PLOS ONE

RESEARCH ARTICLE

# ProfPPIdb: Pairs of physical protein-protein interactions predicted for entire proteomes

**Linh Tran[1,2]*, Tobias Hamp[2], Burkhard Rost[2,3]**

**1** Imperial College London (ICL), Department of Computing, United Kingdom, **2** Technical University of Munich (TUM), Department of Informatics, Bioinformatics & Computational Biology - i12, Boltzmannstr, Germany, **3** Technical University of Munich (TUM), Institute for Advanced Study (TUM-IAS), Lichtenbergstr, Germany

* linh.tran@imperial.ac.uk





**Citation:** Tran L, Hamp T, Rost B (2018) ProfPPIdb: Pairs of physical protein-protein interactions predicted for entire proteomes. PLoS ONE 13(7): e0199988. https://doi.org/10.1371/journal.pone.0199988

**Editor:** Christos A. Ouzounis, CPERI, GREECE

**Received:** January 8, 2018

**Accepted:** June 17, 2018

**Published:** July 18, 2018



**Data Availability Statement:** The database interface representing our results is accessible from https://rostlab.org/services/ppipair/. The data can also be downloaded from https://figshare.com/collections/ProfPPI-DB/4141784.

**Funding:** This work was supported by a grant from the Alexander von Humboldt foundation through the German Ministry for Research and Education (BMBF: Bundesministerium fuer Bildung und Forschung) to BR.

**Competing interests:** The authors have declared that no competing interests exist.

## Abstract

### Motivation

Protein-protein interactions (PPIs) play a key role in many cellular processes. Most annotations of PPIs mix experimental and computational data. The mix optimizes coverage, but obfuscates the annotation origin. Some resources excel at focusing on reliable experimental data. Here, we focused on new pairs of interacting proteins for several model organisms based solely on sequence-based prediction methods.

### Results

We extracted reliable experimental data about which proteins interact (binary) for eight diverse model organisms from public databases, namely from *Escherichia coli*, *Schizosaccharomyces pombe*, *Plasmodium falciparum*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Mus musculus*, *Rattus norvegicus*, *Arabidopsis thaliana*, and for the previously used *Homo sapiens* and *Saccharomyces cerevisiae*. Those data were the base to develop a PPI prediction method for each model organism. The method used evolutionary information through a profile-kernel Support Vector Machine (SVM). With the resulting eight models, we predicted all possible protein pairs in each organism and made the top predictions available through a web application. Almost all of the PPIs made available were predicted between proteins that have not been observed in any interaction, in particular for less well-studied organisms. Thus, our work complements existing resources and is particularly helpful for designing experiments because of its uniqueness. Experimental annotations and computational predictions are strongly influenced by the fact that some proteins have many partners and others few. To optimize machine learning, recent methods explicitly ignored such a network-structure and rely either on domain knowledge or sequence-only methods. Our approach is independent of domain-knowledge and leverages evolutionary information. The database interface representing our results is accessible from https://rostlab.org/services/ppipair/. The data can also be downloaded from https://figshare.com/collections/ProfPPI-DB/4141784.

## Introduction

### Operational definition of physical Protein-Protein Interactions (PPIs)

We define PPIs as interactions that bring two different proteins A and B directly into 'physical contact'. This 'molecular' perspective on PPIs differs from the most frequent view of both associations and permanent complexes. For us the crucial aspect of a PPI is that it brings two proteins into direct physical contact (usually transiently, i.e. for a limited time). Given all PPIs in an organism, the *interactome* comprises all PPIs in the entire proteome; this network contains all non-temporal aspects of associations on the network level.

### Experimental annotations of binary PPI maps

Due to the importance, many experiments establish PPIs. Despite this effort, most pairs of physically interacting proteins remain likely unknown [1]. Statistical models of PPIs can amend the coverage of networks formed from binary PPIs (A binds B) cost-effectively by enriching protein association networks [2–4] or by combining heterogeneous data sources in Bayesian networks [5].

### Predictions important but often over-estimated

Numerous computational methods have been developed to predict protein-protein interactions using different data sources, e.g. secondary structure, phylogenetic tree, phylogenetic profile, and gene expression [6–10]. Most methods employ more than one of the mentioned properties. However, their application is limited due to their specific need of domain knowledge. These specific knowledge is but not universally available, and limit these methods to specific (smaller) datasets.

Further, many methods only use sequence information, such as motifs of co-occurrence on the level of domains [11–13], matching features from protein sequence, structure and evolutionary conservation for binding sites alone [10, 14] and for binding sites and sequence/structure triads [15]. However, none of those sequence-based methods restrict their method to the identification of physical non-permanent PPIs as we defined them. Most of those methods used permanent complexes, the others also associations. This is also true for methods pioneering the use of kernel-based predictions [14, 15]. Evolutionary information embedded in proteins sequence was employed to improve predicting PPIs [10, 14, 16, 17], some in combination with profile kernels [18], by leveraging information available to us which are not domain specific.

Another set of problems with existing methods pertain to the problems in choosing "negatives", i.e. pairs of proteins known not to interact [19]. In fact, negatives have to be carefully considered when setting up the cross-validation process [20]. Moreover, the cross-validation protocol also needs to carefully avoid using the same proteins in training and testing [21, 22], and even allowing for homologues between training and testing over-estimates performance [20]. Overall, it appears that every careful independent review of existing methods has unraveled some substantial over-estimates [20–22]. One recent method combining profile kernels with Support Vector Machines (SVM) to predict pairs of physical, non-permanent PPIs has tried to avoid all known flaws [23]. However, it still awaits critical assessment from independent experts. This method improved particularly for proteins without experimental annotations about their interactions recommending the approach for discovery of novel PPIs [23].

Here, we simply apply the concept of profile-kernel SVMs [23] to the prediction of the entire interactomes in eight model organisms, namely ordered by size: *Escherichia coli*, *Schizosaccharomyces pombe* (fission yeast), *Plasmodium falciparum*, *Drosophila melanogaster* (fruit

fly), *Caenorhabditis elegans* (roundworm), *Mus musculus* (mouse), *Rattus norvegicus* (rat), and *Arabidopsis thaliana* (mouse-ear cress). The choice of applying profile-kernel SVMs is due to its independence of domain knowledge and its usage of evolutionary profiles. Further, in vast evaluation we chose negative interactions by avoiding using the similar proteins in training and testing. Repeated cross-validation was employed to reduce additional over-estimation as stated in [20–22]. We have created a database of the most reliable predictions for each organism, and implemented a versatile online search interface (https://rostlab.org/services/ppipair/). Our new methods and new predictions at least double the number of organisms for which sequence-based PPI predictions are available, and they do this in a more consistent way than other method [24]. On top, our resource contributes the first-ever predictions for many unannotated proteins.

## Materials and methods

### Data Sources

We extracted PPIs from the following databases BioGRID [25], DIP [26], and IntAct [27]. BioGRID is a public curated database that holds 553,827 physical interactions from 58 species. DIP archives 795,534 PPIs from 777 organisms, curated both manually by experts and through computational approaches. IntAct is also public archiving 356,806 PPIs mostly from eight organisms. All PPIs originated either from publications or submissions from experimentalists.

### Data Extraction

We only used PPIs for which their protein identifiers mapped to the EBI reference proteomes [28]. We mapped proteins of each organism to a corresponding reference protein only if their sequences aligned with at least 95% sequence identity. The fraction of PPIs that could not be mapped in this simple manner accounted for about 9% of all data. We grouped the resulting PPIs by organism using taxonomy identifiers and differentiated PPIs from 768 organisms.

To predict PPIs, we needed as much reliable training data as possible. However, we also need to remove redundancy in many non-trivial ways [23]. We used an established expert knowledge-scoring scheme [29] (S4 Fig) to reflect the quality of evidence for a given PPI. The scheme assigned scores from one (lowest reliability) to ten (highest reliability) for each experimental method used to annotate a PPI. High scores were assigned to techniques such as X-ray crystallography or electron tomography, average scores of five were given to, e.g. complementation-based assays and affinity-based technologies. Methodologies that do not directly provide evidence for interaction, such as co-localization or co-sedimentation, were scored lowest. The scores are available online at our service. We applied that scheme to our PPI data and kept only PPIs with at least one experimental evidence $\geq 5$. For instance, the *Escherichia coli* PPI between P0ABB0 and P0ABB4 is supported by two experimental methods: blue native page (score = 3) and pull down (score = 2.5); both below 5, i.e. we discarded this PPI. In contrast the PPI between P0ACF0 and P03004 established by enzyme linked immunosorbent assay (score = 5) was kept. After data filtering, we redundancy reduced the PPI set of each organism set such that no PPI pair was sequence-similar. A PPI pair was considered similar if at least one of the two sequences reached HVAL $> 20$ [30] to any protein already in the data set. Note that HVAL $> 20$ corresponds to $> 40\%$ pairwise sequence identity for alignments over 250 residues.

We applied the above procedure to all 768 organisms for which we found PPIs. Only 8 of the 768 had at least 200 PPIs with strong experimental support. We considered these our 'model organisms'. 200 PPI was the minimum number of data points we assumed to be necessary to train our method. Redundancy reduction shrank our data by over ten-fold for some

**Table 1. Data sets extracted from BioGRID, DIP and IntAct.** Organism: latin name for eight model organisms sorted alphabetically; NPPIs: number of distinct physical pairs of protein-protein interactions extracted by merging the entire BioGRID, DIP, and IntAct; NPPIs with strong evidence: subset of previous column with reliable experimental evidence (according to [29]); NPPIs with strong evidence redundancy reduced: subset of previous column after removing sequence-similar pairs (HVAL > 20).

| Organism | NPPIs | NPPIs with strong evidence | NPPIs with strong evidence redundancy reduced |
|---|---|---|---|
| *A. thaliana* | 38,258 | 8,459 | 814 |
| *C. elegans* | 23,105 | 5,229 | 818 |
| *D. melanogaster* | 79,291 | 19,033 | 1,680 |
| *E. coli* | 27,119 | 8,587 | 998 |
| *M. musculus* | 30,070 | 6,262 | 734 |
| *P. falciparum* | 4,792 | 1,312 | 239 |
| *S. pombe* | 13,478 | 4,396 | 410 |
| *R. norvegicus* | 6,698 | 1,574 | 236 |
| *Sum over all 8* | 222,811 | 54,852 | 5,929 |

https://doi.org/10.1371/journal.pone.0199988.t001

organisms (Table 1). The most extreme attrition was for fly for which we extracted almost 80k PPIs from the databases, and could use only about 1.6k for training/testing.

## Negative interactions

Databases collect positives (A binds B), i.e. PPIs with experimental evidence. For training, we also needed negatives (A does not bind B). We collected negatives as described before in [20, 23]. For each PPI data set, we sampled negatives in a ratio of 1:10 (10 negative for each positive). The 1:10 ratio seemed appropriate to provide enough negatives to sample the reality in a cell. As before in [20] and [23], we obtained negatives by randomly sampling from all possible combinations of proteins of an organism with the restrictions that each protein in a 'negative PPI' needed to differ in sequence (HVAL < 20) to all proteins in the positive training set.

## Profile-kernel SVM parameter optimization and cross-validation

Many advanced sequence-based PPI prediction methods have been developed. Park and Marcotte [22] showed that PIPE2 [24], AutoCorrelation [31], and SigProd [32] performed well compared to other methods. We showed a profile-kernel SVM to improve over these methods for human and yeast [23]. This method is described in detail in [23]. The basic concept is described in the following. Essentially, the profile-kernel finds k-mers of k adjacent residues for which the conservation within a given protein family exceeds some value $\sigma$ and then collects the most informative such k-mers through SVMs. Thus, as for each profile-kernel SVM [33], we needed to optimize two hyperparameters: the k-mer length k and the evolutionary score threshold $\sigma$. Following our previous experience, we sampled $k = 3, 4, 5, 6$ for $\sigma = 4, \ldots, 11$. For all organisms with more than 500 non-redundant PPIs, we optimized the two parameters empirically with a grid-search on two-thirds of the PPI data for each organism (*training set*). The remaining third of each data set (*test set*) was used to confirm generalization. Each training set was split further into five parts. For every hyperparameter combination, we performed a full 5-fold cross-validation using four splits for *cross-validation training* and one for *cross-validation testing*. In this way, each of the five splits of the full non-redundant set was used as cross-training split exactly once. We repeated each 5-fold cross-validation five times from the start, including splitting positives and sampling negatives, in order to minimize sampling noise [34]. Finally, we used the best combination of k and $\sigma$ and the entire training set to

train the method one last time in order to predict the test set. For organisms with more than 200, but fewer than 500 PPIs (Table 1), we did not optimize parameters, but only evaluated their performance in a five times 5-fold cross-validation on the whole data set. As hyperparameters, we used the most common combination found for the larger PPI sets ($k = 5$, $\sigma = 11$).

### Evolutionary profiles

The evolutionary profiles were taken from PredictProtein [35]. They were created by PSI-BLAST-ing [36] queries against an 80% non-redundant database combining UniProt [37] and PDB [38]. Our method never used any information not available through these profiles.

### Recall-precision curves

Each model built from a training data set outputs a score for each prediction. We used these scores to calculate precision-recall-curves. In a cross-validation, we used all precisions at a particular recall to calculate the mean and the standard deviation of the precisions at that point. If only one curve was available (assessment of hold-out sets for organisms with $>$ 200 PPIs), we assumed precision to follow a standard binomial distribution and calculated the precision error at a particular recall as:

$$e = n_{PPI} \cdot p \cdot (1 - p), \tag{1}$$

where $n_{PPI}$ denotes the number and p denotes the precision at that particular recall. In order to assess a particular parameter combination, we needed to condense the associated recall-precision curve into a single point. We did this by collecting all mean precision values until a recall of 20% and then averaging over those values. The best parameter combination optimized this average precision.

### Interactome predictions

For predicting the entire interactomes, we used all available PPI data (training + test set) our models. As the hyper-parameters values $k = 5$, $\sigma = 11$ yielded best performance for almost all organisms, we used those parameters for our interactome model for all organisms. This might not be the optimal solution, but it might provide the most conservative result avoiding more over-fitting. We applied our method to all pairs of proteins for which both proteins were dissimilar to any protein in the positives used for training. We chose to only publish the most reliable PPIs accounting to about 10 PPIs per protein of an organism (numbers given in Table 2).

## Results and discussion

### Similar prediction performances between many organisms

Accumulating all non-redundant PPIs from the curated databases BioGRID, DIP and IntAct with reliable experimental annotations left only five organisms with over 500 PPIs enough to develop and evaluate organism-specific new methods using profile-kernel SVMs to predict PPIs from sequence: *Escherichia coli*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Mus musculus*, and *Arabidopsis thaliana* (Table 1). For each organism, two thirds of the data served for training and one-third as an independent test set. Training revealed that a $k$-mer length of $k = 5$ and conservation threshold $\sigma = 11$ were optimal for all organisms except *Escherichia coli* (Section A in S1 Appendix, Fig A). For simplicity, we used this hyper-parameter combination for all species (S1 Fig). Three other organisms (*Schizosaccharomyces pombe* (fission yeast), *Plasmodium falciparum*, and *Rattus norvegicus* (rat)) have too few experimental PPIs to fully optimize all parameters (Table 1: 236-410 PPIs). We evaluated the performance for these

**Table 2. Whole interactome predictions.** For each organism investigated, we aggregated the data we used for training and testing, trained a final model and predicted the whole interactome of that organism. Organism: latin name for eight model organisms sorted alphabetically; Nprot: number of proteins in proteome (values taken from [28]); NpredPPI: subset of PPIs used for prediction in which both proteins are dissimilar to the proteins in the positive interactions of the training set; NprotPred: corresponding number of proteins for which NpredPPI interactions were predicted, see Eq 1 for calculation; NpredPPI novel: denotes the number of predicted PPIs for which both proteins are dissimilar to any known positive interaction, including redundant and low-quality PPIs; NpredPPI ProfPPIdb: subset with strongest predictions of previous column contained in our resource; Nprot ProfPPIdb: number of unique proteins in the PPIs published at https://rostlab.org/services/ppipair/, as well as https://figshare.com/collections/ProfPPI-DB/4141784.

| Organism | Nprot | NpredPPI (NprotPred) | NpredPPI novel | NpredPPI ProtPPIdb | Nprot ProtPPIdb |
|---|---|---|---|---|---|
| *A. thaliana* | 27,064 | 206,441,040 (20,320) | 71,251,953 | 250,000 | 7,023 |
| *C. elegans* | 20,137 | 142,171,953 (16,863) | 83,301,778 | 200,000 | 7,041 |
| *D. melanogaster* | 13,707 | 31,916,055 (7,990) | 5,410,405 | 100,000 | 7,664 |
| *E. coli* | 4,306 | 2,729,616 (2,337) | 332,520 | 40,000 | 1,341 |
| *M. musculus* | 22,136 | 131,325,321 (16,207) | 58,790,746 | 200,000 | 4,144 |
| *P. falciparum* | 5,159 | 9,041,878 (4,253) | 5,622,981 | 50,000 | 3,576 |
| *S. pombe* | 5,121 | 8,349,741 (4,087) | 2,630,071 | 50,000 | 3,946 |
| *R. norvegicus* | 21,330 | 211,922,578 (20,588) | 174,929,160 | 200,000 | 9,265 |
| *Sum over all 8* | 118,960 | 743,898,182 (92,645) | 402,269,614 | 1,090,000 | 44,000 |

https://doi.org/10.1371/journal.pone.0199988.t002

organisms in a 5-fold cross-validation using the default parameters $k = 5$, $\sigma = 11$ as fixed parameters (S2 Fig).

For three of the five organisms (*Caenorhabditis elegans*, *Arabidopsis thaliana* and *Mus musculus*) our method performed on a similar level as our method predicting PPIs in human (S1 Fig). For low recall ($\leq 0.1$), the average precision for those three organisms appeared to even slightly (and significantly) exceed the values for human. However, our newly developed models for *Escherichia coli* and *Drosophila melanogaster* performed less well than the method for human. For *Escherichia coli*, changing the hyperparameters to $k = 3$, $\sigma = 4$ improved the performance (Section A in S1 Appendix, Fig B). We used the same hyperparameters for all eight models although we knew before using the testing set that this solution was not optimal. We did this as an additional precaution against over-fitting. For *Drosophila melanogaster* (fly) with over 1600 PPIs, we had no explanation for the dip in performance. In fact, the PPI predictions for fly appeared to be the worst amongst all ten organisms for which we applied our formalism (including human and baker's yeast) although we had the highest number of PPIs for training. For fly we also observed by far the highest attrition from PPIs with 'some experimental evidence' to 'non-redundant PPIs with strong experimental evidence' (Table 1: column 'Number of PPIs' vs. column 'Number of PPIs with strong evidence'). However, we see no reason why this attrition should impact the consistency of the PPI data left over.

For organism with fewer than 500 PPIs (*Schizosaccharomyces pombe*, *Plasmodium falciparum* and *Rattus norvegicus*), we only evaluated the model performance with 5-fold cross-validation (S2 Fig). Our PPI prediction model for human appeared to perform better than the prediction models for these three organisms. This was most likely due to a lack of training data.

## Experimental evidences of novel predictions

We analysed our novel predictions by searching for any experimental evidence in databases such as BioGRID [25], DIP [26], IntAct [27], STRING [39], MINT [40] and Mentha [41]. All these databases have aggregated information of PPIs with experimental evidences. STRING [39], MINT [40] and Mentha [41] also provide confidence measures. Although the databases BioGRID [25], DIP [26], and IntAct [27] were already used for our organism-specific models, only a small subset of the databases' PPIs was employed for training. The PPIs published on our online service only include PPIs which have not any experimental evidence from any of these three databases. In order to perform an evaluation of the quality of the predictions, we used the top 1% of all predictions (ranked according to our confidence measure) which were not included in the training set. We compared these predictions against all experimental from BioGRID [25], DIP [26], and IntAct [27]. Overall, we found a total number of 772 PPIs with evidence which results in an average 86.79% accuracy of correctly predicted PPIs. We also found evidences of PPIs for PPIs which our models did not predict any direct physical interaction. However, these evidences were usually experimental evidences with expert knowledge scores of lower or equal 4 [29] and thus highly likely to be false positives. A more detailed description of our findings can be found in the supplementary materials (Section B in S1 Appendix).

While we have only found minor number of PPIs with evidences in MINT [40] and Mentha [41], we found a significant portion of evidence in the STRING [39] database. Table 3 shows the number of evidences found of our evaluation with the STRING [39] database and includes numbers of evidences conforming with our predictions as well as the resulting accuracy. Except for *Mus musculus*, *Plasmodium falciparum* and *Rattus norvegicus*, we have found more than 1000 PPIs per organism with evidence in the database. With a high number of correctly predicted PPIs (both our prediction and STRING score indicate a PPI), we can observe a correlation between our most reliable PPIs and STRING PPI score. The average accuracy of positive predicted PPIs with STRING evidences is at 86.34%, with the lowest accuracy at 75.19% (*Schizosaccharomyces pombe*).

S3 Fig illustrates the distribution of the experimental evidences found in STRING [39] plotted against their STRING scores. For *Arabidopsis thaliana* (S3 Fig, first row, first column), evidences in STRING were found for only positive predicted PPIs (1138 evidences). This results in about ≈ 70% of the predictions having a STRING confidence score between 0.4 and 0.6, and the remaining ≈ 30% having a high confidence score between 0.6 and 1.0. For

**Table 3. Summary of experimental evidences found in STRING [39].** organism: latin name for eight model organisms sorted alphabetically; NpredPPIs: number of PPIs of 1% ranked predictions; NEvidence: number of PPIs for which experimental evidences was found in at least on of the three databases used for training; NcorrectEvidence: number of PPIs with experimental evidence which were correctly classified by our approach; Accuracy: fraction of correct predictions within the predictions with experimental evidence.

| Organism | NpredPPIs | NEvidence | NcorrectEvidence | Accuracy |
|---|---|---|---|---|
| *A. thaliana* | 250,000 | 1,138 | 1,138 | 100.00% |
| *C. elegans* | 200,000 | 1,671 | 1,818 | 91.91% |
| *D. melanogaster* | 100,000 | 1,763 | 2,049 | 86.04% |
| *E. coli* | 40,000 | 1,807 | 2,196 | 82.29% |
| *M. musculus* | 200,000 | 0 | 0 | - |
| *P. falciparum* | 50,000 | 0 | 0 | -% |
| *S. pombe* | 50,000 | 1,088 | 1,447 | 75.19% |
| *R. norvegicus* | 200,000 | 0 | 0 | - |
| *Sum over all 8* | 1,090,000 | 7,467 | 8,648 | 86.34% |

https://doi.org/10.1371/journal.pone.0199988.t003

*Caenorhabditis elegans* (S3 Fig, first row, second column) and *Escherichia coli* (S3 Fig, second row, first column), the accuracy of positive predicted PPIs found in STRING amounts to respectively 91.91% and 82.29%. Plotting the distribution of positive and negative predicted evidences found in STRING, both plots for *Caenorhabditis elegans* and *Escherichia coli* show similar distribution between positive and negative predicted PPI. In both cases, we found equal distribution of lower and higher STRING confidence score for both positive and negative predicted PPIs. In contrast, *Drosophila melanogaster* (S3 Fig, first row, third column) and *Schizosaccharomyces pombe* (S3 Fig, second row, second column) show a difference in distribution between positive and negative predicted PPIs. We observe a high percentage of STRING scores (below 0.5 for more than 80% of the evidences) for negative PPIs, and a high percentage of high STRING scores (above 0.7 for 50% of the evidences found). The negative predictions which were still found in STRING are likely to be false positive, as according to [39]: "A score of 0.5 would indicate that roughly every second interaction might be erroneous (i.e., a false positive)."

## Insights from novel predictions

The majority of PPIs predicted by our models has not been reported in any of the three databases that we used at any level of reliablity (BioGRID, DIP, and IntAct). Column 4 of Table 2 (NpredPPI novel) summarizes the number of novel PPIs predicted for each organism; novel means that they differ from all experimentally known PPIs, including redundant and low-quality PPIs. Even if we assumed that only one in 20 of the positive predictions were right, these large numbers demonstrated that even for the best studied organisms, millions of PPIs without a close homolog from which interactions could be inferred remain unknown.

What can be stated about those newly predicted PPIs? While there is no answer for the millions, we investigated the most reliable 100 PPI predictions for *Escherichia coli* (note 'only' about 300k PPIs were predicted novel in *Escherichia coli*). 79 of these 100 PPIs were annotated to involve DNA-binding proteins. We are aware of very few DNA-binding proteins that do not bind to other proteins. Thus, the fact that DNA-binding proteins are involved in almost 80% of all our top predictions of PPIs that were never seen before seemed at least encouraging. However, we did not find any clear evidence supporting any one of those 79 PPIs explicitly. 15 of the 100 top PPIs were annotated to involve repressing molecular binding. For example, *Escherichia coli* proteins P0ACP7 and P0ACQ0 were predicted with strong reliability (probability = 0.999977). Both proteins were classified as repressors by UniProt [42]. Transcriptional repression is an important aspect of gene regulation. As in most areas of molecular biology, studies of *Escherichia coli* have provided the model for subsequent investigations of transcription in different organisms, in particular in eukaryotic cells [43]. We were, therefore, surprised that some of our strongest predictions of PPIs never seen before involved *Escherichia coli* repressors. Again, we did not find any explicit experimental data to support or refute these 15 novel PPI predictions.

Further findings include Zinc finger (ZnF) domains, which are widely distributed in eukaryotic genomes. It has been estimated that around 1% of all genes encode proteins containing ZnFs and those proteins often contain multiple repeats of ZnFs [44]. Their functions are extraordinarily diverse and include DNA recognition, RNA packaging, transcriptional activation, regulation of apoptosis, protein folding and assembly, and lipid binding. Zinc finger structures are as diverse as their functions. In general, little is known about these protein—protein interactions [45]. We analysed the molecular function using Gene Ontology (GO, [46]). Interestingly, zinc ion binding is a molecular function which 81 of the top 5000 *Drosophila melanogaster* protein pairs of positive predicted PPIs have in common as well as 5 of the top

1000 *Caenorhabditis elegans* PPIs. However, protein pairs both being zinc ion binding in *Arabidopsis thaliana* (181 of the top 5000) and in *Schizosaccharomyces pombe* (7 of the top 1000) are common functions of protein pairs highly unlikely to interact. Similar to our findings about *Escherichia coli*, we did not find any explicit experimental data to support or refute these interactions relating to zinc ion binding protein pairs.

## Limitation of performance evaluation

Several problems were in the way to providing a completely convincing comprehensive performance assessment. Specific to our problem were the rather small data sets of experimentally characterized PPIs: fewer than 6,000 non-redundant PPIs for all 8 organisms. In order to avoid severe problems from database bias, we had to focus on high-quality non-redundant PPIs [23]. As our profile-kernel based SVM requires at least 200 reliable PPIs, the number of acquired non-redundant PPIs reduced the set of organisms to only 8. The additional challenges were not specific to our work: it remains uncertain by more than an order of magnitude how many interactions are to be expected in an organism. Related to this: what is the fraction of positives (PPIs) to negatives (proteins that do not interact) is in a living cell? Yet another crucial problem is that positives are much more reliable than negatives. For molecular biology in general it is much more accurate to state that an event happens than to rule out that it does not. All these issues magnify each other to render even the most careful performance estimates to become speculative approximations at best. Many authors use ROC-curves that relate the number of true positives (correctly predicted PPIs) to that of false positives (PPIs predicted but not observed). These plots depend heavily on the negatives in particular on the ratio of positives-to-negatives. Given that the truth for this number remains uncertain even within an order of magnitude, we decided to focus on curves that show precision-vs-recall, i.e. only values directly related to the observed PPIs. Although one of the axes still is strongly influenced by the assumption that 'not observed' means 'not interacting'. AUC, the area under the ROC-curve, is another simple and popular score for performance evaluation. Given the argument against ROC-curve, we might still vary this and compile an analogous area under the precision-recall curve. However, such a value would constitute another major problem: arguably, most users of prediction methods are most interested in the most reliable predictions. In other words, when predicting whether protein X interacts with any other human protein, the N-strongest predictions (for some N might be 1 for others 1000) matter more than all 20k scores against all 20k human proteins. But those 20k-N would exactly dominate the AUC-type performance measures.

## Database of predictions

Table 2 summarizes the results of the full interactome predictions. We only predicted PPIs which are dissimilar to proteins in our positive training set (Table 2, column NprotPred). Most proteins of the reference proteomes were dissimilar (Table 2: difference between columns Nprot, number of proteins, and NprotPred, number of predicted proteins). Overall, the eight new methods predicted PPIs for most of all possible pairs of proteins in an organism, i.e. at least 73% of all possible pairs (only exception: *Escherichia coli* and *Drosophila melanogaster*). Even after excluding all proteins previously reported in low-quality or redundant PPIs from the set of predicted PPIs, millions of predicted PPIs remained (Table 2, column NpredPPI novel). Due to our large mistake in the prediction of all PPIs proposed by the model at the default threshold, the ProfPPIdb resource only reported the most reliably predicted, non-redundant predictions (top $\sim$ 10% of all predicted PPIs) as novel PPIs (Table 2, column NpredPPI novel). For most of the 8 model organisms, this subset excludes most proteins in the

organism (Table 2, numbers in column NprotPred more than twice those in column Nprot ProfPPIdb). The exceptions were Plasmodium falciparum, Schizosaccharomyces pombe and Drosophila melanogaster for which PPI predictions remained for almost all proteins with predictions (Table 2, column NprotPred) after the application of these filters (Table 2, column Nprot ProfPPIdb). Hence, although our resource adds over one million newly predicted PPIs (sum over 8 rows of column NpredPPI ProfPPIdb in Table 2: 1,090,000 PPIs), many proteins in those organisms remain without annotation and without predictions.

## Conclusions

We applied the concept of profile-kernel SVMs for the prediction of physical protein-protein interactions (PPIs), i.e. we leverage information available for all proteins for which the sequence is known. The profile-kernel SVM-based methods appeared to achieve state-of-the-art performance for sequence-based PPI predictions. In fact, for most model organisms, the predictions were not inferior to those for human for which we had most experimental data and developed our initial approach. We put the most reliable predictions into a freely available database where users can access predictions for all proteins in the entire proteomes of eight different organisms (eukaryotes and prokaryotes, multi-cellular and single cellular, animals and plants, mammals, fly and worm).

## Supporting information

**S1 Appendix. Supplementary online materials.** Detailed results from training and testing the PPIs.
(PDF)

**S1 Fig. PPI test set for five organisms with $\geq$ 500 PPIs performed similar to human.** The y-axes give precision (number of PPIs correctly predicted at threshold), the x-axes the recall (number of experimental interactions predicted at that threshold). The precision-recall curves of each organism describe the performance of the test data set. The model for that was trained with two-thirds of the PPI data. Bars give the standard binomial deviation; negatives were sampled at a rate of 10:1 (ten negatives for one positive). The gray values compare the model organisms to the PPI prediction performance for human. *H. sapiens* (test) denotes the performance of the same method described here for human PPIs.
(EPS)

**S2 Fig. PPI test set for three organisms with $<$ 500 PPIs inferior to human.** The y-axes give precision (number of PPIs correctly predicted at threshold), the x-axes the recall (number of experimental interactions predicted at that threshold). The precision-recall curves of each organism describe the performance of the 5x5 cross validation of train data set. Bars give the standard deviation; negatives were sampled at a rate of 10:1 (ten negatives for one positive). The gray values compare the model organisms to the PPI prediction performance for human. *H. sapiens* (train) denotes the results of cross validation set.
(EPS)

**S3 Fig. Percentages of predictions as a function of STRING [39] (confidence) score.** The fractions of positive and negative predicted PPIs are each plotted against their STRING database confidence score. The plots show the plots for evidences for *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Escherichia coli* and *Schizosaccharomyces pombe*. For *Mus musculus* and *Rattus norvegicus*, no evidence was found.
(EPS)

**S4 Fig. Scores for experiment types taken from [29].** These scores were used for selection of high-quality (reliable) PPIs.
(TIF)

## Acknowledgments

Linh Tran is now at Imperial College London, but did this work earlier as a Technical University of Munich student. Thanks primarily to Tim Karl, but also to Guy Yachdav (all TUM) for invaluable help with hardware and software; to Inga Weise (TUM) for support with many other aspects of this work; to Jonas Reeb and Lothar Richter (both TUM) for help in supervising some of the work. This work was supported by a grant from the Alexander von Humboldt foundation through the German Ministry for Research and Education (BMBF: Bundesministerium fuer Bildung und Forschung). Last, not least, thanks to all those who deposit their experimental data in public databases, and to those who maintain these databases. In particular thanks to the BIND team headed by Gary Bader and Christopher Hogue, the DIP team around David Eisenberg, Ioannis Xenarios, and to the IntAct team co-ordinated by Henning Hermjakob.

## Author Contributions

**Conceptualization:** Tobias Hamp.

**Formal analysis:** Linh Tran.

**Methodology:** Linh Tran, Tobias Hamp.

**Software:** Linh Tran, Tobias Hamp.

**Supervision:** Tobias Hamp, Burkhard Rost.

**Visualization:** Linh Tran.

**Writing – original draft:** Linh Tran, Tobias Hamp.

**Writing – review & editing:** Linh Tran, Burkhard Rost.

## References

1. Rolland T, Taşan M, Charloteaux B, Pevzner SJ, Zhong Q, Sahni N, et al. A proteome-scale map of the human interactome network. Cell. 2014; 159(5):1212–1226. https://doi.org/10.1016/j.cell.2014.10.050 PMID: 25416956
2. Huang J, Niu C, Green CD, Yang L, Mei H, Han JDJ. Systematic prediction of pharmacodynamic drug-drug interactions through protein-protein-interaction network. PLoS Comput Biol. 2013; 9(3):e1002998. https://doi.org/10.1371/journal.pcbi.1002998 PMID: 23555229
3. Liu Y, Kim I, Zhao H. Protein interaction predictions from diverse sources. Drug discovery today. 2008; 13(9):409–416. https://doi.org/10.1016/j.drudis.2008.01.005 PMID: 18468558
4. Mosca R, Pons T, Céol A, Valencia A, Aloy P. Towards a detailed atlas of protein—protein interactions. Current opinion in structural biology. 2013; 23(6):929–940. https://doi.org/10.1016/j.sbi.2013.07.005 PMID: 23896349
5. Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, et al. A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science. 2003; 302(5644):449–453. https://doi.org/10.1126/science.1087361 PMID: 14564010
6. Li S, You ZH, Guo H, Luo X, Zhao ZQ. Inverse-free extreme learning machine with optimal information updating. IEEE transactions on cybernetics. 2016; 46(5):1229–1241. https://doi.org/10.1109/TCYB.2015.2434841 PMID: 26054082
7. You ZH, Yin Z, Han K, Huang DS, Zhou X. A semi-supervised learning approach to predict synthetic genetic interactions by combining functional and topological properties of functional gene network. Bmc Bioinformatics. 2010; 11(1):343. https://doi.org/10.1186/1471-2105-11-343 PMID: 20573270

8. Lan X, Bonneville R, Apostolos J, Wu W, Jin VX. W-ChIPeaks: a comprehensive web application tool for processing ChIP-chip and ChIP-seq data. Bioinformatics. 2010; 27(3):428–430. https://doi.org/10.1093/bioinformatics/btq669 PMID: 21138948
9. Ideker T, Ozier O, Schwikowski B, Siegel AF. Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics. 2002; 18(suppl 1):S233–S240. https://doi.org/10.1093/bioinformatics/18.suppl_1.S233 PMID: 12169552
10. Aytuna AS, Gursoy A, Keskin O. Prediction of protein—protein interactions by combining structure and sequence conservation in protein interfaces. Bioinformatics. 2005; 21(12):2850–2855. https://doi.org/10.1093/bioinformatics/bti443 PMID: 15855251
11. Sprinzak E, Margalit H. Correlated sequence-signatures as markers of protein-protein interaction. Journal of molecular biology. 2001; 311(4):681–692. https://doi.org/10.1006/jmbi.2001.4920 PMID: 11518523
12. Nye TM, Berzuini C, Gilks WR, Babu MM, Teichmann SA. Statistical analysis of domains in interacting protein pairs. Bioinformatics. 2004; 21(7):993–1001. https://doi.org/10.1093/bioinformatics/bti086 PMID: 15509600
13. Riley R, Lee C, Sabatti C, Eisenberg D. Inferring protein domain interactions from databases of interacting proteins. Genome biology. 2005; 6(10):R89. https://doi.org/10.1186/gb-2005-6-10-r89 PMID: 16207360
14. Chen Xw, Jeong JC. Sequence-based prediction of protein interaction sites with an integrative method. Bioinformatics. 2009; 25(5):585–591. https://doi.org/10.1093/bioinformatics/btp039 PMID: 19153136
15. Shen J, Zhang J, Luo X, Zhu W, Yu K, Chen K, et al. Predicting protein—protein interactions based only on sequences information. Proceedings of the National Academy of Sciences. 2007; 104(11):4337–4341. https://doi.org/10.1073/pnas.0607879104
16. Yu J, Guo M, Needham CJ, Huang Y, Cai L, Westhead DR. Simple sequence-based kernels do not predict protein—protein interactions. Bioinformatics. 2010; 26(20):2610–2614. https://doi.org/10.1093/bioinformatics/btq483 PMID: 20801913
17. Zahiri J, Yaghoubi O, Mohammad-Noori M, Ebrahimpour R, Masoudi-Nejad A. PPIevo: Protein—protein interaction prediction from PSSM based evolutionary information. Genomics. 2013; 102(4):237–242. https://doi.org/10.1016/j.ygeno.2013.05.006 PMID: 23747746
18. Liu B, Zhang D, Xu R, Xu J, Wang X, Chen Q, et al. Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection. Bioinformatics. 2013; 30(4):472–479. https://doi.org/10.1093/bioinformatics/btt709 PMID: 24318998
19. Smialowski P, Pagel P, Wong P, Brauner B, Dunger I, Fobo G, et al. The Negatome database: a reference set of non-interacting protein pairs. Nucleic acids research. 2009; 38(suppl_1):D540–D544. https://doi.org/10.1093/nar/gkp1026 PMID: 19920129
20. Hamp T, Rost B. More challenges for machine-learning protein interactions. Bioinformatics. 2015; 31(10):1521–1525. https://doi.org/10.1093/bioinformatics/btu857 PMID: 25586513
21. Park Y. Critical assessment of sequence-based protein-protein interaction prediction methods that do not require homologous protein sequences. BMC bioinformatics. 2009; 10(1):419. https://doi.org/10.1186/1471-2105-10-419 PMID: 20003442
22. Park Y, Marcotte EM. Flaws in evaluation schemes for pair-input computational predictions. Nature methods. 2012; 9(12):1134–1136. https://doi.org/10.1038/nmeth.2259 PMID: 23223166
23. Hamp T, Rost B. Evolutionary profiles improve protein—protein interaction prediction from sequence. Bioinformatics. 2015; 31(12):1945–1950. https://doi.org/10.1093/bioinformatics/btv077 PMID: 25657331
24. Pitre S, Hooshyar M, Schoenrock A, Samanfar B, Jessulat M, Green JR, et al. Short co-occurring polypeptide regions can predict global protein interaction maps. Scientific reports. 2012; 2. https://doi.org/10.1038/srep00239 PMID: 22355752
25. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. Nucleic acids research. 2006; 34(suppl 1):D535–D539. https://doi.org/10.1093/nar/gkj109 PMID: 16381927
26. Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM, Eisenberg D. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. Nucleic acids research. 2002; 30(1):303–305. https://doi.org/10.1093/nar/30.1.303 PMID: 11752321
27. Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, Derow C, et al. IntAct—open source resource for molecular interaction data. Nucleic acids research. 2007; 35(suppl 1):D561–D565. https://doi.org/10.1093/nar/gkl958 PMID: 17145710
28. Sonnhammer EL, Gabaldón T, da Silva AWS, Martin M, Robinson-Rechavi M, Boeckmann B, et al. Big data and other challenges in the quest for orthologs. Bioinformatics. 2014; p. btu492. https://doi.org/10.1093/bioinformatics/btu492

29. Schaefer MH, Fontaine JF, Vinayagam A, Porras P, Wanker EE, Andrade-Navarro MA. HIPPIE: Integrating protein interaction networks with experiment based quality scores. PloS one. 2012; 7(2): e31826. https://doi.org/10.1371/journal.pone.0031826 PMID: 22348130

30. Rost B. Twilight zone of protein sequence alignments. Protein engineering. 1999; 12(2):85–94. https://doi.org/10.1093/protein/12.2.85 PMID: 10195279

31. Guo Y, Yu L, Wen Z, Li M. Using support vector machine combined with auto covariance to predict protein—protein interactions from protein sequences. Nucleic acids research. 2008; 36(9):3025–3030. https://doi.org/10.1093/nar/gkn159 PMID: 18390576

32. Martin S, Roe D, Faulon JL. Predicting protein—protein interactions using signature products. Bioinformatics. 2005; 21(2):218–226. https://doi.org/10.1093/bioinformatics/bth483 PMID: 15319262

33. Kuang R, Ie E, Wang K, Wang K, Siddiqi M, Freund Y, et al. Profile-based string kernels for remote homology detection and motif extraction. Journal of bioinformatics and computational biology. 2005; 3(03):527–550. https://doi.org/10.1142/S021972000500120X PMID: 16108083

34. Witten IH, Frank E. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann; 2005.

35. Yachdav G, Kloppmann E, Kajan L, Hecht M, Goldberg T, Hamp T, et al. PredictProtein—an open resource for online prediction of protein structural and functional features. Nucleic acids research. 2014; p. gku366. https://doi.org/10.1093/nar/gku366

36. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research. 1997; 25 (17):3389–3402. https://doi.org/10.1093/nar/25.17.3389 PMID: 9254694

37. Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic acids research. 2000; 28(1):45–48. https://doi.org/10.1093/nar/28.1.45 PMID: 10592178

38. Berman RM, Cappiello A, Anand A, Oren DA, Heninger GR, Charney DS, et al. Antidepressant effects of ketamine in depressed patients. Biological psychiatry. 2000; 47(4):351–354. https://doi.org/10.1016/S0006-3223(99)00230-9 PMID: 10686270

39. Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, et al. The STRING database in 2017: quality-controlled protein—protein association networks, made broadly accessible. Nucleic acids research. 2016; p. gkw937.

40. Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, Cesareni G. MINT: a Molecular INTeraction database. FEBS letters. 2002; 513(1):135–140. https://doi.org/10.1016/S0014-5793(01)03293-8 PMID: 11911893

41. Calderone A, Castagnoli L, Cesareni G. Mentha: a resource for browsing integrated protein-interaction networks. Nature methods. 2013; 10(8):690. https://doi.org/10.1038/nmeth.2561 PMID: 23900247

42. Consortium U, et al. Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic acids research. 2011; p. gkr981.

43. Cooper GM, Hausman RE. The cell. vol. 85. Sinauer Associates Sunderland; 2000.

44. Matthews JM, Kowalski K, Liew CK, Sharpe BK, Fox AH, Crossley M, et al. A class of zinc fingers involved in protein—protein interactions. The FEBS Journal. 2000; 267(4):1030–1038.

45. Laity JH, Lee BM, Wright PE. Zinc finger proteins: new insights into structural and functional diversity. Current opinion in structural biology. 2001; 11(1):39–46. https://doi.org/10.1016/S0959-440X(00)00167-6 PMID: 11179890

46. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nature genetics. 2000; 25(1):25. https://doi.org/10.1038/75556 PMID: 10802651