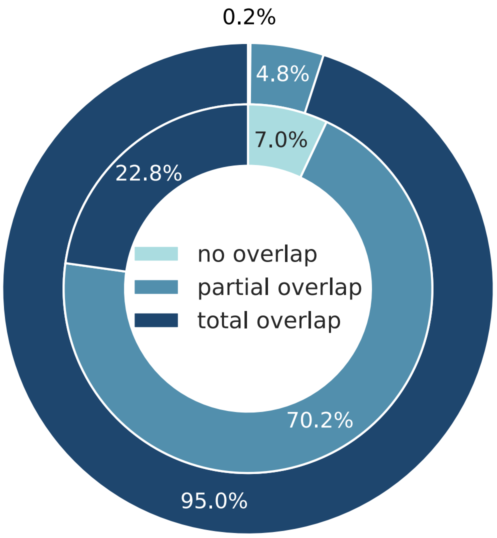


Figure 2: The impact of train/test splitting strategies on protein-level overlap. The common splitting strategy is to allocate pairs randomly (outer ring) while here we set aside proteins for testing (inner ring).

The choice of metrics is a crucial element of a benchmark, and summarising the results by a single number, such as accuracy or AUROC, is often misleading [27]. We report both the Receiver Operating Characteristic (ROC) and the Precision-Recall (PR) curves which highlight nuanced and complementary aspects of PPI models. In addition, to address the environmental impact of bioinformatic tools [28], we also reported the carbon footprint of training models, measured using the Green Algorithms calculator [29].

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.07.479382; this version posted February 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The elements described above represent the minimum needed for reproducible benchmarks and researchers who wish to use their own input features can evaluate their models on these partitions. However, to rapidly test a new model, it is useful to have access to carefully selected and highly accurate protein properties. The two main categories of features used are amino acid sequences and functional genomics annotations, such as subcellular localisation and biological functions. These are available with B4PPI, from the professionally curated databases UniProt, the Human Protein Atlas (HPA) [30], [31] and Bgee [32] (**Methods** and **Table 1** for the full list of features).

With B4PPI, we could compare different models in a consistent manner to better understand what aspects of the underlying biology are captured by each method. We focused here on FG-based and sequence-based models as they have been widely used and rarely compared, despite attempts at combining them.

FG-based linear models achieved top performance

In FG-based models, FG annotations are pre-processed to compute similarity measures, such as colocalisation, between proteins (**Methods**). The low dimensionality of the transformed problem explains the success of standard machine learning algorithms; in particular, Naïve Bayes Classifiers [33], decision trees [34] and Random Forests [35] have been the most popular choices [5]–[7], [36]. Despite the proven track record of such tools, the more recent XGBoost algorithm [37] has been shown to outperform them in other situations like kidney disease diagnostic [38], which motivated its inclusion in this analysis.

Using logistic regression as a baseline, we reported PR and ROC curves on the two test sets (**Figure 1**). A list of coordinates for these two curves was made available so that future models can be compared without unnecessary re-training. We also reported the training time, 6 seconds, and the carbon footprint, close to 0 gCO_2e. We then compared other models to this baseline and produced similar performance sheets (**Supplementary Figure 2**).

We found that more complex algorithms brought little improvement over logistic regression, as most models performed similarly on *T1* (**Figure 3** and **Supplementary Figure 3**). XGBoost and Random Forest showed minor improvement in AUROC and AUPRC, but the difference between the ROC curves of the logistic regression and XGBoost is non-significant (p=0.27) (**Methods**). Moreover, XGBoost was more efficient than Random Forest as it had nearly half the runtime (30s vs 54s). When studying the coefficients of the linear regression, we found that most decisions are based on common biological processes, co-localisation (cellular compartment) and common domains, all three coefficients being significant (p<0.001) (**Supplementary Figure 4**).

The reporting standard also enabled us to look at finer performance metrics, broken down by protein-level overlap (i.e. individual proteins common to the training and testing sets). Comparing PR and ROC curves showed that both logistic regression and XGBoost were

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.07.479382; this version posted February 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

unaffected by the level of overlap (**Figure 1** and **Supplementary Figure 2**), and can therefore transfer effectively to new proteins.

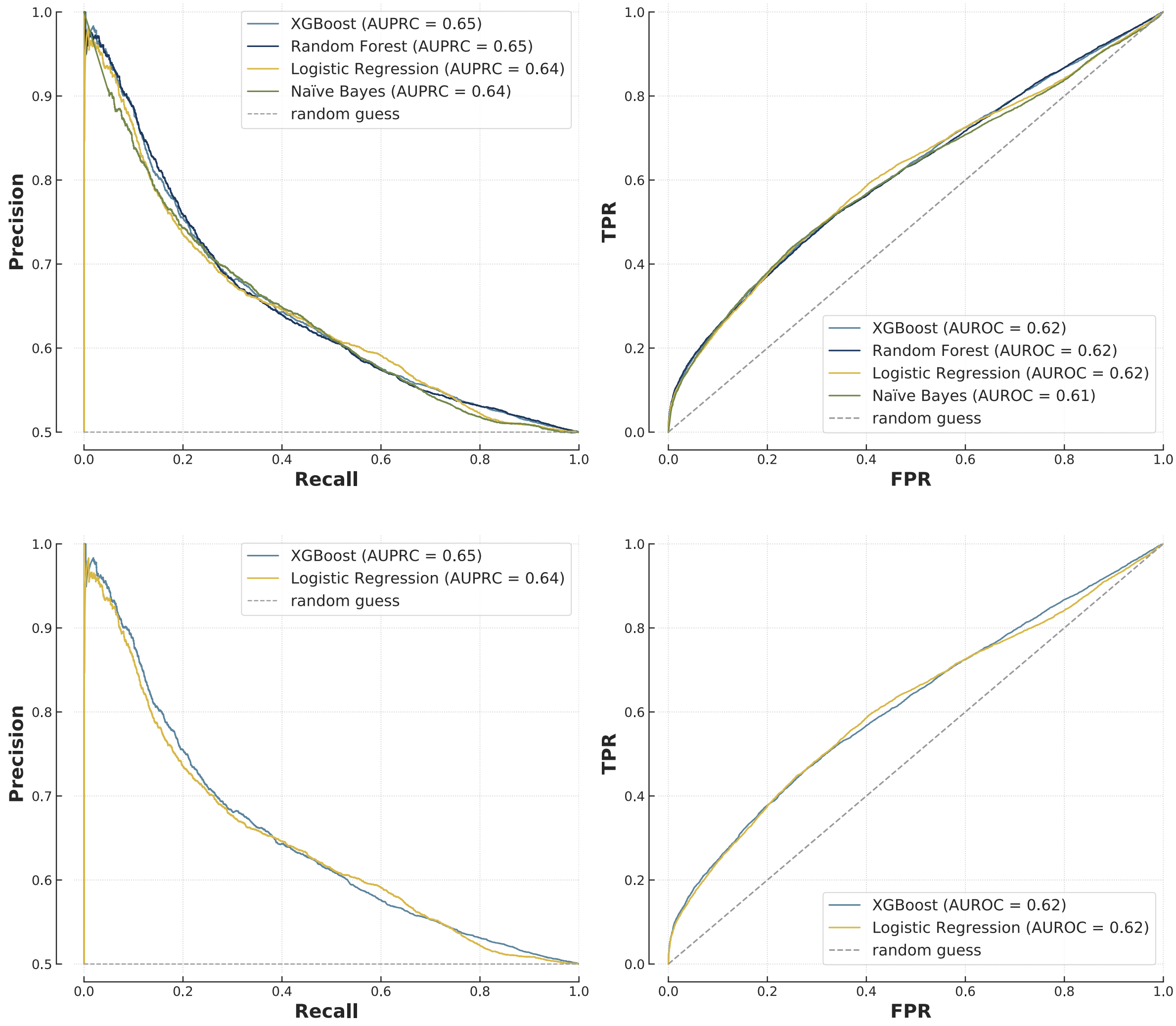


Figure 3: Comparison of FG-based models on *T1* (n=24,898, 50% positive), with PR curves (left) and ROC curves (right) for all the models tested (top) and then only XGBoost and Logistic Regression for clarity (bottom).

Sequence models outperformed FG-based algorithms on known proteins

The alternative to FG-based models is to use amino acids sequences as the input for a PPI prediction algorithm. We compared several deep learning architectures and reported the performance of an optimised Siamese neural network (**Methods**, **Figure 9** and **Supplementary Figure 5**). Despite having access to no functional information about the proteins, the sequence model outperformed the best performing FG-based model, XGBoost, except at low recall and high precision (AUPRC=0.68 vs 0.65 and ROC curves significantly different, $p=9x10^{-47}$) (**Figure 4**). However, while XGBoost was trained in only 30s with less than 0.01 kWh of energy, the deep learning approach trained for 1h10 with 0.62 kWh, emitting 22,000 times more greenhouse gases (GHGs). In addition, the performance of the sequence model was heavily

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.07.479382; this version posted February 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

affected by the choice of deep learning architecture and its hyper-parameters, such as number of layers or learning rate. These require extensive (and expensive) optimisation. Protein-level overlap had a significant impact on these results. The model had an AUPRC of 0.68 on average, but 0.75 when restricted to proteins present in both training and testing sets, and only 0.62 when there was no overlap (**Supplementary Figure 5**). This demonstrates that (1) in the absence of specific adjustments, such deep learning models are poorly suited to make predictions on previously unseen proteins and (2) in-depth benchmarks like B4PPI are important to reliably measure performances. While this comparison of FG-based and sequence-based models could indicate that deep learning is the best approach to PPI prediction, and support the numerous similar claims in the literature, it could also be the consequence of unaccounted-for biological properties of PPIs.

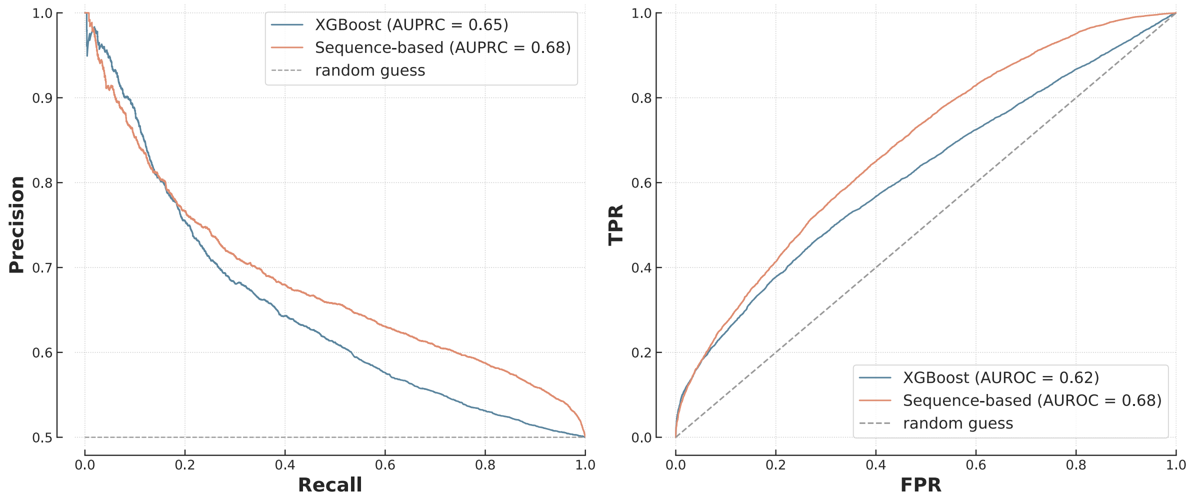


Figure 4: Siamese network vs XGBoost on *T1*. The difference between the ROC curves was statistically significant (p = $9x10^{-47}$).

The role of network hubs is essential to PPI prediction

A scale-free topology has important biological implications [25] so we hypothesised that a one-fits-all approach for hubs and lone proteins is unlikely to be optimal. In assessing interactions between protein hubs (hub-hub), between a protein hub and a lone protein (hub-lone) and between lone proteins (lone-lone) (**Methods**), we found a distinct pattern whereby FG-based models had greater AUPRC and AUROC for interactions involving only lone proteins while sequence-based models performed better for hubs (**Figure 5**).

These findings can be explained by the pre-processing of similarity measures for FG models. Because of their central role in biological pathways, hubs are highly studied and therefore annotated for many processes and localisations. For example in the training set, hubs have on average 11.6 annotations for biological processes (significant feature in the logistic regression model discussed above) while non-hubs only have 5.8 (median 6 vs 3). The same phenomenon

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.07.479382; this version posted February 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

is observed for cellular compartments (6.2 vs 3.6 annotations on average). Because the similarity measures used by the FG models quantify overlaps in annotations, hubs annotated for a large number of processes provide little information about the probability of interaction, which can explain why FG-based models perform best when hubs are not involved.

These results provide insight into the strengths of each approach and, importantly, show that a PPI approach should be context specific, particularly with respect to the network topologies of interest. Indeed, the apparent superiority of the deep learning model shown on **Figure 4** is largely due to the composition of *T1*, made up of 70% of hub-hub or hub-lone interactions.

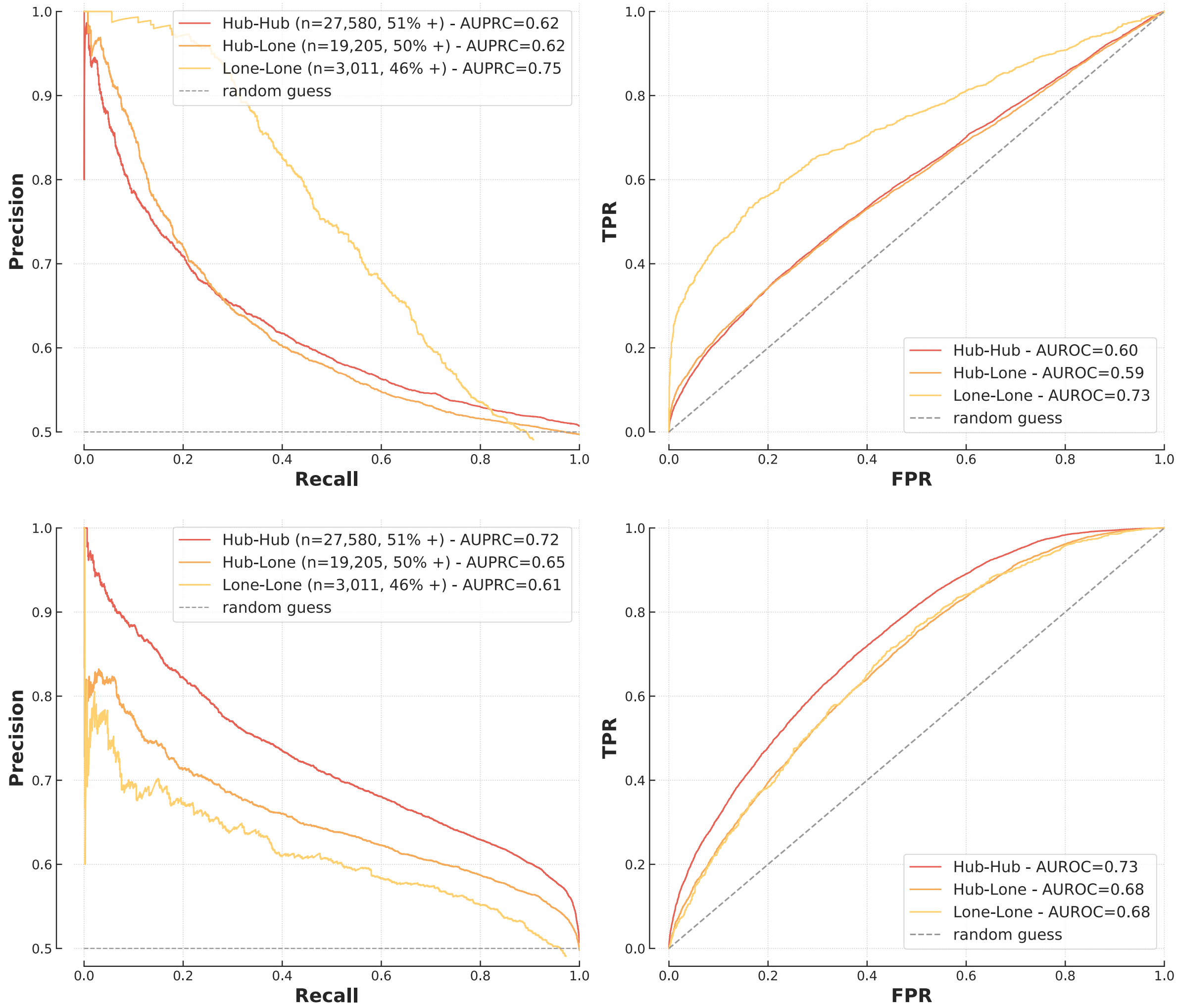


Figure 5: Performance of XGBoost (top) and sequence-based model (bottom) on hubs and lone proteins.

Cross-species validation of PPI prediction models and relative performances

S. Cerevisiae is a well-studied model organism with a known interactome and has been used extensively for *in silico* PPI predictions [8], [36], [39]. We replicated the analyses presented above on *S. Cerevisiae* proteins and found that our findings regarding network topology and

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.07.479382; this version posted February 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

models' relative performances were robust across species. The data was selected similarly to previously, extracted and curated from IntAct and UniProt, but without data from HPA and Bgee as these databases do not curate yeast (**Methods**).

As shown previously, all FG-based models had similar performances with AUPRC between 0.71 and 0.73 (**Figure 6**); however, in this analysis, the differences between XGBoost and other models were statistically significant (p = 2x10^{-12} for Naïve Bayes and p = 9x10^{-31} for logistic regression). The sequence model outperformed FG-based models in most cases (p = 2x10^{-6}), except at high recall (**Figure 6**). Second, similar to humans, FG-based models were not sensitive to protein-level overlap while sequence-based models had different performances depending on the level of overlap (**Supplementary Figure 6**). Finally, we found consistency regarding the role of network hubs; FG-based models were better able to predict lone-lone protein interactions while the sequence-based model was better at predicting interactions amongst protein hubs (**Supplementary Figure 7**).

While experimental data on PPIs is readily available for humans and *S. Cerevisiae*, many non-model organisms lack data despite their biological relevance [40]. For these, cross-species predictions – i.e. training a model on a species to make predictions on another – are of particular interest. We showed that FG-based models are generally more suitable than sequence-based ones for this task.

We investigated whether models trained on yeast could be used to predict human PPIs, finding that the yeast-trained FG-based models (logistic regression and XGBoost) achieved similar AUPRC and AUROC as those which were human-trained to predict human PPIs (p = 0.26 for XGBoost) (**Figure 7**). Conversely, the yeast-trained sequence model was unable to predict human PPIs (AUPRC = 0.52 vs 0.68, p = 3x10^{-272}). We observed the same phenomenon when using human-trained models to make predictions on yeast (**Supplementary Figure 8**).

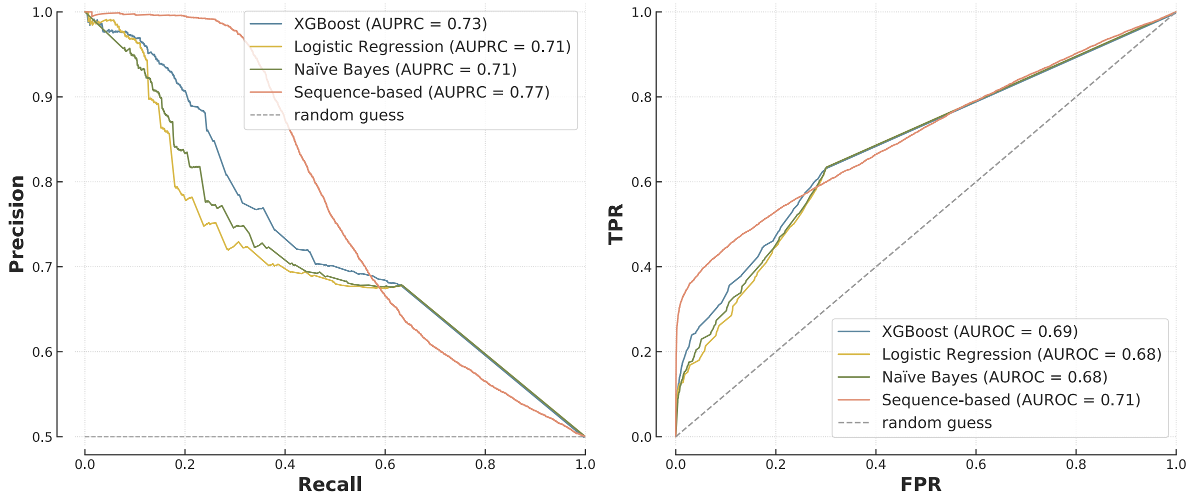


Figure 6: Comparison of a range of models on the yeast testing set.

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.07.479382; this version posted February 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

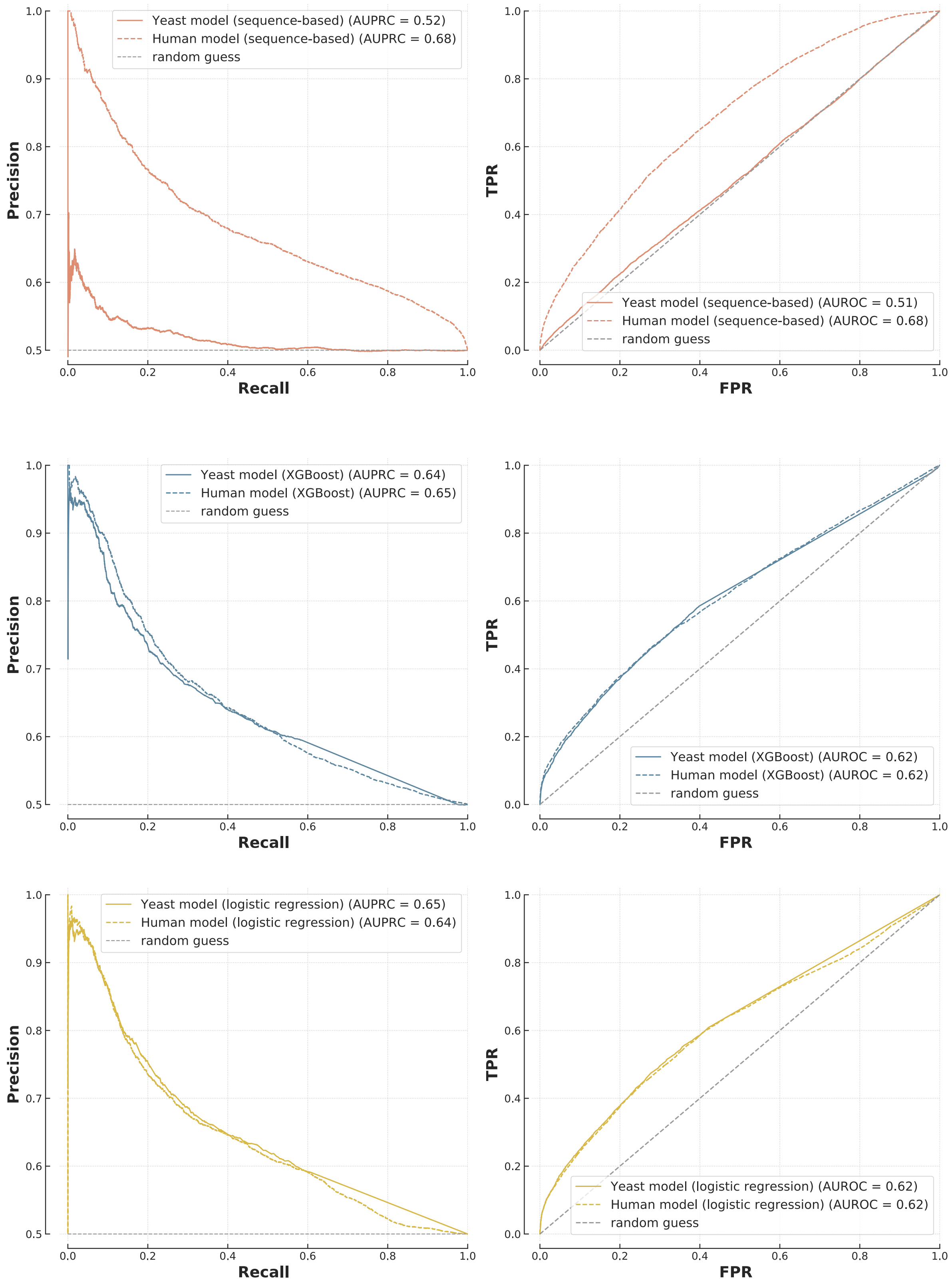


Figure 7: Cross-species predictions. Models trained on human PPIs (dotted lines) and yeast PPIs (solid lines) were used to make predictions on the human testing set. The top plot is the sequence-based model, the other ones are FG-based (XGBoost in the middle, logistic regression at the bottom).

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.07.479382; this version posted February 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Discussion

In this work, we sought to identify and explain the strengths and weaknesses of a wide selection of approaches to PPI prediction, and thereby provide the community with both a benchmarking resource, insight into which PPI approach to select or trust in a particular scenario, and finally with a set of well-studied PPI models which have also been validated across species.

The FG-based and sequence-based models are common for PPI prediction but are rarely directly compared. In particular, it was unclear where the differences lie and if one approach should be preferred today. We found that when using FG annotations, the choice of algorithm has little impact on the predictions and a logistic regression performs close to the state-of-the-art while providing clear insight into the decision-making process; here, colocalisation, common biological processes and shared domains are the main indicators of interaction. The fact that a highly flexible and non-linear model such as XGBoost performs similarly to logistic regression, making identical predictions in 93% of cases, shows that performance is likely driven by the quality and the pre-processing of the FG annotations instead of the modelling; once the similarity measures have been calculated, there are limited non-linearities and a simple logistic regression achieves top performance. Sequence-based models on the other hand need specifically optimised architectures but achieve similarly high performances, if not higher in some settings, without any biological information apart from amino acid sequences.

We found that the two approaches adapt to the presence of hubs and lone proteins differently and show complementary strengths. While sequence-based models are mostly useful when hubs are involved, FG-based models perform well for interactions between lone proteins. This simple result offers important insight into the specificities of each approach and explains discrepancies in reported performances in the literature, as the topology of the testing set has a large impact on metrics. These results are not specific to human PPIs as the same conclusions were drawn from analysis on *S. Cerevisiae*. Cross-species predictions are instrumental to study non-model organisms, and we showed that FG-based decision rules translate well to new species while sequence-based models do not.

These observations are consistent with the way each algorithm learns. FG-based models make predictions based on general, but less complex, rules about PPIs which translate well to new proteins and new species. This is particularly useful considering that many proteins are still not represented in interaction databases. On the other hand, sequence-based models have millions of parameters which give them the flexibility to recognise individual proteins and learn specific interaction patterns. Although this enables such models to make predictions without functional information, it also limits high performance to proteins present in the training set. This likely explains the poor results of sequence models on previously unseen proteins and cross-species datasets. It is also consistent with the high performance of these models on network hubs, which are overrepresented in training sets and therefore well captured by the models.

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.07.479382; this version posted February 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

These analysis and results required a robust and reliable benchmarking pipeline. We designed the open-source B4PPI, which accounts for a range of biological and statistical pitfalls. By being freely accessible and using standard identifiers for proteins, B4PPI can be used by any researcher working on *in silico* PPI prediction to assess performances and compare their approaches to the state-of-the-art. An example reporting sheet is presented that includes relevant metrics, from PR and ROC curves to runtime and carbon footprint, to ensure the models released can be trusted and encourage wider use of PPI imputation for downstream analysis. B4PPI also comes with pre-processed features to enable rapid development of new approaches.

Our study has limitations. We focused on the two most widely used approaches to PPI prediction, namely FG-based and sequence-based; however, some alternative approaches have also been proposed, using, for example, higher-level protein structures [36], phylogeny [41] and the topology of existing networks [42]–[44], but the latter depends heavily on the quality of the existing PPI networks. Most FG annotations are from gene ontologies which have a hierarchical structure which we do not account for here, contrary to Armean et al. [45] for example. Moreover, we analysed two common interactomes, human and yeast, yet there are many more. As demonstrated though with the yeast dataset, similar benchmarks and analysis can be transferred to other model organisms in a relatively straightforward manner.

We showed here the limits of classic sequence-based deep learning models for cross-species predictions, but it is worth noting some recent deep learning models that have been successfully used for cross-species predictions [46], [47] by including biological and chemical information about amino acids as well as structural knowledge. The results presented in this work can hopefully guide similar future work and help move this area further.

While a benchmarking standard for PPI prediction is needed, it is important to remember the downsides of benchmarks, as demonstrated in computer vision or natural language processing. A fixed set of metrics can motivate the community to overly focus on those, at the expense of applicability and usefulness. To limit this, B4PPI includes a range of metrics but the relevant indicators for each use-case should nonetheless be carefully considered.

The size and complexity of the PPI network makes *in silico* prediction tools indispensable, but it is important to ensure that the models developed are reliable and readily available to the community for downstream analysis and to give insights into biological pathways. For this, consistent and reliable evaluation pipelines are necessary as well as a better understanding of what machine learning models learn. The results presented here make key progress in both areas and facilitate the development, evaluation and reliability of future PPI models.

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.07.479382; this version posted February 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

METHODS

B4PPI-Human

The data was obtained from large and professionally curated databases. This limits experimental bias, as each interaction is based on several experiments, and leverages experts' knowledge in the curation process. Standard UniProt IDs are used throughout to ensure maximum compatibilities. Most of the manipulations were done in Python [48] with Jupyter Notebooks [49] using the Pandas library [50], [51] and Numpy [52]. The plots were drawn using Matplotlib [53], Seaborn [54] and the MetBrewer colour palettes [55]. All the code and final data are available on GitHub (https://github.com/Llannelongue/B4PPI); some intermediary pre-processed datasets are not available online due to file size limits but they can be recreated using the code available. Data is available under Creative Commons Attribution (CC BY 4.0) License.

Protein-protein interaction data

The train machine learning algorithms, the quality of the gold standard is paramount. Data on PPIs was obtained from IntAct [22] and downloaded from the EMBLE-EBI FTP server (timestamp: 15/10/2021). We restricted the data to human protein-protein interactions with UniProt IDs. To reduce the risk of false positives, we removed spoke complex expansions (where the pairwise interactions within a complex are unreliable) and interactions based on colocalisation only. This quality control step leaves 128,790 PPIs, covering 15,506 proteins (out of 20,386 in UniProt). Based on this dataset, we created an index of the number of recorded interactions per protein and made a list of hubs (highly connected proteins). In line with the literature, hubs are defined as the 20% of proteins with the most interactions [56], which here is equivalent to proteins with more than 21 partners. The quality of the interactions is assessed further by looking at the MIscore [57], a quality score based on the manual curation of the interactions that takes into account the detection method, the interaction type and the number of publications reporting it. In case of duplicated PPIs, the highest MIscore was used. When looking at the distribution of the MIscores in the dataset (**Figure 8**), a natural threshold of 0.47 is visible, which restrict the dataset to 78,229 interactions, covering 12,026 proteins.

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.07.479382; this version posted February 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

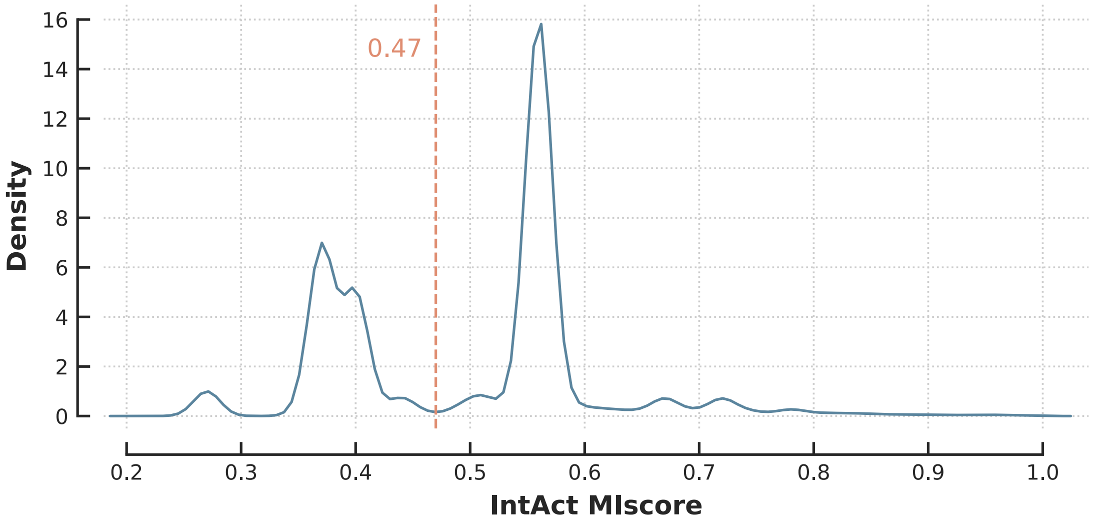


Figure 8: Distribution of the MIscore in IntAct.

Functional genomics annotations and amino acids sequences

Protein sequences in humans are well documented and can be obtained from UniProt, but FG features can be more challenging as they should be diverse (i.e. cover a wide range of properties), of high-quality and have high coverage (i.e. few missing proteins). For the same reasons as described above, aggregated, manually curated and professionally reviewed databases are preferred. Based on features that have been successfully used for the task before, it is relevant to include information about cellular and tissue localisation, biological functions and gene expression patterns [5], [8], [10], [36].

One of the main databases on proteins is UniProt [58] and in particular its knowledgebase UniProtKB. Swiss-Prot, the section of UniProtKB that is reviewed and manually curated, is used in this work to ensure optimal quality. The data from Swiss-Prot is downloaded through their API by restricting to reviewed, non-obsolete, human proteins (last download is 09/11/2021). The different columns are then cleaned to extract the information of interest in a standardised format, and we use UniProt IDs throughout. There is information for 20,386 proteins and more details about each feature are in **Table 1**. UniProt's API is also used to map UniProt IDs between different databases and to map outdated IDs. In particular, we extract amino acids sequences for each protein, more than 95% of which come from the translation of coding sequences submitted to the International Nucleotide Sequence Database Collaboration [59]. Annotated domains and motifs are also included in the database. Additionally, we extract gene ontology (GO) annotations of biological processes, cellular components and molecular functions. For each protein, each of the FG features is represented as a bag-of-words, i.e. a sparse vector of length the number of annotations in the database.

When working with gene expression data, both biological and technical noise need to be accounted for correctly. The Bgee public repository [32] does that by regrouping curated healthy wild-type standardised gene expression patterns. The human data is mainly from GTEx

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.07.479382; this version posted February 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

v6 (phs000424.v6.p1), with an added layer of manual curation to remove unhealthy subjects. For a gene, the final data provides binary calls of presence or absence of expression for each combination of anatomical entity and developmental stage. We downloaded the database from their FTP server (version 14.2) and obtained information for 59,777 genes, 320 anatomical entities and 33 developmental stages, which leads to 1,147 stage/entity combinations. The Bgee entries are matched to the UniProt IDs using UniProt's own mapping table.

The Human Protein Atlas (HPA) [30], [31] provides data mapping human proteins to tissues and cells. In particular, we used the Tissue Atlas [30] that presents the distribution of proteins in tissues and cell types and the Cell Atlas [31] that contains the distribution across subcellular locations. The Tissue Atlas contains data similar to Bgee, but the overlap is likely to be limited as the two databases only share GTEx RNA-seq data. While Bgee has a more thorough curation process, HPA contains a lot of original in-house experimental results, which justifies the inclusion of both data sources. We downloaded the HPA data from their website (release 20.1, Ensembl version 92.38). Despite its name, the data in HPA is identified by Ensembl gene IDs, which are mapped to UniProt IDs using UniProt's API. We restricted the dataset to the reviewed proteins present in Swiss-Prot and to ensure the quality of annotations, we discarded the entries HPA annotated as "uncertain". For the tissue IHC data, we mapped expression levels to numerical values (high=3, medium=2, low=1 and "not detected"=-1) with untested tissues being mapped to 0. Similar pre-processing was used for the consensus RNA-seq data and the subcellular location.

Table 1: Features used to train human models (GO = Gene Ontology).

Feature	Number of different annotations	Missing values (/20,386)	Source
Biological processes (GO)	12,248	3,338	UniProt [58]
Cellular components (GO)	1,754	1,765	UniProt
Molecular functions (GO)	4,346	4,552	UniProt
Domains	2,313	11,815	UniProt
Motifs	819	18,103	UniProt
Sequence	N/A	0	UniProt
Gene expression profile	1,147	1,296	Bgee [32]
Tissue IHC data	62	9,536	HPA [30], [31]
Tissue and cell type	189	9,536	HPA
RNA-seq	61	1,448	HPA
Subcellular location	33	7,820	HPA

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.07.479382; this version posted February 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Pre-processing to measure features similarity

For a protein, each FG feature was represented as a vector, of length the number of annotations. To measure the feature-specific similarity between two proteins, we compared the two vectors using cosine similarity [60], a popular tool widely used for similar tasks in Natural Language Processing. For two vectors $A = (A_i)$ and $B = (B_i)$, their cosine similarity $CS(A, B)$ is:

$$CS(A,B) = \frac{A \cdot B}{\|A\| \, \|B\|} = \frac{\sum_i A_i B_i}{\sqrt{\sum_i A_i^2} \sqrt{\sum_i B_i^2}}$$

As a result, for each of the 207,784,305 possible pairs of proteins, we obtained 12 similarity features: biological processes, cell components, molecular function, domains and motifs from UniProt, gene expression from Bgee, tissue/cell expression, tissue expression, RNA-seq expression and subcellular locations from the Human Protein Atlas (**Table 1**).

Creation of the gold standard

The PPIs obtained from IntAct are divided between a training set and two testing sets. First, a set of 1,562 proteins (13%) was randomly set aside to ensure some unseen proteins are present in the testing set; the necessity of this is shown in **Figure 2**. The dataset was then randomly divided under this constraint and included 53,331 PPIs in the training set, 12,449 in *T1* and the same in *T2* (**Supplementary Table 1**).

The negative examples (i.e. non interacting proteins) are obtained using random sampling among all the possible pairs, excluding any pair that has been observed experimentally to limit the risk of false negative. For the training set, balanced sampling is used [26] to favour learning, which means that the probably of sampling a protein for the negative set is proportionate to its frequency in the positive set. For *T1* and *T2*, we used uniform sampling (all proteins have the same probability of sampling) to limit the risk of bias. The training set and *T1* both have 50% of positive examples, while *T2* has ten times more non-interacting proteins than interacting ones (**Supplementary Table 1**).

To investigate how models deal with different network topologies, especially hubs and lone proteins, we had to create a separate testing set to ensure sufficient sample size in each category (hub-hub, hub-lone and lone-lone interactions). We do so by aggregating PPIs from *T1* and *T2*, and using balanced sampling for the non-interacting proteins. This results in 49,796 pairs (50% positive) (

Supplementary Table 2).

S. Cerevisiae **data**

The pipeline describe above was also followed for the *S. Cerevisiae* data. UniProt lists 6,721 yeast proteins and the same information as for humans (

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.07.479382; this version posted February 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplementary Table 3) but HPA and Bgee do not include data for this organism. PPIs were obtained from IntAct following the same procedure, although no selection based on MIscores was made considering the absence of an obvious choice when looking at the distribution (**Supplementary Figure 9**). The final PPI dataset comprised 43,068 interactions covering 5,679 proteins.

The split between training and testing sets was done similarly by setting aside 737 proteins for testing and then randomly allocating PPIs to keep 30,369 PPIs for training. Because there is fewer data on yeast, and only one testing set is needed to replicate the analysis conducted on humans, dividing the remaining 12,699 further between *T1* and *T2* is not suitable here. But if the goal was to measure generalisability of a yeast model, this could be easily done.

Training

FG-based machine learning models were trained using the scikit-learn library [61]. For models that cannot deal with missing data, mean imputation was used (**Supplementary Table 4**). Hyper-parameter search was done using Weight-and-Bias's Bayesian method [62] to find the optimal settings of each algorithm in a reasonable time. All hyperparameter choices are in **Supplementary Table 4** and **Supplementary Table 5**.

Deep learning models were trained using PyTorch Lightning [63], [64]. The Siamese architecture [65], [66] was composed of a bidirectional Gated Recurrent Unit (GRU) [67] followed by a linear output (**Figure 9**). Long-Short Term Memory networks (LSTM) [68] and Convolutional Neural Networks (CNN) [69] were also tested, but GRU was preferred because of runtime efficiency and its ability to account for proteins of various lengths. Full parameters are in **Supplementary Table 4**, **Supplementary Table 5** and in the open-source code.

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.07.479382; this version posted February 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

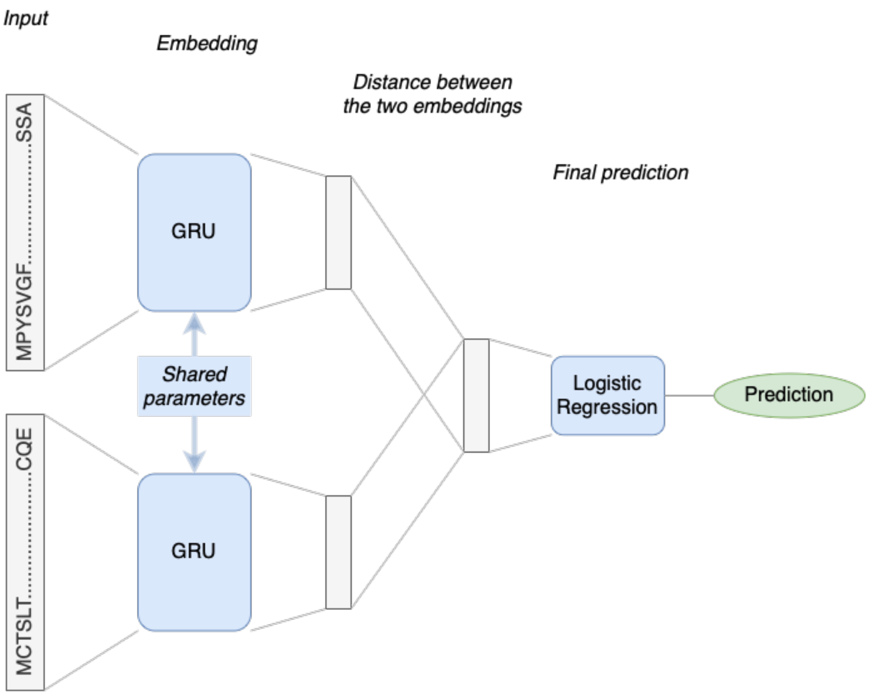


Figure 9: Diagram of the deep learning architecture used to predict interactions from a pair of protein sequences.

Evaluation

The Receiver Operating Characteristic (ROC) and the Precision-Recall (PR) curves are complementary options for PPI prediction. While the ROC curve is unaffected by the prevalence of interacting proteins, a benefit as the true prevalence of PPIs is mostly unknown, it also means that both classes are considered equally, whereas often, PPIs are more interesting than non-interacting proteins. This is addressed by the PR curve where precision puts an emphasis on positive examples.

Both curves are reported alongside their respective Areas Under the Curve (AUC). To statistically compare ROC curves for a same testing set, we used a DeLong nonparametric test [70] and reported the p-value. We corrected for multiple testing by using a conservative significance threshold of 5×10^{-4}, corresponding to a Bonferroni correction for 100 pairwise comparisons [71].

Carbon footprint of this project

We used the Green Algorithms calculator (v2.1) [29] and estimated that the carbon footprint of this project was 51 $kgCO_2e$, which corresponds to 4.7 tree-years. We did our best to minimise greenhouse gas emissions in the first place, and as a commitment to the reduction of the carbon footprint of computational research, we funded tree planting in the east of

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.07.479382; this version posted February 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

England region through carbonfootprint.com. These trees are estimated to sequester 1 tonne of CO_2 in their lifetime, almost 20 times the emissions of this study.

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.07.479382; this version posted February 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

ACKNOWLEDGMENTS

L.L. was supported by the University of Cambridge MRC DTP (MR/S502443/1). M.I. was supported by the Munz Chair of Cardiovascular Prediction and Prevention and the NIHR Cambridge Biomedical Research Centre (BRC-1215-20014)[*]. This work was supported by core funding from the British Heart Foundation (RG/13/13/30194; RG/18/13/33946). This work was also supported by Health Data Research UK, which is funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation and Wellcome. This work was performed using resources provided by the Cambridge Service for Data Driven Discovery (CSD3) operated by the University of Cambridge Research Computing Service (www.csd3.cam.ac.uk), provided by Dell EMC and Intel using Tier-2 funding from the Engineering and Physical Sciences Research Council (capital grant EP/P020259/1), and DiRAC funding from the Science and Technology Facilities Council (www.dirac.ac.uk).

*The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.07.479382; this version posted February 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

References

[1] H. Goehler et al., 'A Protein Interaction Network Links GIT1, an Enhancer of Huntingtin Aggregation, to Huntington's Disease', Molecular Cell, vol. 15, no. 6, pp. 853–865, Sep. 2004, doi: 10.1016/j.molcel.2004.09.016.

[2] A. Vinayagam et al., 'A Directed Protein Interaction Network for Investigating Intracellular Signal Transduction', Science Signaling, vol. 4, no. 189, pp. rs8–rs8, Sep. 2011, doi: 10.1126/scisignal.2001699.

[3] M. Bakail and F. Ochsenbein, 'Targeting protein–protein interactions, a wide open field for drug design', Comptes Rendus Chimie, vol. 19, no. 1–2, pp. 19–27, Jan. 2016, doi: 10.1016/j.crci.2015.12.004.

[4] S. Rapposelli, E. Gaudio, F. Bertozzi, and S. Gul, 'Editorial: Protein–Protein Interactions: Drug Discovery for the Future', Front. Chem., vol. 9, p. 811190, Nov. 2021, doi: 10.3389/fchem.2021.811190.

[5] R. Jansen, 'A Bayesian Networks Approach for Predicting Protein-Protein Interactions from Genomic Data', Science, vol. 302, no. 5644, pp. 449–453, Oct. 2003, doi: 10.1126/science.1087361.

[6] L. V. Zhang, S. L. Wong, O. D. King, and F. P. Roth, 'Predicting co-complexed protein pairs using genomic and proteomic data integration', BMC Bioinformatics, p. 15, 2004.

[7] X.-W. Chen and M. Liu, 'Prediction of protein–protein interactions using random decision forest framework', Bioinformatics, vol. 21, no. 24, pp. 4394–4400, Dec. 2005, doi: 10.1093/bioinformatics/bti721.

[8] A. Ben-Hur and W. S. Noble, 'Kernel methods for predicting protein-protein interactions', Bioinformatics, vol. 21, no. Suppl 1, pp. i38–i46, Jun. 2005, doi: 10.1093/bioinformatics/bti1016.

[9] M. S. Scott and G. J. Barton, 'Probabilistic prediction and ranking of human protein-protein interactions', BMC Bioinformatics, vol. 8, no. 1, p. 239, Jul. 2007, doi: 10.1186/1471-2105-8-239.

[10] M. Kotlyar et al., 'In silico prediction of physical protein interactions and characterization of interactome orphans', Nature Methods, vol. 12, no. 1, pp. 79–84, Jan. 2015, doi: 10.1038/nmeth.3178.

[11] Y. Murakami and K. Mizuguchi, 'Homology-based prediction of interactions between proteins using Averaged One-Dependence Estimators', BMC Bioinformatics, vol. 15, no. 1, p. 213, Jun. 2014, doi: 10.1186/1471-2105-15-213.

[12] Z.-H. You, M. Zhou, X. Luo, and S. Li, 'Highly Efficient Framework for Predicting Interactions Between Proteins', IEEE Transactions on Cybernetics, vol. 47, no. 3, pp. 731–743, Mar. 2017, doi: 10.1109/TCYB.2016.2524994.

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.07.479382; this version posted February 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[13] M. Chen et al., 'Multifaceted protein–protein interaction prediction based on Siamese residual RCNN', Bioinformatics, vol. 35, no. 14, pp. i305–i314, Jul. 2019, doi: 10.1093/bioinformatics/btz328.

[14] M. Kotlyar, A. E. M. Rossos, and I. Jurisica, 'Prediction of Protein-Protein Interactions', Current Protocols in Bioinformatics, vol. 60, no. 1, p. 8.2.1-8.2.14, 2017, doi: 10.1002/cpbi.38.

[15] T. Sun, B. Zhou, L. Lai, and J. Pei, 'Sequence-based prediction of protein protein interaction using a deep-learning algorithm', BMC Bioinformatics, vol. 18, no. 1, p. 277, May 2017, doi: 10.1186/s12859-017-1700-2.

[16] F. Li, F. Zhu, X. Ling, and Q. Liu, 'Protein Interaction Network Reconstruction Through Ensemble Deep Learning With Attention Mechanism', Front Bioeng Biotechnol, vol. 8, p. 390, May 2020, doi: 10.3389/fbioe.2020.00390.

[17] A. Ben-Hur and W. Noble, 'Choosing negative examples for the prediction of protein-protein interactions', BMC Bioinformatics, vol. 7, no. Suppl 1, p. S2, 2006, doi: 10.1186/1471-2105-7-S1-S2.

[18] P. Blohm et al., 'Negatome 2.0: a database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis', Nucleic Acids Res, vol. 42, no. Database issue, pp. D396–D400, Jan. 2014, doi: 10.1093/nar/gkt1079.

[19] Y. Park and E. M. Marcotte, 'Revisiting the negative example sampling problem for predicting protein–protein interactions', Bioinformatics, vol. 27, no. 21, pp. 3024–3028, Nov. 2011, doi: 10.1093/bioinformatics/btr514.

[20] L. Hu, X. Wang, Y.-A. Huang, P. Hu, and Z.-H. You, 'A survey on computational models for predicting protein–protein interactions', Briefings in Bioinformatics, vol. 22, no. 5, Sep. 2021, doi: 10.1093/bib/bbab036.

[21] Y. Park and E. M. Marcotte, 'Flaws in evaluation schemes for pair-input computational predictions', Nature Methods, vol. 9, no. 12, pp. 1134–1136, Dec. 2012, doi: 10.1038/nmeth.2259.

[22] S. Orchard et al., 'The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases.', Nucleic Acids Res, vol. 42, no. Database issue, pp. D358-63, Jan. 2014, doi: 10.1093/nar/gkt1115.

[23] S. Orchard et al., 'Protein interaction data curation: the International Molecular Exchange (IMEx) consortium', Nature Methods, vol. 9, no. 4, pp. 345–350, Apr. 2012, doi: 10.1038/nmeth.1931.

[24] M. Agrawal, M. Zitnik, and J. Leskovec, 'Large-scale analysis of disease pathways in the human interactome', Pac Symp Biocomput, vol. 23, pp. 111–122, 2018.

[25] EMBL-EBI, 'Properties of PPINs: scale-free networks | Network analysis of protein interaction data'. https://www.ebi.ac.uk/training/online/courses/network-analysis-of-

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.07.479382; this version posted February 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

protein-interaction-data-an-introduction/protein-protein-interaction-networks/properties-of-ppins-scale-free-networks/ (accessed Nov. 29, 2021).

[26] J. Yu, M. Guo, C. J. Needham, Y. Huang, L. Cai, and D. R. Westhead, 'Simple sequence-based kernels do not predict protein-protein interactions', Bioinformatics, vol. 26, no. 20, pp. 2610–2614, Oct. 2010, doi: 10.1093/bioinformatics/btq483.

[27] F. J. Provost, T. Fawcett, and R. Kohavi, 'The Case against Accuracy Estimation for Comparing Induction Algorithms', in Proceedings of the Fifteenth International Conference on Machine Learning, San Francisco, CA, USA, Jul. 1998, pp. 445–453.

[28] J. Grealey et al., 'The carbon footprint of bioinformatics', Bioinformatics, preprint, Mar. 2021. doi: 10.1101/2021.03.08.434372.

[29] L. Lannelongue, J. Grealey, and M. Inouye, 'Green Algorithms: Quantifying the Carbon Footprint of Computation', Advanced Science, vol. 8, no. 12, p. 2100707, 2021, doi: 10.1002/advs.202100707.

[30] M. Uhlén et al., 'Tissue-based map of the human proteome', Science, vol. 347, no. 6220, Jan. 2015, doi: 10.1126/science.1260419.

[31] P. J. Thul et al., 'A subcellular map of the human proteome', Science, vol. 356, no. 6340, May 2017, doi: 10.1126/science.aal3321.

[32] F. B. Bastian et al., 'The Bgee suite: integrated curated expression atlas and comparative transcriptomics in animals', Nucleic Acids Research, vol. 49, no. D1, pp. D831–D847, Jan. 2021, doi: 10.1093/nar/gkaa793.

[33] M. E. Maron, 'Automatic Indexing: An Experimental Inquiry', J. ACM, vol. 8, no. 3, pp. 404–417, Jul. 1961, doi: 10.1145/321075.321084.

[34] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, Classification and Regression Trees. Taylor & Francis, 1984. [Online]. Available: https://books.google.fr/books?id=JwQx-WOmSyQC

[35] L. Breiman, 'Random Forests', Machine Learning, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[36] Q. C. Zhang et al., 'Structure-based prediction of protein–protein interactions on a genome-wide scale', Nature, vol. 490, no. 7421, pp. 556–560, Oct. 2012, doi: 10.1038/nature11503.

[37] T. Chen and C. Guestrin, 'XGBoost: A Scalable Tree Boosting System', in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16, San Francisco, California, USA, 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[38] A. Ogunleye and Q.-G. Wang, 'XGBoost Model for Chronic Kidney Disease Diagnosis', IEEE/ACM Trans. Comput. Biol. and Bioinf., vol. 17, no. 6, pp. 2131–2140, Nov. 2020, doi: 10.1109/TCBB.2019.2911071.

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.07.479382; this version posted February 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[39] E. Sprinzak and H. Margalit, ‘Correlated sequence-signatures as markers of protein-protein interaction1 1Edited by G. von Heijne’, Journal of Molecular Biology, vol. 311, no. 4, pp. 681–692, Aug. 2001, doi: 10.1006/jmbi.2001.4920.

[40] J. J. Russell et al., ‘Non-model model organisms’, BMC Biol, vol. 15, no. 1, pp. 55, s12915-017-0391–5, Dec. 2017, doi: 10.1186/s12915-017-0391-5.

[41] G. Marmier, M. Weigt, and A.-F. Bitbol, ‘Phylogenetic correlations can suffice to infer protein partners from sequences’, PLoS Comput Biol, vol. 15, no. 10, p. e1007179, Oct. 2019, doi: 10.1371/journal.pcbi.1007179.

[42] X. Zhong and J. C. Rajapakse, ‘Graph embeddings on gene ontology annotations for protein–protein interaction prediction’, BMC Bioinformatics, vol. 21, no. Suppl 16, p. 560, Dec. 2020, doi: 10.1186/s12859-020-03816-8.

[43] L. Becchetti, A. Fazzone, and L. Martini, ‘Network and Sequence-Based Prediction of Protein-Protein Interactions’, arXiv:2107.03694 [cs, q-bio], Jul. 2021, Accessed: Aug. 16, 2021. [Online]. Available: http://arxiv.org/abs/2107.03694

[44] I. A. Kovács et al., ‘Network-based prediction of protein interactions’, Nat Commun, vol. 10, no. 1, p. 1240, Dec. 2019, doi: 10.1038/s41467-019-09177-y.

[45] I. M. Armean, K. S. Lilley, M. W. B. Trotter, N. C. V. Pilkington, and S. B. Holden, ‘Co-complex protein membership evaluation using Maximum Entropy on GO ontology and InterPro annotation’, Bioinformatics, vol. 34, no. 11, pp. 1884–1892, Jun. 2018, doi: 10.1093/bioinformatics/btx803.

[46] S. Mahapatra and S. S. Sahu, ‘Improved prediction of protein–protein interaction using a hybrid of functional-link Siamese neural network and gradient boosting machines’, Briefings in Bioinformatics, no. bbab255, Jul. 2021, doi: 10.1093/bib/bbab255.

[47] S. Sledzieski, R. Singh, L. Cowen, and B. Berger, ‘D-SCRIPT translates genome to phenome with sequence-based, structure-aware, genome-scale predictions of protein-protein interactions’, cels, vol. 0, no. 0, Sep. 2021, doi: 10.1016/j.cels.2021.08.010.

[48] ‘The Python Language Reference — Python 3.10.1 documentation’. https://docs.python.org/3/reference/ (accessed Jan. 10, 2022).

[49] ‘Jupyter Project Documentation — Jupyter Documentation 4.1.1 alpha documentation’. https://docs.jupyter.org/en/latest/ (accessed Jan. 10, 2022).

[50] J. Reback et al., pandas-dev/pandas: Pandas 1.0.3. Zenodo, 2020. doi: 10.5281/zenodo.3715232.

[51] W. McKinney, ‘Data Structures for Statistical Computing in Python’, Austin, Texas, 2010, pp. 56–61. doi: 10.25080/Majora-92bf1922-00a.

[52] C. R. Harris et al., ‘Array programming with NumPy’, Nature, vol. 585, no. 7825, pp. 357–362, Sep. 2020, doi: 10.1038/s41586-020-2649-2.

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.07.479382; this version posted February 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[53] J. D. Hunter, ‘Matplotlib: A 2D Graphics Environment’, Comput. Sci. Eng., vol. 9, no. 3, pp. 90–95, 2007, doi: 10.1109/MCSE.2007.55.

[54] M. Waskom, ‘seaborn: statistical data visualization’, JOSS, vol. 6, no. 60, p. 3021, Apr. 2021, doi: 10.21105/joss.03021.

[55] ‘GitHub - BlakeRMills/MetBrewer: Color palette package in R inspired by works at the Metropolitan Museum of Art in New York’. https://github.com/BlakeRMills/MetBrewer (accessed Jan. 10, 2022).

[56] G. Jin, S. Zhang, X.-S. Zhang, and L. Chen, ‘Hubs with Network Motifs Organize Modularity Dynamically in the Protein-Protein Interaction Network of Yeast’, PLoS ONE, vol. 2, no. 11, p. e1207, Nov. 2007, doi: 10.1371/journal.pone.0001207.

[57] B. Aranda et al., ‘PSICQUIC and PSISCORE: accessing and scoring molecular interactions’, Nat Methods, vol. 8, no. 7, pp. 528–529, Jun. 2011, doi: 10.1038/nmeth.1637.

[58] The UniProt Consortium, ‘UniProt: the universal protein knowledgebase in 2021’, Nucleic Acids Research, vol. 49, no. D1, pp. D480–D489, Jan. 2021, doi: 10.1093/nar/gkaa1100.

[59] M. Arita, I. Karsch-Mizrachi, G. Cochrane, and on behalf of the International Nucleotide Sequence Database Collaboration, ‘The international nucleotide sequence database collaboration’, Nucleic Acids Research, vol. 49, no. D1, pp. D121–D124, Jan. 2021, doi: 10.1093/nar/gkaa967.

[60] A. Singhal, ‘Modern Information Retrieval: A Brief Overview’, p. 9, 2001.

[61] F. Pedregosa et al., ‘Scikit-learn: Machine Learning in Python’, J. Mach. Learn. Res., vol. 12, no. null, pp. 2825–2830, Nov. 2011.

[62] L. Biewald, ‘Experiment Tracking with Weights and Biases’. 2020. [Online]. Available: https://www.wandb.com/

[63] W. Falcon and The PyTorch Lightning team, PyTorch Lightning. 2019. doi: 10.5281/zenodo.3828935.

[64] A. Paszke et al., ‘PyTorch: An Imperative Style, High-Performance Deep Learning Library’, in Advances in Neural Information Processing Systems, 2019, vol. 32. Accessed: Jan. 10, 2022. [Online]. Available: https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html

[65] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, ‘Signature Verification using a “Siamese” Time Delay Neural Network’, in Advances in Neural Information Processing Systems 6, J. D. Cowan, G. Tesauro, and J. Alspector, Eds. Morgan-Kaufmann, 1994, pp. 737–744. Accessed: Oct. 21, 2019. [Online]. Available: http://papers.nips.cc/paper/769-signature-verification-using-a-siamese-time-delay-neural-network.pdf

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.07.479382; this version posted February 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[66] S. Chopra, R. Hadsell, and Y. LeCun, ‘Learning a Similarity Metric Discriminatively, with Application to Face Verification’, in 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05), San Diego, CA, USA, 2005, vol. 1, pp. 539–546. doi: 10.1109/CVPR.2005.202.

[67] K. Cho et al., ‘Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation’, arXiv:1406.1078 [cs, stat], Sep. 2014, Accessed: Nov. 29, 2021. [Online]. Available: http://arxiv.org/abs/1406.1078

[68] S. Hochreiter and J. Schmidhuber, ‘Long Short-Term Memory’, Neural Computation, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[69] Y. LeCun et al., ‘Handwritten Digit Recognition with a Back-Propagation Network’, in Advances in Neural Information Processing Systems, 1990, vol. 2. Accessed: Feb. 04, 2022. [Online]. Available: https://papers.nips.cc/paper/1989/hash/53c3bce66e43be4f209556518c2fcb54-Abstract.html

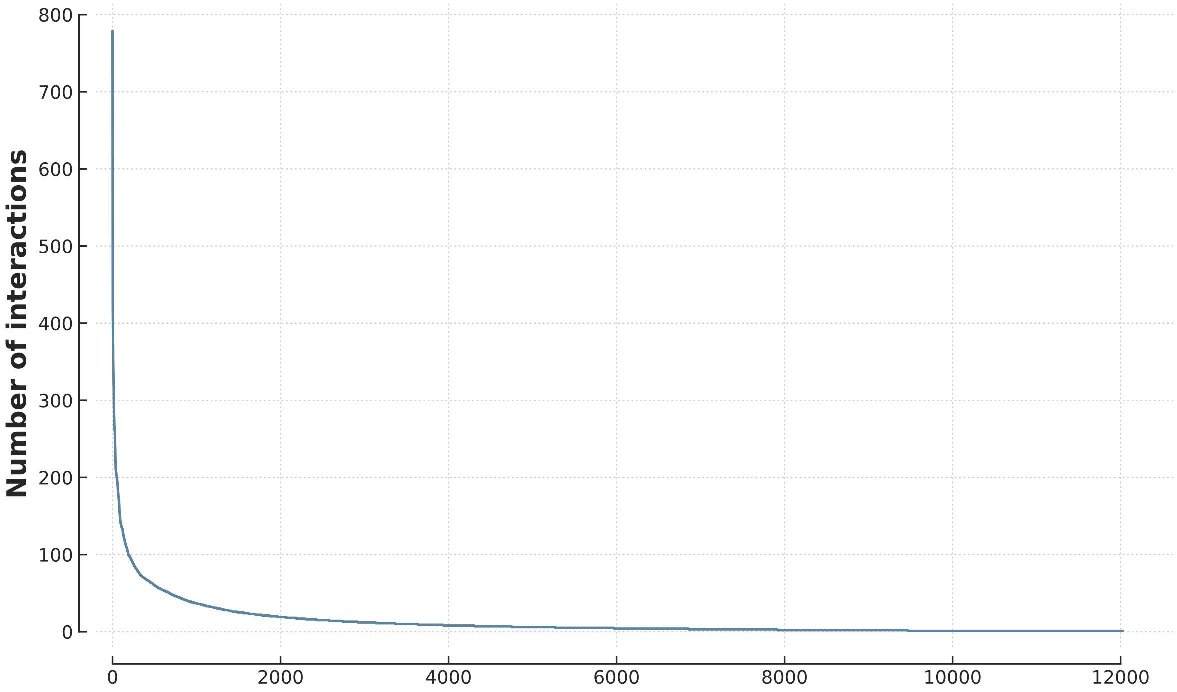
[70] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, ‘Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach’, Biometrics, vol. 44, no. 3, p. 837, Sep. 1988, doi: 10.2307/2531595.

[71] J. Neyman and E. S. Pearson, ‘On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference: Part I’, Biometrika, vol. 20A, no. 1/2, pp. 175–240, 1928, doi: 10.2307/2331945.

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.07.479382; this version posted February 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

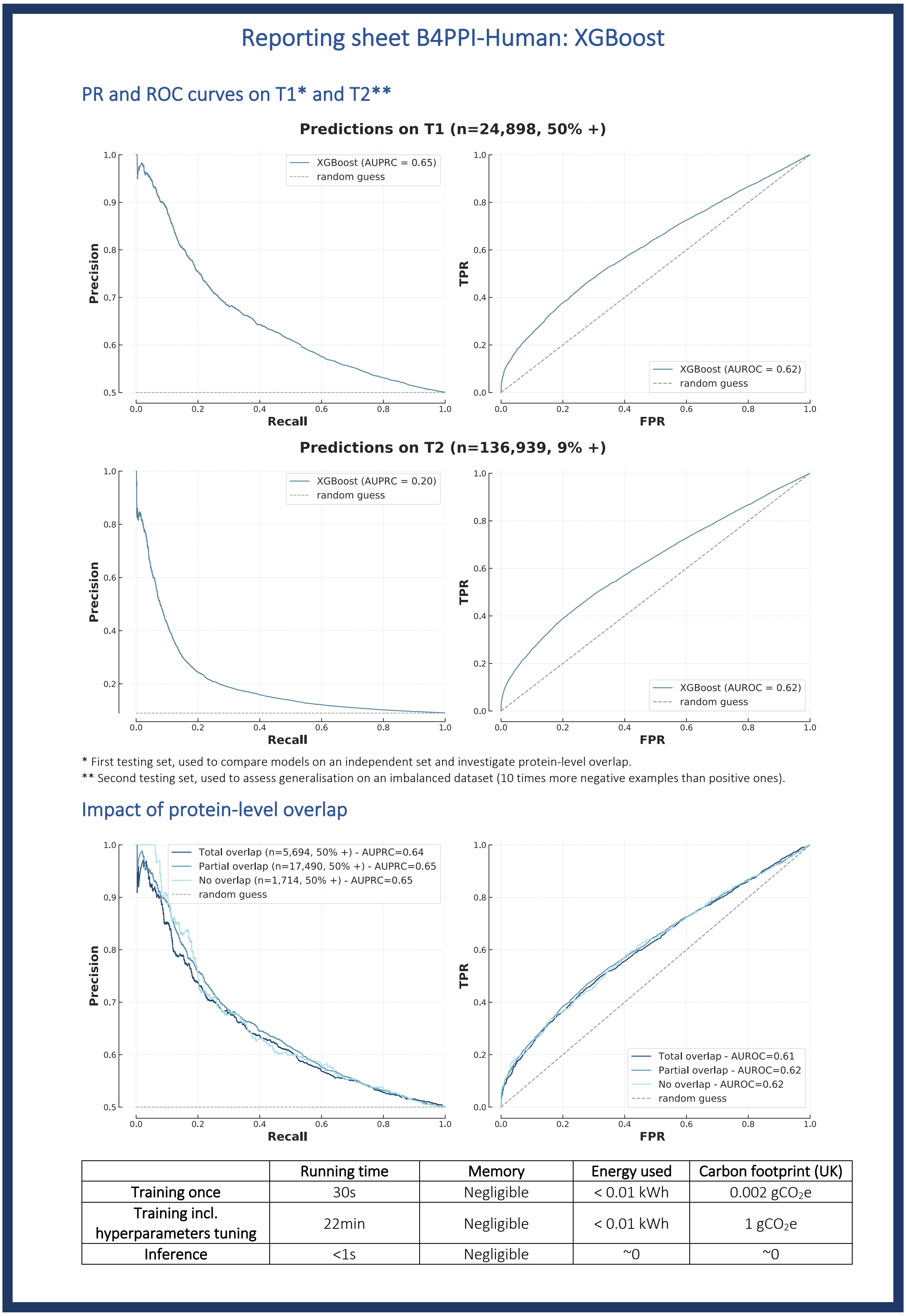
SUPPLEMENTARY MATERIAL

Supplementary Figures



Supplementary Figure 1: Distribution of proteins' degree in IntAct. The exponential decrease is characteristic of a scale-free network.

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.07.479382; this version posted February 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



Reporting sheet B4PPI-Human: XGBoost

PR and ROC curves on T1* and T2**

* First testing set, used to compare models on an independent set and investigate protein-level overlap.

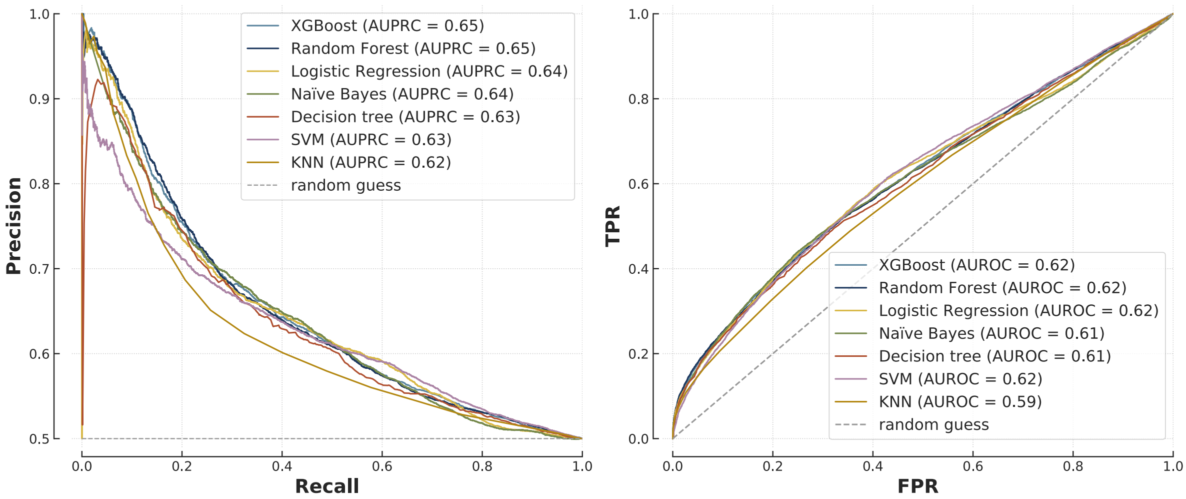
** Second testing set, used to assess generalisation on an imbalanced dataset (10 times more negative examples than positive ones).

Impact of protein-level overlap

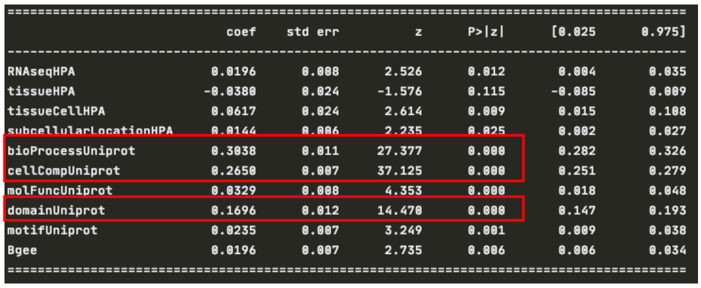
	Running time	Memory	Energy used	Carbon footprint (UK)
Training once	30s	Negligible	< 0.01 kWh	0.002 gCO_2e
Training incl. hyperparameters tuning	22min	Negligible	< 0.01 kWh	1 gCO_2e
Inference	<1s	Negligible	~0	~0

Supplementary Figure 2: Performance sheet of XGBoost on B4PPI-Human.

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.07.479382; this version posted February 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



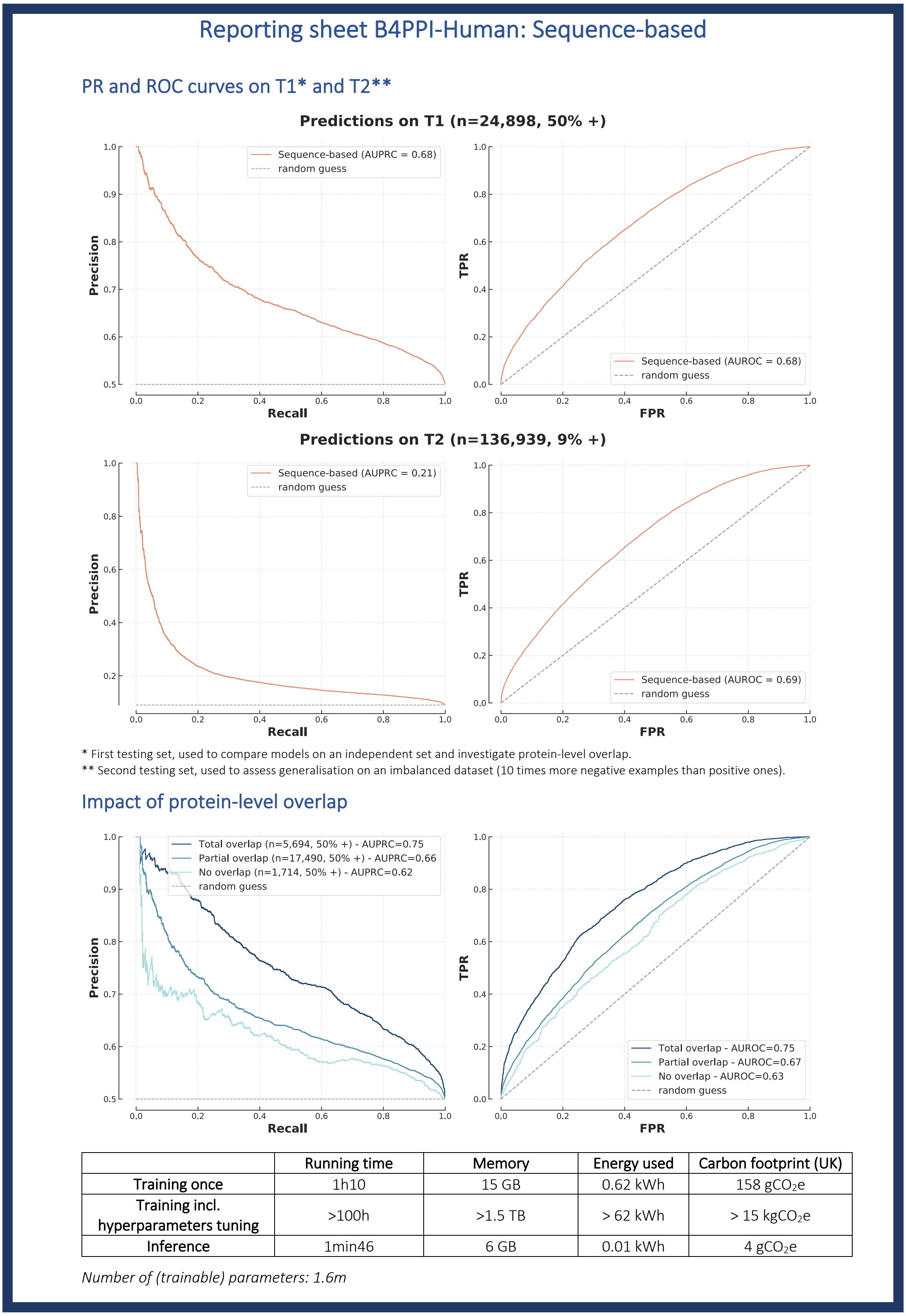
Supplementary Figure 3: Comparison of a broader range of FG-based models.



```
=========================================================================================
                            coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------------
RNAseqHPA                 0.0196      0.008      2.526      0.012       0.004       0.035
tissueHPA                -0.0380      0.024     -1.576      0.115      -0.085       0.009
tissueCellHPA             0.0617      0.024      2.614      0.009       0.015       0.108
subcellularLocationHPA    0.0144      0.006      2.235      0.025       0.002       0.027
bioProcessUniprot         0.3038      0.011     27.377      0.000       0.282       0.326
cellCompUniprot           0.2650      0.007     37.125      0.000       0.251       0.279
molFuncUniprot            0.0329      0.008      4.353      0.000       0.018       0.048
domainUniprot             0.1696      0.012     14.470      0.000       0.147       0.193
motifUniprot              0.0235      0.007      3.249      0.001       0.009       0.038
Bgee                      0.0196      0.007      2.735      0.006       0.006       0.034
=========================================================================================
```

Supplementary Figure 4: Output of the logistic regression on the training set.

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.07.479382; this version posted February 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



Reporting sheet B4PPI-Human: Sequence-based

PR and ROC curves on T1* and T2**

Predictions on T1 (n=24,898, 50% +)

Predictions on T2 (n=136,939, 9% +)

* First testing set, used to compare models on an independent set and investigate protein-level overlap.

** Second testing set, used to assess generalisation on an imbalanced dataset (10 times more negative examples than positive ones).

Impact of protein-level overlap

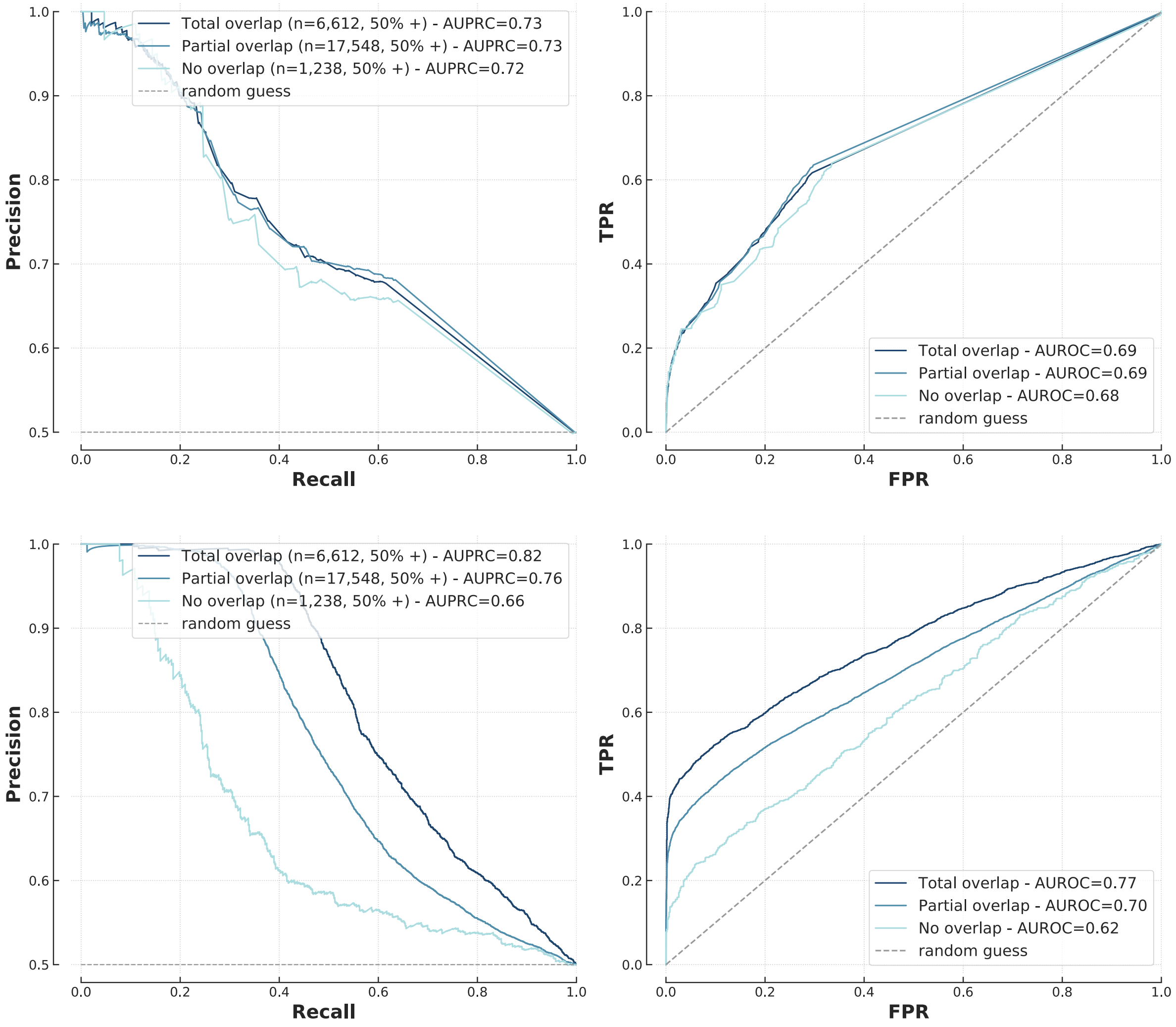
Precision
Recall
Total overlap (n=5,694, 50% +) - AUPRC=0.75
Partial overlap (n=17,490, 50% +) - AUPRC=0.66
No overlap (n=1,714, 50% +) - AUPRC=0.62
random guess
TPR
FPR
Total overlap - AUROC=0.75
Partial overlap - AUROC=0.67
No overlap - AUROC=0.63
random guess
0.0
0.2
0.4
0.6
0.8
1.0
0.5
0.6
0.7
0.8
0.9

	Running time	Memory	Energy used	Carbon footprint (UK)
Training once	1h10	15 GB	0.62 kWh	158 gCO_2e
Training incl. hyperparameters tuning	>100h	>1.5 TB	> 62 kWh	> 15 $kgCO_2e$
Inference	1min46	6 GB	0.01 kWh	4 gCO_2e

Number of (trainable) parameters: 1.6m

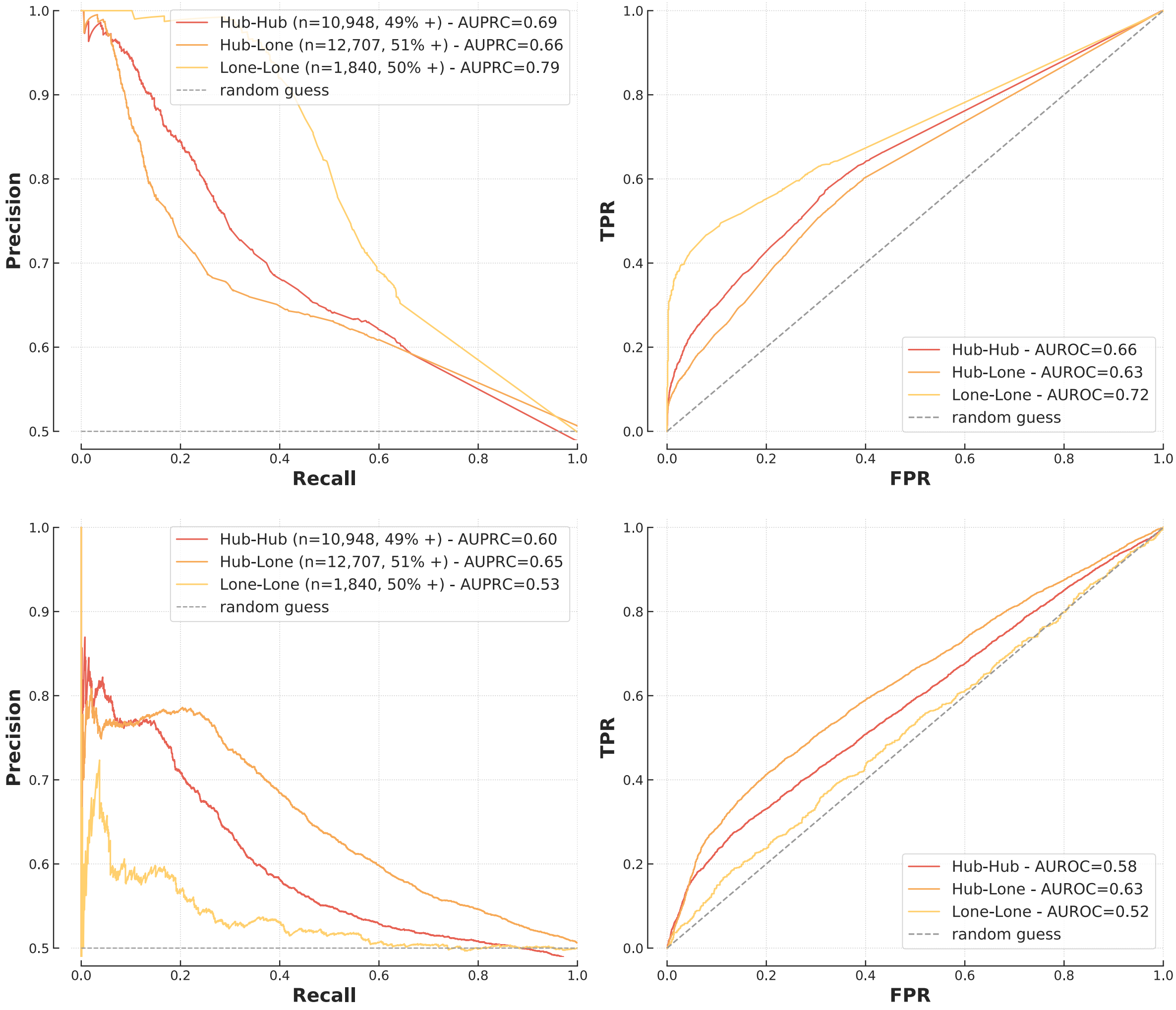
Supplementary Figure 5: Performance sheet of the sequence-based model.

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.07.479382; this version posted February 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



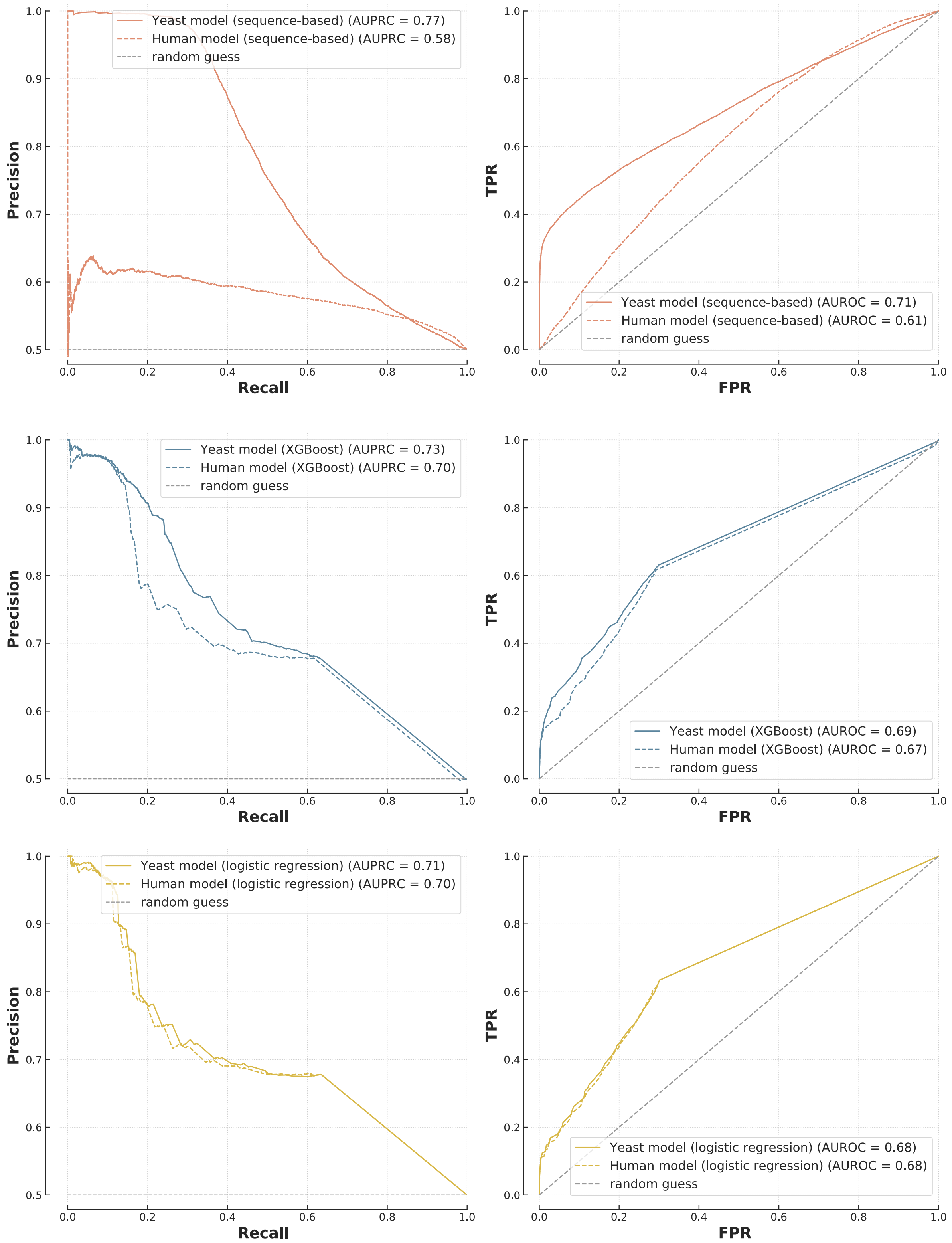
Supplementary Figure 6: Impact of protein-level overlap on the yeast dataset for XGBoost (top, FG-based) and the sequence-based model (bottom).

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.07.479382; this version posted February 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



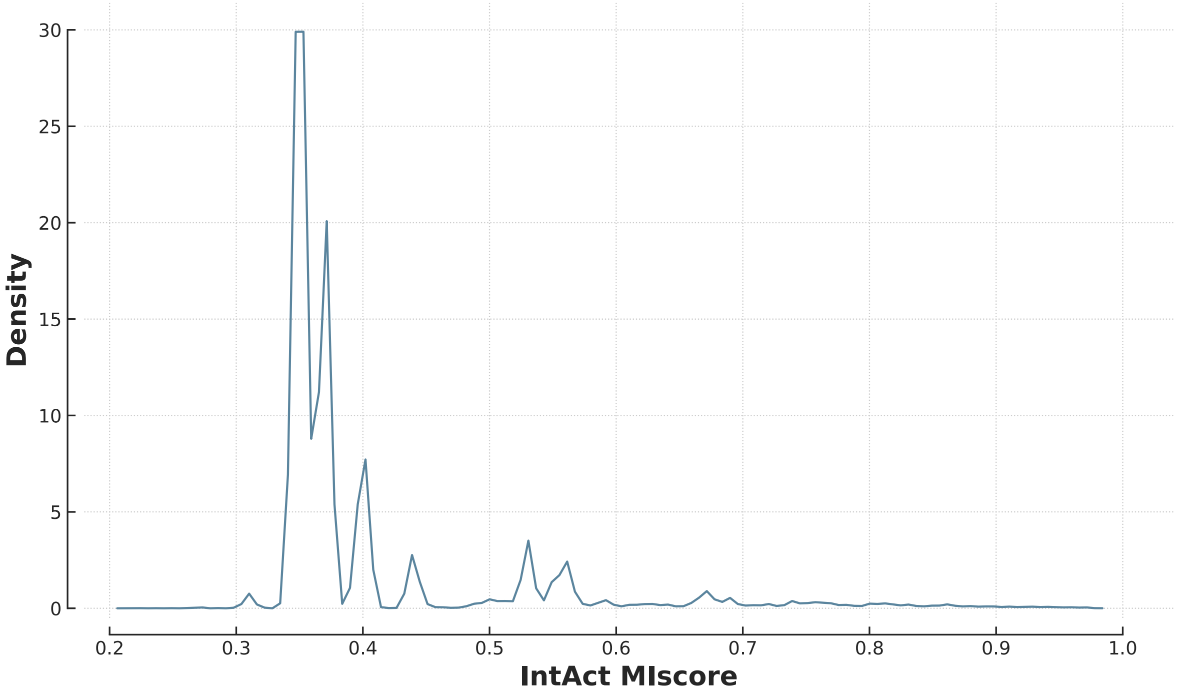
Supplementary Figure 7: Impact of hubs for FG-based (XGBoost, top) and sequence-based (bottom) model on yeast interactions.

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.07.479382; this version posted February 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



Supplementary Figure 8: Cross-species predictions. Models trained on human PPIs (dotted lines) and yeast PPIs (solid lines) were used to make predictions on the yeast testing set (n=25,398, 50% positive). The top plot is the sequence-based method, the others the FG-based ones (XGBoost in the middle, logistic regression at the bottom).

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.07.479382; this version posted February 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



Supplementary Figure 9: distribution of IntAct's MIscore in the yeast dataset.

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.07.479382; this version posted February 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplementary Tables

Supplementary Table 1: Sample size in B4PPI.

Set	Number of examples (% of positive)	
	B4PPI-Human	Yeast dataset
Training	106,662 (50%)	60,738 (50%)
T1	24,898 (50%)	25,398 (50%)
T2	136,939 (9%)	N/A

Supplementary Table 2: Sample size in each category in the dataset used to investigate networks topology.

Type of interaction	Number of pairs	% of PPIs
Hub-hub	27,580	50.69 %
Hub-lone	19,205	49.70 %
Lone-lone	3,011	45.67 %

Supplementary Table 3: Details of the features used for B4PPI-Yeast.

Feature	Number of different annotations	Missing values (/6,721)	Source
Biological processes (GO)	3,114	1,510	UniProt [58]
Cellular components (GO)	820	839	UniProt
Molecular functions (GO)	2,079	2,242	UniProt
Domains	606	5,135	UniProt
Motifs	181	6,273	UniProt
Sequence	N/A	0	UniProt

bioRxiv preprint doi: https://doi.org/10.1101/2022.02.07.479382; this version posted February 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplementary Table 4: Optimal parameters for the models trained on B4PPI-Human

Algorithm	Missing data imputation	Scaling	Optimal hyperparameters
Logistic Regression	Yes (mean)	Yes	Penalty = none, tol = 0.0001
XGBoost	No	No	colsample_bytree = 0.8059, learning_rate = 0.00002186, max_depth = 29, min_child_weight = 25, n_estimators = 116, subsample = 0.4595
Decision Tree	Yes (mean)	No	Criterion = entropy, min_samples_split = 895, splitter = random
SVM	Yes (mean)	No	C = 1, degree = 3, gamma = scale, kernel = rbf (default values were used due to long runtime)
Random Forest	Yes (mean)	No	Criterion = gini, max_features = log2, min_sample_split = 487, n_estimators = 336
KNN	Yes (mean)	No	Algorithm = brute, leaf_size = 53, n_neighbors = 35, p = 2, weights = uniform
Naïve Bayes	Yes (mean)	No	N/A
Sequence-based Siamese architecture	N/A	N/A	Batch size = 200, gradient_clip_val = 10, RNN = bidirectional GRU, output = linear, hidden size = 512, n_layers = 1, learning rate = 0.001 (GRU) and 0.0001 (output)

Supplementary Table 5: Optimal parameters for the models trained on the yeast dataset.

Algorithm	Missing data imputation	Scaling	Optimal hyperparameters
Logistic Regression	Yes (mean)	Yes	Penalty = none, tol = 0.0001
XGBoost	No	No	colsample_bytree = 0.7087, learning_rate = 0.00001129, max_depth = 26, min_child_weight = 3, n_estimators = 244, subsample = 0.9318
Naïve Bayes	Yes (mean)	No	N/A
Sequence-based Siamese architecture	N/A	N/A	Batch size = 200, gradient_clip_val = 10, RNN = bidirectional GRU, output = linear, hidden size = 512, n_layers = 1, learning rate = 0.001 (GRU) and 0.0001 (output)