Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Subject Section

Supplementary Material: OPA2Vec: combining formal and informal content of biomedical ontologies to improve similarity-based prediction

Fatima Zohra Smaili [1], Xin Gao [1,*] and Robert Hoehndorf [1,*]

Annotation Property	Label	Human	Yeast
$< http://purl.obolibrary.org/obo/IAO\_0000115 >$	definition	0.8215	0.8298
$< http://purl.obolibrary.org/obo/IAO\_0100001 >$	term replaced by	0.7611	0.7703
$< http://purl.org/dc/elements/1.1/creator >$	creator	0.7611	0.7701
$< http://purl.org/dc/elements/1.1/date >$	date	0.7619	0.7693
$< http://www.geneontology.org/formats/oboInOwl\#consider >$	consider	0.7614	0.7701
$< http://www.geneontology.org/formats/oboInOwl\#created_by >$	created_by	0.7697	0.7729
$< http://www.geneontology.org/formats/oboInOwl\#creation_date >$	creation_date	0.7494	0.7536
$< http://www.geneontology.org/formats/oboInOwl\#hasAlternativeId >$	has_alternative_id	0.7616	0.7703
$< http://www.geneontology.org/formats/oboInOwl\#hasBroadSynonym >$	has_broad_synonym	0.7683	0.7764
$< http://www.geneontology.org/formats/oboInOwl\#hasDbXref >$	database_cross_reference	0.7617	0.7695
$< http://www.geneontology.org/formats/oboInOwl\#hasExactSynonym >$	has_exact_synonym	0.7711	0.7792
$< http://www.geneontology.org/formats/oboInOwl\#hasNarrowSynonym >$	has_narrow_synonym	0.7731	0.7788
$< http://www.geneontology.org/formats/oboInOwl\#hasOBONamespace >$	has_obo_namespace	0.7587	0.7683
$< http://www.geneontology.org/formats/oboInOwl\#hasRelatedSynonym >$	has_related_synonym	0.7680	0.7772
$< http://www.geneontology.org/formats/oboInOwl\#id >$	id	0.7610	0.7700
$< http://www.geneontology.org/formats/oboInOwl\#SynonymTypeProperty >$	synonym_type_property	0.7609	0.7703
$owl:deprecated$		0.7621	0.7709
$rdfs:comment$	comment	0.7600	0.7692
$rdfs:label$	label	0.7973	0.8006
none	No annotation property	0.7614	0.7701
All	All annotation properties	0.8792	0.8971

Table 1. Contribution of each annotation property in GO ontology to PPI prediction. We list the AUC values for predicting interacting proteins in human and yeast.

Annotation Property	Human	Mouse
$< http://purl.obolibrary.org/obo/core\#provenance\_notes >$	0.7844	0.8377
$< http://purl.obolibrary.org/obo/core\#somite\_number >$	0.7837	0.8409
$< http://purl.obolibrary.org/obo/core\#tooth\_number >$	0.7836	0.8411
$< http://purl.obolibrary.org/obo/core\#vertebra\_number >$	0.7838	0.8433
$http://purl.obolibrary.org/obo/HP\_0040005$	0.7852	0.8446
$http://purl.obolibrary.org/obo/hp.owl\#layperson >$	0.7849	0.8440
$< http://purl.obolibrary.org/obo/hsapdv\#editor\_notes >$	0.7837	0.8435
$< http://purl.obolibrary.org/obo/IAO\_0000115 >$	0.8071	0.8704
$< http://purl.obolibrary.org/obo/IAO\_0000116 >$	0.7840	0.8431
$< http://purl.obolibrary.org/obo/IAO\_0000232 >$	0.7841	0.8428
$< http://purl.obolibrary.org/obo/IAO\_0000589 >$	0.7841	0.8431
$< http://purl.obolibrary.org/obo/IAO\_0100001 >$	0.7850	0.8452
$< http://purl.obolibrary.org/obo/namespace >$	0.7837	0.8433
$< http://purl.obolibrary.org/obo/synonymtype >$	0.7851	0.8423
$< http://purl.obolibrary.org/obo/uberon/core\#ABBREVIATION >$	0.7840	0.8430
$< http://purl.obolibrary.org/obo/uberon/core\#EXACT\_PREFERRED >$	0.7840	0.8430
$< http://purl.obolibrary.org/obo/uberon/core\#LATIN >$	0.7841	0.8430
$< http://purl.obolibrary.org/obo/uberon/core\#RLATED >$	0.7841	0.8431
$< http://purl.obolibrary.org/obo/UBPROP\_0000001 >$	0.7852	0.8425
$< http://purl.obolibrary.org/obo/UBPROP\_0000002 >$	0.7843	0.8426
$< http://purl.obolibrary.org/obo/UBPROP\_0000003 >$	0.7834	0.8430
$< http://purl.obolibrary.org/obo/UBPROP\_0000005 >$	0.7842	0.8430
$< http://purl.obolibrary.org/obo/UBPROP\_0000006 >$	0.7837	0.8429
$< http://purl.obolibrary.org/obo/UBPROP\_0000007 >$	0.7842	0.8433
$< http://purl.obolibrary.org/obo/UBPROP\_0000008 >$	0.7839	0.8430
$< http://purl.obolibrary.org/obo/UBPROP\_0000009 >$	0.7840	0.8427
$< http://purl.obolibrary.org/obo/UBPROP\_0000010 >$	0.7839	0.8432
$< http://purl.obolibrary.org/obo/UBPROP\_0000011 >$	0.7834	0.8424
$< http://purl.obolibrary.org/obo/UBPROP\_0000012 >$	0.7829	0.8430
$< http://purl.obolibrary.org/obo/UBPROP\_0000013 >$	0.7842	0.8434
$< http://purl.obolibrary.org/obo/UBPROP\_0000014 >$	0.7833	0.8429
$< http://purl.obolibrary.org/obo/UBPROP\_0000015 >$	0.7831	0.8428
$< http://purl.obolibrary.org/obo/UBPROP\_0000103 >$	0.7824	0.8424
$< http://purl.obolibrary.org/obo/UBPROP\_0000104 >$	0.7837	0.8430
$< http://purl.obolibrary.org/obo/UBPROP\_0000105 >$	0.7831	0.8424
$< http://purl.obolibrary.org/obo/UBPROP\_0000106 >$	0.7837	0.8422
$< http://purl.obolibrary.org/obo/UBPROP\_0000107 >$	0.7829	0.8423
$< http://purl.obolibrary.org/obo/UBPROP\_0000111 >$	0.7831	0.8424
$< http://purl.org/dc/elements/1.1/contributor >$	0.7840	0.7843
$< http://purl.org/dc/elements/1.1/date >$	0.7812	0.8331
$< http://purl.org/dc/elements/1.1/description >$	0.7841	0.8431
$< http://www.geneontology.org/formats/oboInOwl\#consider >$	0.7832	0.8445
$< http://www.geneontology.org/formats/oboInOwl\#created\_by >$	0.7864	0.8470
$< http://www.geneontology.org/formats/oboInOwl\#creation\_date >$	0.7822	0.8401
$< http://www.geneontology.org/formats/oboInOwl\#date\_retrieved >$	0.7803	0.8403
$< http://www.geneontology.org/formats/oboInOwl\#editor >$	0.7782	0.8428
$< http://www.geneontology.org/formats/oboInOwl\#editor\_note >$	0.7839	0.8426
$< http://www.geneontology.org/formats/oboInOwl\#external\_class >$	0.7844	0.8441
$< http://www.geneontology.org/formats/oboInOwl\#external\_class\_label >$	0.7840	0.8430
$< http://www.geneontology.org/formats/oboInOwl\#external\_ontology >$	0.7851	0.8438
$< http://www.geneontology.org/formats/oboInOwl\#hasAlternativeId >$	0.7879	0.8452
$< http://www.geneontology.org/formats/oboInOwl\#hasBroadSynonym >$	0.7869	0.8513
$< http://www.geneontology.org/formats/oboInOwl\#hasDbXref >$	0.7832	0.8409
$< http://www.geneontology.org/formats/oboInOwl\#hasExactSynonym >$	0.7856	0.8504
$< http://www.geneontology.org/formats/oboInOwl\#hasNarrowSynonym >$	0.7871	0.8519
$< http://www.geneontology.org/formats/oboInOwl\#hasOBONamespace >$	0.7837	0.8433
$< http://www.geneontology.org/formats/oboInOwl\#hasRelatedSynonym >$	0.7864	0.8514
$< http://www.geneontology.org/formats/oboInOwl\#hasScope >$	0.7840	0.8430
$< http://www.geneontology.org/formats/oboInOwl\#id >$	0.7833	0.8410
$< http://www.geneontology.org/formats/oboInOwl\#inSubset >$	0.7823	0.8372
$< http://www.geneontology.org/formats/oboInOwl\#is\_about >$	0.7840	0.8430
$< http://www.geneontology.org/formats/oboInOwl\#is\_anonymous >$	0.7840	0.8430
$< http://www.geneontology.org/formats/oboInOwl\#note >$	0.7834	0.8425
$< http://www.geneontology.org/formats/oboInOwl\#ontology >$	0.7846	0.8438
$< http://www.geneontology.org/formats/oboInOwl\#ontology\_class >$	0.7841	0.8431
$< http://www.geneontology.org/formats/oboInOwl\#reference >$	0.7843	0.8425
$< http://www.geneontology.org/formats/oboInOwl\#seeAlso >$	0.7831	0.8417
$< http://www.geneontology.org/formats/oboInOwl\#source >$	0.7848	0.8449
$< http://www.geneontology.org/formats/oboInOwl\#src >$	0.7840	0.8429
$< http://www.geneontology.org/formats/oboInOwl\#stage >$	0.7840	0.8430
$< http://www.geneontology.org/formats/oboInOwl\#status >$	0.7840	0.8430
$< http://www.geneontology.org/formats/oboInOwl\#taxon >$	0.7847	0.8415
$< http://www.geneontology.org/formats/oboInOwl\#url >$	0.7832	0.8428

Annotation Property	Human	Mouse
$< http://www.geneontology.org/formats/oboInOwl\#version >$	0.7840	0.8430
$< http://www.geneontology.org/formats/oboInOwl\#xref >$	0.7848	0.8442
$owl : deprecated$	0.7821	0.8411
$rdfs : comment$	0.7849	0.8433
$rdfs : label$	0.7988	0.8661
All	0.8411	0.8962
None	0.7841	0.8431

Table 2. Contribution of each annotation property in Phenomenet ontology to gene-disease association prediction, shown through the AUC value.

Parameter	Definition	Default value
	Choice of training algorithm	
	sg= 1 skip-gram	
sg	sg= 0 CBOW	1
$size$	Dimension of the obtained vectors	200
min_count	Words with frequency lower than this value will be ignored	1
$window$	Maximum distance between the current and the predicted word	5
$iter$	Number of iterations	5
$negative$	Whether negative sampling will be used and how many "noise words" would be drawn	5

Table 3. Parameters used for training the Word2Vec model.

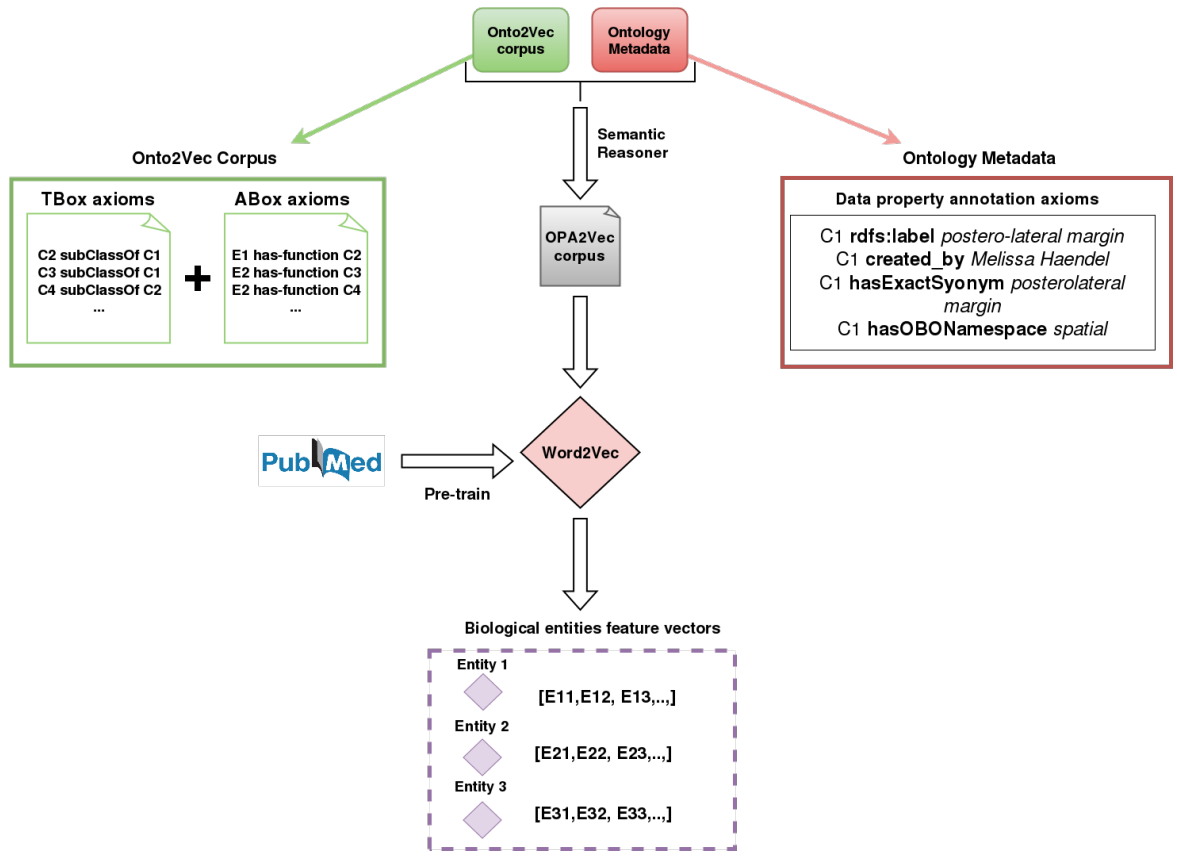


Fig. 1: The detailed workflow of the feature vector generation pipeline of OPA2Vec

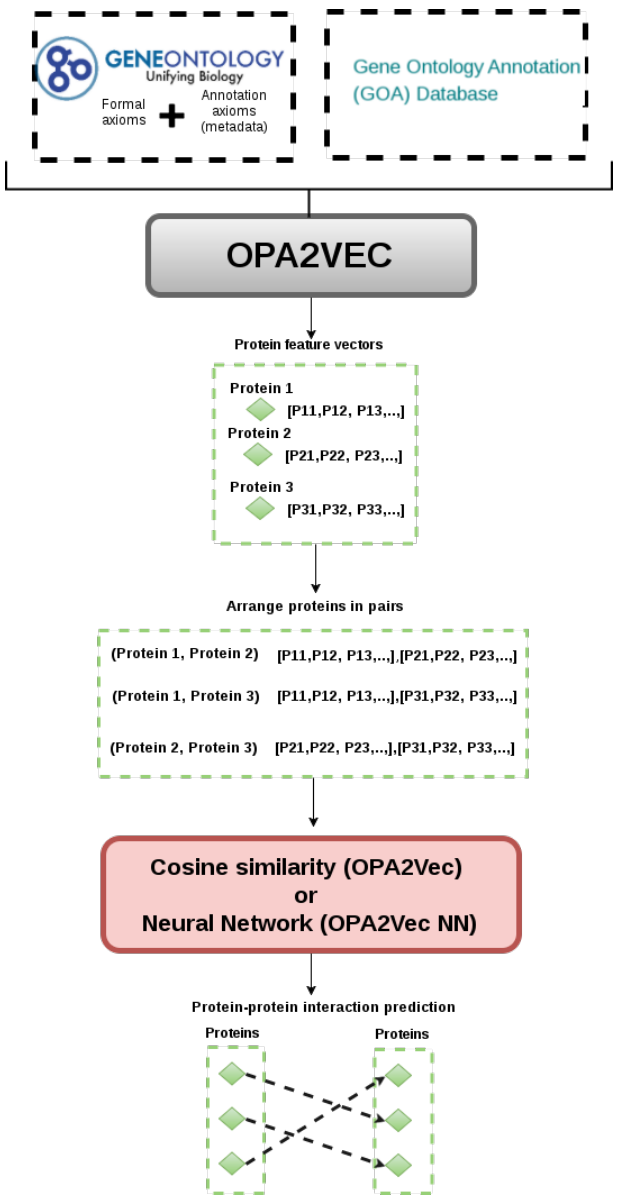
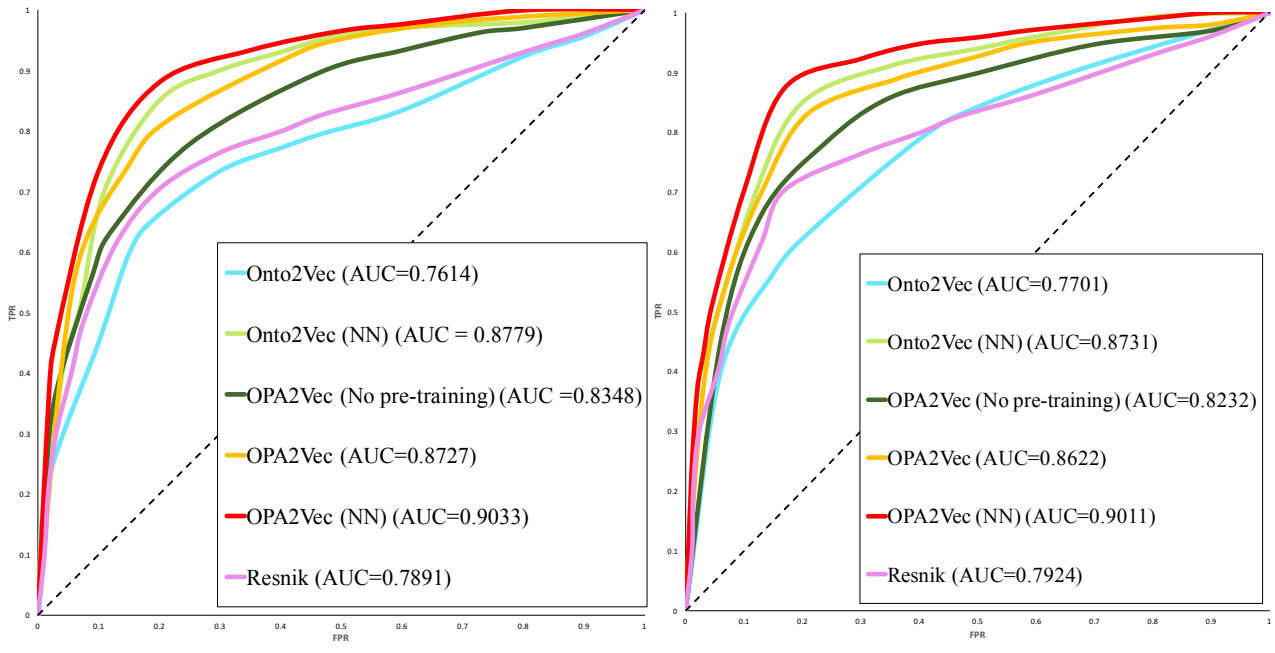


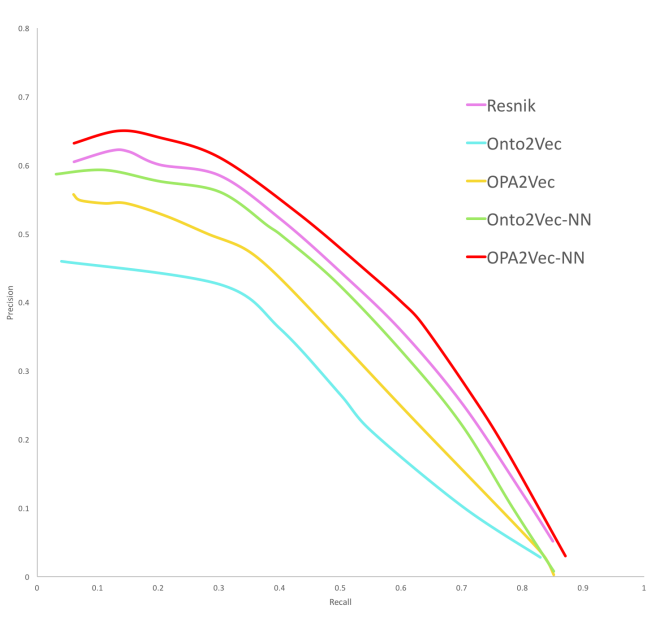
Fig. 2: Workflow for protein–protein interaction (PPI) prediction using OPA2Vec.



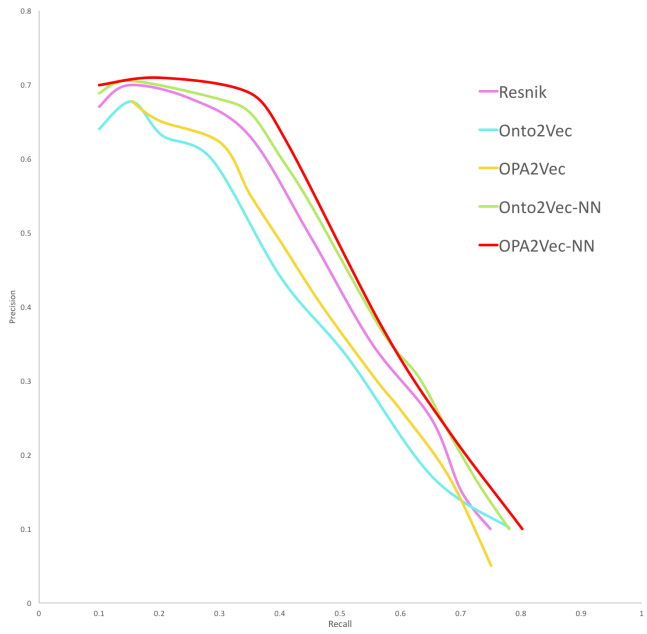
(a) Human

(b) Yeast

Fig. 3: ROC curves for each prediction method for PPI prediction accuracy for human and yeast.

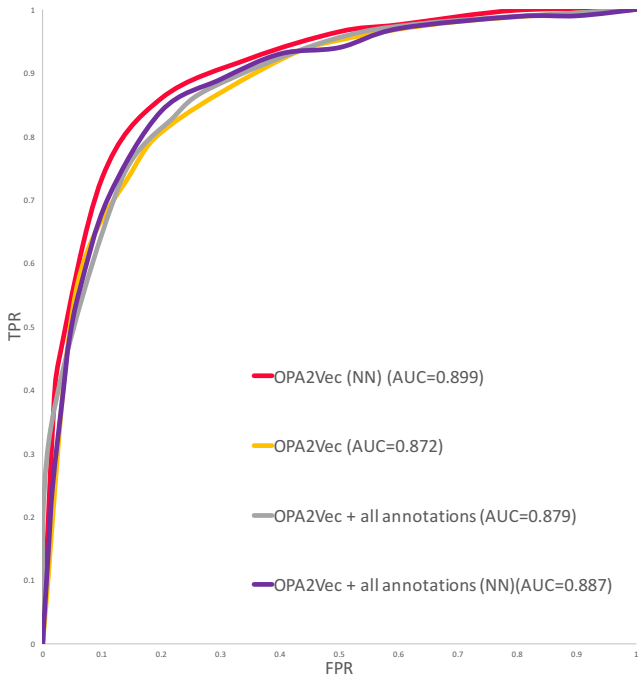


(a) Human

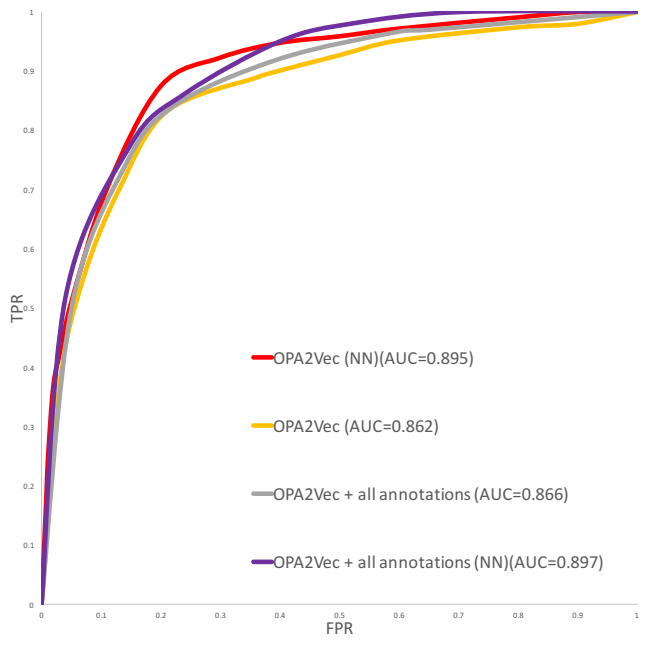


(b) Yeast

Fig. 4: Precision–recall curves for PPI prediction for human and yeast.

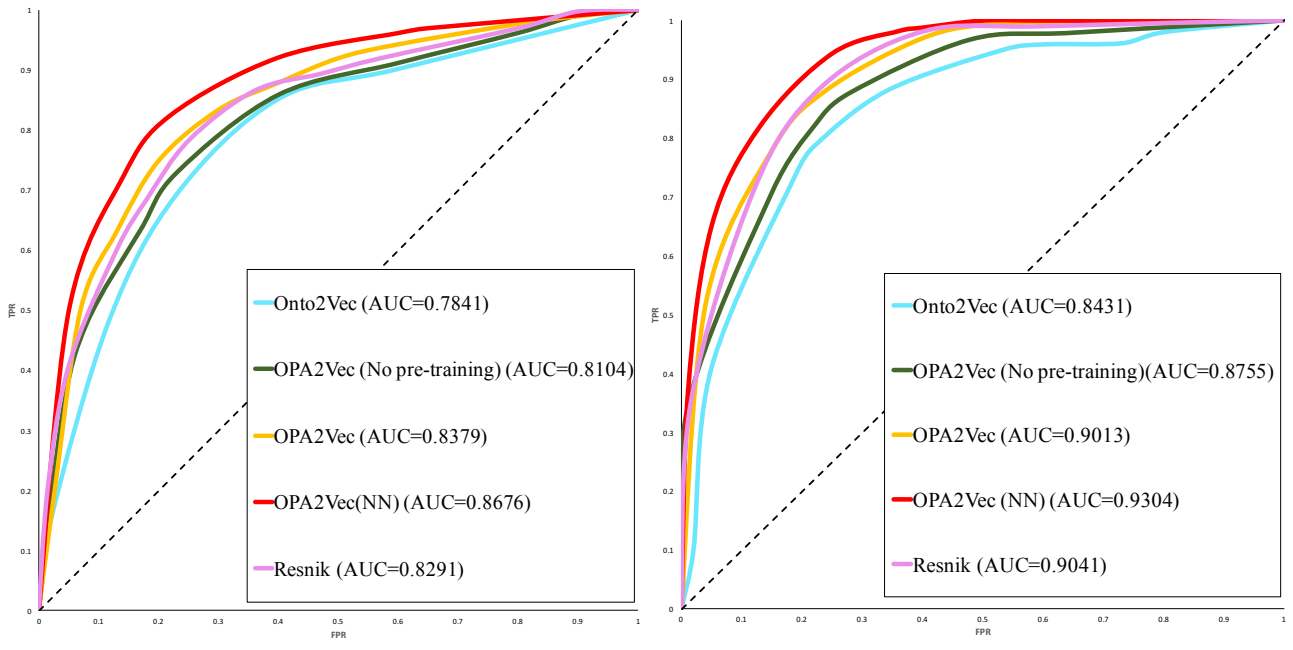


(a) Human



(b) Yeast

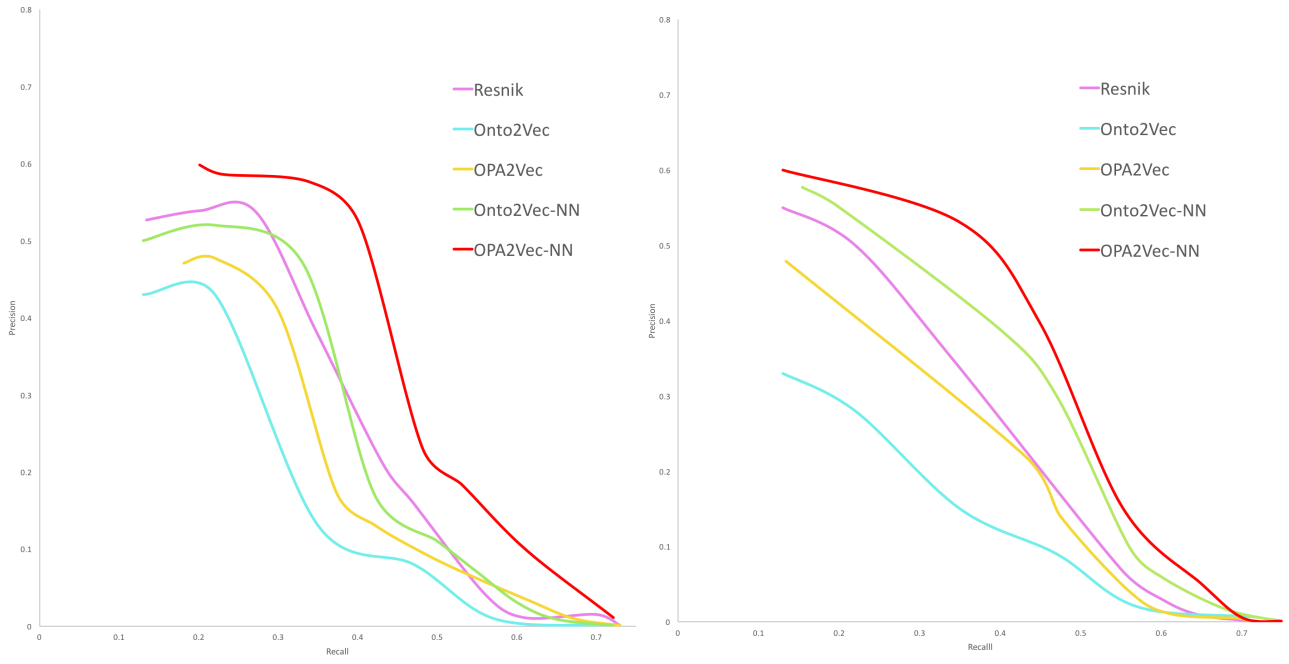
Fig. 5: Performance of OPA2Vec when trained on all annotation properties (OPA2Vec + all annotations/ OPA2Vec + all annotations (NN)) compared to the performance of OPA2Vec trained on a selected subset of properties (OPA2Vec / OPA2Vec (NN)) for PPI prediction for human and yeast.



(a) Human

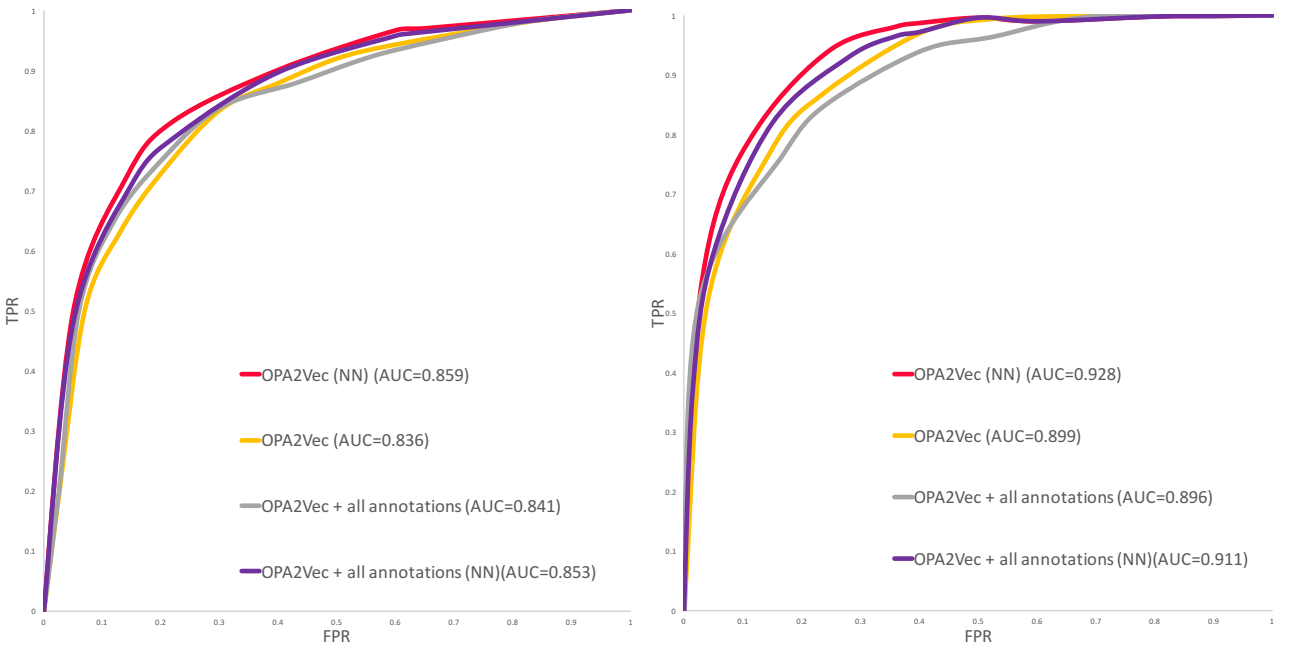
(b) Mouse

Fig. 6: ROC curves of each prediction method for gene–disease prediction for human and mouse.



(a) Human (b) Mouse

Fig. 7: Precision–recall curves for gene–disease association for human and mouse.



(a) Human (b) Mouse

Fig. 8: Performance of OPA2Vec using all annotation properties (OPA2Vec + all annotations/ OPA2Vec + all annotations (NN)) compared to the performance of OPA2Vec using a selected subset of properties (OPA2Vec / OPA2Vec (NN)) for gene-disease association prediction for human and mouse.

Using precision-recall curves as evaluation metric

In addition to using ROC curves to evaluate our method in different experiment, we also use precision–recall curves as an additional evaluation metric. Our precision and recall pairs are calculated for each rank and the resulting curves are shown in Figure 4 and Figure 7 for prediction of interacting proteins and gene–disease associations. The values of the area under precision-recall curve (AUPR) are available in Table 4 and Table 5 for predicting interacting proteins and gene–disease associations.

We also report the precision value at rank 1, 20 and 50 for OPA2Vec for PPI prediction for human and yeast and gene-disease association prediction for human and mouse in Table 6.

	Human	Yeast
$Resnik$	0.3422	0.3140
$Onto2Vec$	0.2300	0.2741
$OPA2Vec$	0.2753	0.2481
$Onto2Vec\_NN$	0.3375	0.3364
$OPA2Vec\_NN$	**0.3675**	**0.3476**

Table 4. Area under PR curve (AUPR) for PPI prediction for human and yeast.

	Human	Mouse
$Resnik$	0.1522	0.1461
$Onto2Vec$	0.0929	0.0772
$OPA2Vec$	0.0992	0.1214
$Onto2Vec\_NN$	0.1490	0.1687
$OPA2Vec\_NN$	**0.1724**	**0.2098**

Table 5. Area under PR curve (AUPR) for gene–disease association prediction for human and mouse.

	Protein–protein Interaction		**Gene–Disease Association**	
	Human	**Yeast**	**Human**	**Mouse**
Rank 1	0.11	0.12	0.23	0.15
Rank 10	0.27	0.33	0.35	0.29
Rank 20	0.53	0.47	0.41	0.42
Rank 50	0.59	0.56	0.58	0.50

Table 6. Recall at ranks 1, 10, 20, and 50 using OPA2Vec for prediction interacting proteins (on human and yeast) and gene–disease associations (human and mouse).

Protein–protein interaction prediction based on experimental interactions only

As an additional experiment, we predict protein–protein interactions using STRING's experimental interactions only (selected by chosing pairs with a confidence score greater than 700). We select positive pairs to be all pairs with confidence score greater than 700 in STRING, while our negatives are sub-sampled from the set of pairs not occurring in STRING in such a way that the cardinality of the positives and negatives is the same. Table 7 shows the AUC of the ROC curves resulting from this experiment.

	Human	Yeast
$Resnik$	0.7521	0.7701
$Onto2Vec$	0.7311	0.7474
$OPA2Vec$	0.7899	0.7961
$Onto2Vec\_NN$	0.8104	0.8211
$OPA2Vec\_NN$	**0.8316**	**0.8523**

Table 7. AUC values of PPI prediction for human and yeast.