Appendix I – Additional results

Part 1	a) Model parameters: Choose values for model parameters	Analysis yeast data Analysis human data Analysis chicken data	**Choose parameter values** k, GO-size filter, use of non-vsalidated data Number of replicates, k, GO-size filter, use of non-vsalidated data Pearson correlation threshold for co-expression network	
	b) Network characteristics: Choose PFP improvement strategy	Overview data species Impact of network characteristics Correlation between parameters Nature of the network (experiments)	**Choose method** epp/tpepp seems to be the most important network characteristic Decreasing the number of edges negative-negative or positives-negatives improves AUC Increasing the quality of the experiment does not seem to help	PU-learning could be a good strategy
Part 2	BMRF with chicken data		**Current state of predictions** Due to poor annotations, few GO terms can be predicted. The use of computational methods can improve this Although the average prediction is high, accuracy of prediction is never larger than 85%. BMRF may improve this	
Part3	PU-BMRF with chicken data			
Part4	Biological support of the approach			

Illustration 1: Diagram of results

Part 1- Impact of the different model parameters and network characteristics in the prediction performance using BMRF. Choose parameter values.

1a. **Iimpact of the model parameters in the prediction performance with BMRF**

Model parameter used in the original code (https://github.com/jwbargsten/bmrf)	
name parameter	**Description**
minGOsize	Minimum # of labels per GO term in the train set
minDFsize	Minimum # of labels per domain term in the train set
maxGOsize	Maximum # of labels per GO term in the train set. (i.e. 0.9 means 90% of the #labels in network)
maxDFsize	Maximum # of labels per domain term in the train set. (i.e. 0.9 means 90% of the #labels in network)
k	Number of folds in the BMRF cross-validation
Additional parameters considered	
network size	Subsets of netowrk used: **coexpresion** (#conexions)*
only EES	Whether only associations of category "Biological process" and with Experimnetral-evidence-scores are considered

Table 1: Description of model parameters

* The following network sizes were considered for yeast co-expession data (# of associations): 10,973; 26,879; 26,774; 64,519; 111,390; 242,504; 598,194

Note on data used:
We used yeast and human data to choose the value of the model parameters. Some analysis were carried on yeast data because it was easily accessible, whereas some other analysis were carried on human as we though that chicken data would not become available and it resembles more the situation in chickens. Finally, some analysis were carried on chicken data, after this was made available.

Note on parameter values used:
Unless specified, the analysis were carried with the following values:
-k:10
-20 replicates
-30 iterations for the Gibbs-sampling
-GO-size filter of 20 and 0.1, for minGOsize and maxGOsize, respectively
-Using domain information as well as non-validated associations.

1a(i): Analysis with yeast-coexpression data.

Using uyeast co-exopresion data we studied the impatc of the GO-size filter, the addition of non-validated data and the num,bver of k-folds on the prediction performance.

scenarios				data				
scenario name	Min GO-size	Max GO-size	only EES	Network size (#conn)	#unkown genes*	#assoc.	#GO-terms	AUC mean(sd) [median]
normal	20	0.1	F	598,174	655	132,249	1,104	0.779 (0.08) [0.778]
only validated associations			**T**		1,307	104,303	1,104	0.762 (0.083) [0.762]
default**		**0.9**			4	264,279	1,187	0.775 (0.08) [0.775]
more GO-terms	**10**	**0.9**			4	273,977	1,738	0.769 (0.1) [0.771]
Only large GO-terms	**30**	**0.07**			688	104,582	832	0.783 (0.075) [0.779]

Table 2: Impact of GO-szie in the data and the prediction performance, with yeast data

In blue, the parameter that were changed with respect to the "normal" scenario

*** default value in original BMRF code: https://github.com/jwbargsten/bmrf*

The number of protein was 5760 in all five scenarios.

From table 2, we conclude, that using non-experimental evidence scores helps to achieve higher performance (Increase in AUC from 0.762 to 0.779). Default value for maxGOsize was 0.9, however, for this thesis we are not intereted in predictions for the most general GO terms. For this reason, the "normal" scenarion has 0.1 as maxGOszie instead of 0.9

Then, the effect of the GO-size filter was investigated at the level of individual GO terms and we observed a slight increase in the prediction perfomance as more GO-terms were considered. Table 3 ilustrates this with 10 randomly chosen GO terms.

GO-term	#labels/#validated labels of the GO term	AUC filters 20,0.1	AUC filters 5,0.9
GO:0006417	100/144	0.738	0.747
GO:0031670	50/61	0.800	0.802
GO:0006414	40/65	0.752	0.766
GO:0051054	30/30	0.508	0.510
GO:0045931	25/31	0.642	0.641
GO:0007533	30/30	0.758	0.760
GO:0000209	36/23	0.863	0.869

Table 3: Impact of GO-size filter on individual GO-terms, with yeast data

In principle, we would expect that the prediction performance of one GO term would be independet of the other GO terms considered, however, we observed that this is not exactly the case. We observed an slight increase in AUC as the filter became less strict. This is due to the fact that when the filter is less strict, less genes will enter the category of unknown (see Appendix I - definitions).

Other anlaysis showed:

- The standard deviation across 5 runs of 20 replicates each was slighly lower for a GO-sizxe filter of (20,0.1) than for (4,0.9): 0.008 vs 0.01, respectivelly. This is logical since the standard deviation is larger for those GO terms with fewer genes and those GO terms were only considered in the analysis when the GO-size filter was 2,0.9 (less strict).

- Increasing the number of iterations from 10 to 20 did not impriove the prediction performance, even though the training samples was sligly larger. AUC was 0.75 vs 0.751, respectivelly. Increasing the number o fiterations, is not recomendable since the test-sample may become very small and this can caus problemsn in tehe computation of AUC.

1a(i): Analysis with human data.

Using human co-exopresion data, we studied the impact of the GO-size filter, the addition of non-validated data and domian infomration, the number of k-folds on the prediction performance and the number of replicates required for reproducible results.

Approach	AUC
AUC domains and nonValid (normal approach)	0.705
AUC domains but nonValid	0.701
AUC not domains and nonvalid	0.657

Table 4: Impact of domain information and non-validated gene-GO associations in the prediction performance with BMRF, using human data

AUC: area under the curve. Mean AUC of all GO terms that pass the filter.

Table 1 shows a significant increase in the prediction performance when the domain information was used. There was no increases, however, when the non-validated were added to the model.

Filter of GO terms for BMRF	#GO terms	average AUC
MinGOsize:9,maxGOsize=0.1	3328	0.701
MinGOsize:20,maxGOsize=0.1	1982	0.705
MinGOsize:20,maxGOsize=1	2069	0.704

Table 5: Ompact of the filters of "GO-term-size" in the prediction swith BMRF, with human data.

From table 2, we learn that the size of the GO terms does not seem to have an impact on the prediction performance. Results, however, may not apply in species for which the number of GO terms using different filters differ more.

We investigated the effect of the number of k-folds in the cross-validation.

AUC k:2	AUC k:5	AUC k:10	AUC k:20
0.668	0.695	0.705	0.705

Table 6: Impact of the number of folds in BMRF, with human data

From Table 3, we learn that the prediction performace increases with the number of iterations up to a point. This makes sense since the size of the trainning set increases for higher AUC and may be, insufficent for lowre values of k. Passed a ccertain value of k the AUC does not increase further.

The effect of the standrad deviation across runs of 10 or 20 replicates was used as an indicator of which num,bnerf replicates is required to achieve reproducble results. Here we considered as reprooduible, results with less below 0.002 standrad deviations in AUC

GO-term (#genes,#validated genes)	sd across runs (i.e. one run is the average of all replicates considered)		Difference
	10 replicates	20 reaplicates	
GO:0006417 (100/144)	0.007	0.002	0.005
GO:0006417 (100/144)	0.007	0.005	0.002
GO:0006414 (40/65)	0.013	0.007	0.006
GO:0051054 (30/30)	0.020	0.008	0.012
GO:0045931 (25/31)	0.024	0.010	0.014
GO:0007533 (30/30)	0.011	0.006	0.005
GO:0000209 (36/23)	0.012	0.014	-0.002

1a(iii): Analysis with chicken-coexpression data.

1b. **Differences in data between chickens, yeast and humans.**

Yeast coexpression
Network file: http://www.inetbio.org/yeastnet/downloadnetwork.php
GO file: http://www.yeastgenome.org/download-data/curation
Domains file: http://www.uniprot.org/docs/yeast

yeast_ppi
Network file: /mnt/scratch/dijk097/Fernando/BMRF-R/
GO file: http://www.yeastgenome.org/download-data
Domains file: http://www.uniprot.org/docs/yeast

Humans
Network file: http://mostafavilab.stat.ubc.ca/gnat/
GO file: http://www.geneontology.org/page/download-annotations
Domains file: http://www.uniprot.org/help/homo_sapiens

Chickens
Network file: http://coxpresdb.jp/download.shtml
GO file: http://www.geneontology.org/page/download-annotation
Domains file: http://www.uniprot.org/help/homo_sapiens

Table 4: Data sources for the different species

In all cases, co-expression data, except for "yeast ppi": yeast protein-protein iinteraction data

		total data	validated	validated after filter	Portion of data that is validated and passes the filter
#GO	yeast ppi	8,680	4,723	1,073	12.36
	yeast	8,680	4,723	1,104	12.72
	humans	19,549	10,271	1,982	10.14
	Chickens_07	9,247	877	9	0.10
	Chickens_035	16,205	2,350	142	0.88
#labels	yeast ppi	5,757	4,488	4,168	72.40
	yeast	5,757	4,488	4,453	77.35
	humans	8,574	5,582	5,535	64.56
	Chickens_07	2,152	53	53	2.46
	Chickens_035	9,038	300	296	3.28
#assoc	yeast ppi	474,389	227,420	98,192	20.70
	yeast	474,389	227,420	104,303	21.99
	humans	1,213,376	410,215	219,796	18.11
	Chickens_07	181,735	2,253	263	0.14
	Chickens_035	734,840	14,733	7,892	1.07

Table 5: Data available for the different species

ppi: protein-protein-interaction
#assoc: # of associations between GO terms and labels; Chickens_07 and Chickens_05: Network data for Chicken when the pearson correlation was 0.7 and 0.5, respectively.

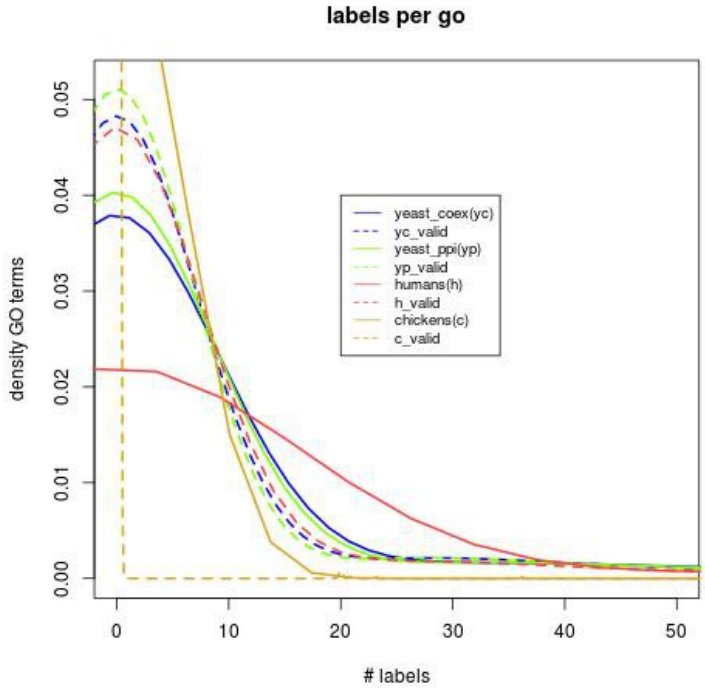
Filter of GO terms for BMRF	#GO terms humans	Chickens_0.35	yeast	yeast_ppi
MinGOsize:9,maxGOsize=0.1	3328	307	1772	1734
MinGOsize:20,maxGOsize=0.1	1982	138	1104	1057
MinGOsize:20,maxGOsize=1	2069	138	1187	1153

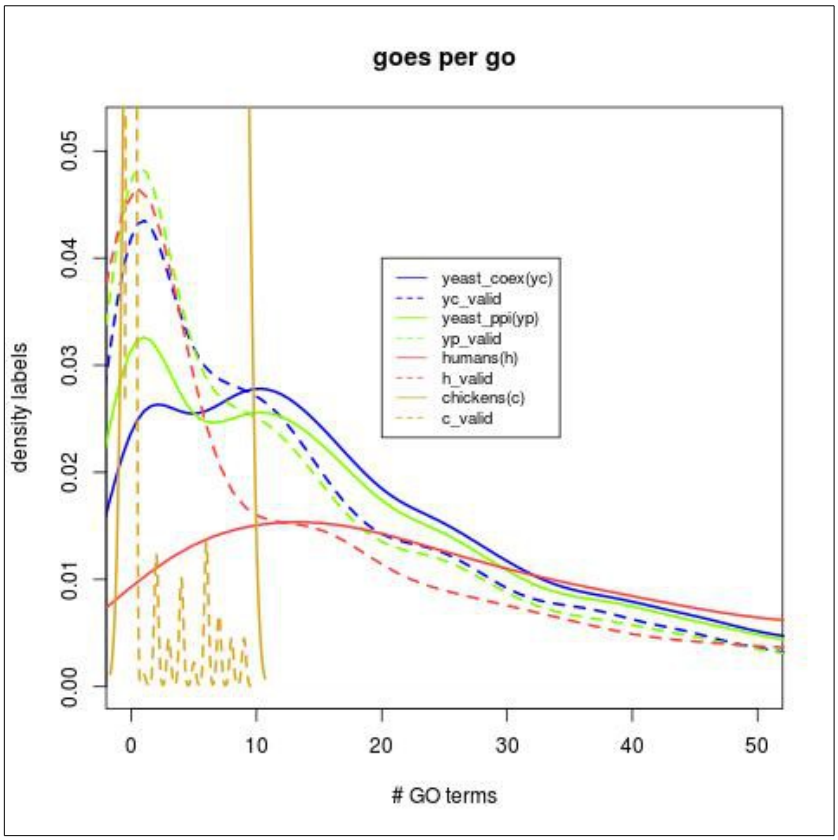
Table 7: Number of GO terms with differnet GO-term-size filters, for the differnet species.

The number of GO-terms after passing the filter is still low for chickens when a minimumGOsize of 9 was used.

We now look at the portion of labels per GO term, the portion of GO-terms per label and the number of edges per label

Illustration 2: Number of labels per GO-term in the different species.





From Illustration 1 and 2, we learn that the portion of labels per GO and number of GO terms per gene are much lower in chickens than in the other species and the differences becomes larger as we compare the validated data. Also we observe that for humans, since it is a more complex organism, the annotations are larger than for yeast but that the portion of data that is validated is considerably less for humans than for yeast.

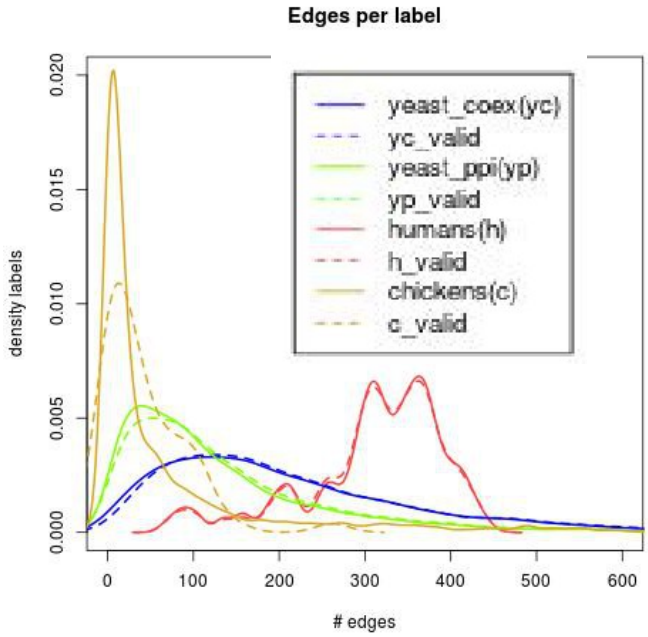


Illustration 3: Number of edges per label in the differnet species

The number of edges per gene is very similar for the validated genes and for the non-validated genes in all four cases: humans, chickens, yeast and yeast-ppi. In humans the number of edges per gene is considerably higher. We expect that this is the case, since humans is a more complex organism than yeast and a large portion of the data is available. In chickens the number of edges per gene is very lue doe to scarce annotation. As in Iluustrations 1 and 2, the coexpression data for yeast is more completed than the protein-protein interaction data.

We alsoo obsevred that, for yeast dat, the number of validated proteins decreases almost linearlt with the number of edges

Scenario	mean(#edges)	mean(#validated labels)
“stress”	4200.727	86.05
only validated associations	18845.56	94.477
“normal”	24007.55	94.477
focus on top	25333.17	98.704
more goes	31496.92	123.806
default	44175.61	173.613

Table 8: Relationship between the number of edes and thenumber of valitead genes, for yeast data.

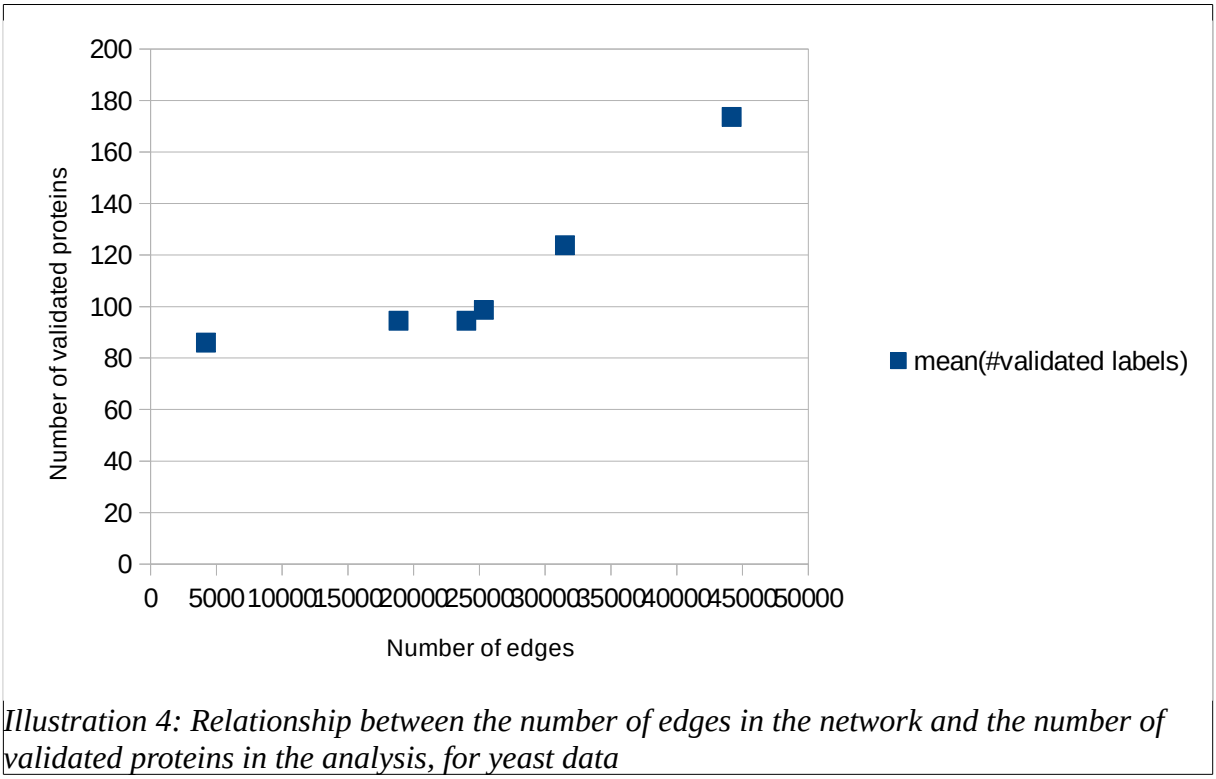


Illustration 4: Relationship between the number of edges in the network and the number of validated proteins in the analysis, for yeast data

1c. **Impact of network characteristics in the prediction performance using BMRF**

By comparing the characteristics of the network of the different species and the prediction p[erformance in each case, we can gain some understandig ion the network properties that are more relevant for protein function prediction via BMRF. Prediction performance was as follows:

	yeast ppi	yeast	humans	Chicken_07	Chicken_035
AUC	0.734	0.775	0.712	0.728	0.762

Table 9: Overall prediction performance for the differnet speciesusing BMRF.

Tables 6 and 7 summarize the main differences in the characteristics network of the different species.

	#te	**#epp (% te)**	**#epn (% te)**	**#enn (% te)**	**AUC**
yeast ppi	401,820	264,347 (65.79)	123,152 (30.65)	14,321 (3.56)	0.734
yeast	598,174	382,450 (63.94)	186,722 (31.22)	29,002 (4.85)	0.775
humans	1,548,622	481,792 (31.11)	754,276 (48.71)	312,554 (20.18)	0.712
Chicken_07	100,764	24 (0.02)	2,232 (2.22)	98,508 (97.76)	0.728
Chicken_035	2,094,870	576 (0.03)	51,610 (2.46)	2,042,684 (97.51)	0.762

Table 10: #edges and AUC

#te: total number of edges

epp: edges positive-positive. epn: edges positive-negative. enn: edges negative-negative

The total number of edges of the network may be of limited importance for PFP because it may be that most of these edges are linking genes that are not known to have the function, or genes that are known to have a given function with genes that are not known to have the same function. A more important network parameter therefore may be the epp (edges of positive-positive). These are edges that are linking genes that are known to have a common function. We compare the epp, epn and enn for the differnet species and we study a relationship between these parameters and the prediction performance (AUC-Area under the curve)

In table 2, we observe that the #epp may not be as important as the ratio epp/te, as AUC is higher for yeast (higher ratio epp/te) than for humans (higher epp). This makes sense since, in principle, epn and enn make more difficult the task of PFP. Note that enn may be inafct edges between positives and negatives. Results also suggest that the ratio epp/te may be related to the portion of associations that are validated, as both quantities are higher in yeast co-expression.

We then studied the degree of connections between the genes of a given GO terms, in the different species. One way to do this is by comparing the portion of epp with respect to the total possible number of epps (tpepp). Tpepp is a constant different for each GO term that refers to the total number of edges if all the genes associated with the GO term were interconnected. Tpepp is calcuated as: n*(n-1)/2, where n is the number of genes that are associated with the GO Term.

	epp/tpepp*1000	**epp/tpepp*1000 corrected by epp and standarized**	**AUC**
yeast ppi	47.88	-0.449	0.734
yeast	63.37	-0.449	0.775
humans	38.63	-0.449	0.712
Chickens_07	210.56	1.789	0.728
Chickens_035	28.15	-0.442	0.762

Table 11: epp by tpepp

epp: edges positive-positive; tpepp: total possible epp

AUC: area under the curve. Mean AUC of all GO terms that pass the filter considering only validated associations between the GO term and genes.

From table 3, we learn that with the exception of chickens data, there seem to be a favorable relation between epp/tpepp and AUC. For chickens_07, epp/tpepp is higher than expected. A possible explanation is that for chickens_07, the number of tpepp is very low and, since the pearson correlation is large, t is more likely that a large portion of the genes associated with the same GO term are interconnected. AUC, nevertheless, is not larger for chickens_07; the relationshupo between epp/tpepp asn AUC is not straignt. Thus we investigated the relationship between Epp/GO and other GO-specific network parameter swith AUC. We considered: the number of labels per GO, the number of GO terms per label, the number of edges per label and the number of epp per label.

The numbetr of labels per GO and Goes per gene is larger for humans. (Appendix I)

Note: The curved corresponding to yeast valid ppi (noit visible) falls just under the curve of humans-non-validated.

From illustration 2, we learn that the postion of epp per GO is much larger for yeast coexpression than in the other 3 cases. If we consider that the number of genes per labe, however was much

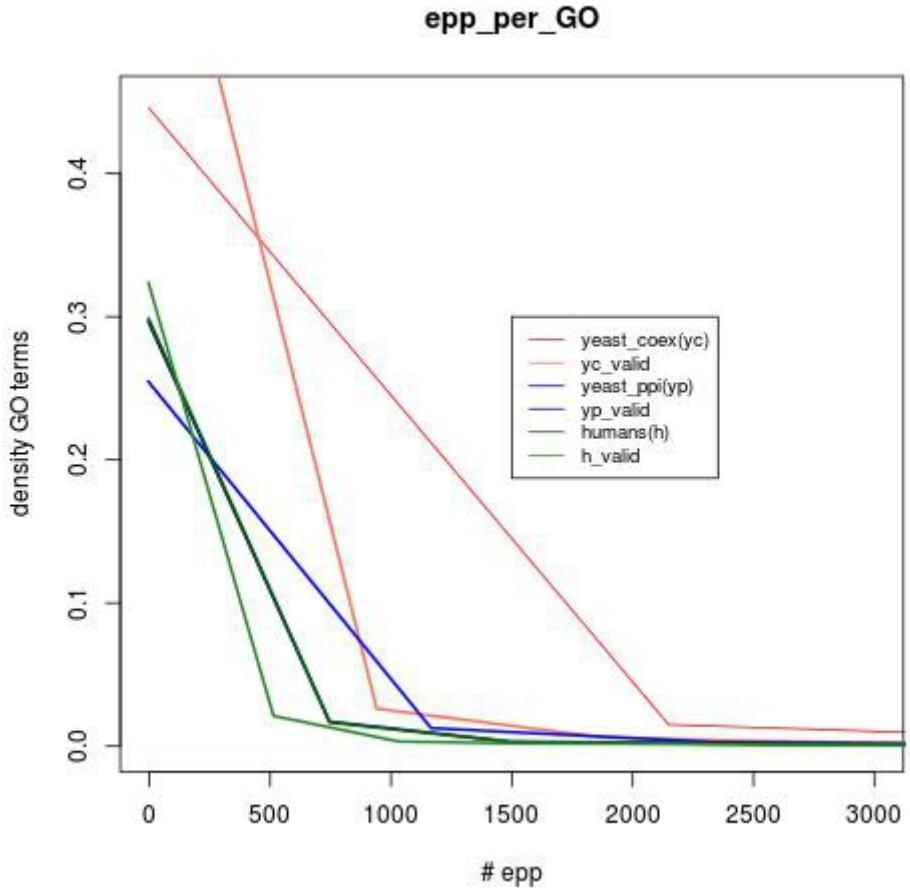
larger for humasn than for yeast (Innlustraions 1 in Appendix), we observe that the portion of data that is annotaed is much large for yeast than for humans.

and therefore we expect that prediction swill be better in this case. The differneces between the validated and non validtaed data is also larger for yeast coexpression. This may be due to the fact that under the name “non-validated” we include not onlt y the non-validated data frpom the Biological processGO category but also the validated nad non-validated data from Molecular function and Cell componets GO categories

	mean (sd)[median]				AUC
	labels/go	**go/labels**	**edges/label**	**Epp/GO**	
yeast_ppi	11.31 (46.29)	17.05 (22.67)	156.39 (179.3) [104]	960.12 (10844.65) [1]	0.734
yeast	12.01 (48.78)	18.12 (22.77)	213.7 (146.61) [178]	1311.15 (15209.25)[1]	0.775
humans	11.24(59.85)	25.63 (42.51)	310.36 (80.50) [323]	954.16 (11417.71) [0]	0.712
Chickens_07	0.28 (0.97)	0.12(0.85)	43.02(49.12)[24]	0.14642(1.199437)[0]	0.728
Chickens_035	24.55 (26.14)	0.87 (6.23)	811.29 (1097.3) [364]	6778.91 (110825.9) [0]	0.762

Table 12: Differences between the network data of the different species

Co-expression data for yeast has higher #labels per GO and epp/GO, whereas human co-expression data has higher in #GO/labels and #edges/label. Since we achieve a higher overall AUC for yeast, we can expect that the first two parameters are more related to AUC. Further, we expect that in order to achieve higher AUC (>0.75) data should have a large #labels/GO (~12) and ~1000 epp/GO. We observe that co-expression data for chickens_07 is still way far from this numbers. However, when we use chicken_035 data “#labels per GO”, #edges/label and “Epp/GO” increase to levels higher than in other species (#go/labels reminds much lower than for the other species). It is therefore not surprising that the mean AUC is higher for chickens_035 than for yeast_ppi and humans.



1b2. **Correlations**

However, we observed that,to some extend, the correlation between th enumber of egdges-AUC increases as the size of the network decreases. We did not observet any pattern on ho wthe size of the netwok alters the correlaion between AU Cnad the #labls, exept for the fact that the value of these correlation sis larger when the netwok is very small. the scenario “similar size network_1” the correlation betwee n#edges and AUC reached 0.376. We observed the differneces in AUC in 5 bins with differenet # of edges:

scenario name	#assoc.	CorrAUC_ #conn	CorrAUC_ #val_labels	CorrAUC_ #labels
very small network	8,862	0.219	-0.223	-0.125
similar_size network_1	32,336	0.356	0.072	0.124
similar_size network_2	58,358	0.376	0.223	0.272
only validated associations	104,303	0.133	0.05	0.05
Only large GO-terms	104,582	0.174	0.015	0.071
“stress” co-expression	110,682	0.255	0.113	0.161
“oxidiation” co-expression	111,480	0.252	0.113	0.162
normal*	132,249	0.124	0.014	0.053
default**	264,279	0.025	-0.008	0.002
more GO-terms	273,977	0.042	NA	NA

Table 14: Impact of the network size on the correlations AUC-#edges and AUC-#l(validated)able, using yeast data.

Also, in this scenario the corre;lation between AUC and the #lables is high (0.272). The difefrences in AUC bwteen 5 bins of GO terms with differnet number of preoteoins was also significant.

Bin of GO terms	AUC
1th/5	0.649
2th/5	0.656
3th/5	0.689
4th/5	0.696
5th/5	0.720

Table 15: Table 11: Differncenes on AUC between groups of GO terms with different # of edges in scenario "similar size network_1". The first fith refers to the 1/5th of the GO terms with a lowest numbver of edges, and so on

1e. **Impact of quality of data in the prediction performance.**

I) using yeast data

Portion of edges extracted from data	Mean AUC
0% (all network data used)	0.744
10%	0.738
30%	0.733
50%	0.738
90%	0.719
95%	0.719

Illustration 5: Impact of number of edges in th eprediction performance, using yeast data.

Removing random edges from the data dd not seem to affect much the prediction performance. We observed a large impact after removing 10% of the edges and after removing more than 50% of the

edges.
When we looked at inividual GO terms, we did not observe differences in the effect between differenyt GO terms.

GO-term	total_labels	valid_labels	portion of edges substracted					
			0% (all data used)	**10%**	**30%**	**50%**	**90%**	**95%**
GO:0042981	30	30	0.741	0.726	0.734	0.778	0.685	0.699
GO:0014068	30	30	0.48	0.479	0.514	0.495	0.504	0.493
GO:0045931	31	25	0.649	0.628	0.632	0.63	0.665	0.682
GO:0000209	36	23	0.862	0.872	0.773	0.837	0.857	0.837
GO:0006664	39	32	0.844	0.853	0.837	0.821	0.77	0.775
GO:0031670	61*	50*	0.811	0.789	0.796	0.789	0.796	0.79
GO:0036503	62*	49*	0.855	0.844	0.85	0.827	0.803	0.819
GO:0006414	65*	40	0.756	0.752	0.755	0.757	0.733	0.714
GO:0006417	144	100*	0.728	0.732	0.741	0.745	0.73	0.731
GO:0044270	195*	166*	0.714	0.703	0.701	0.705	0.644	0.649

Illustration 6: Impact of the extraction of edges in predicion performance for individuyaol GO terms

1f. Nature of the networks

Using yeast data, we investigated whether the nature of the co-expressionnetwork (cahracteristics of the experimnet) have any impact on th eprediction performance.

scenarios	**data**				**AUC mean(sd) [median]**	**AUC mean(sd) [median]**
scenario name	**Network size (#conn)**	**#unkown genes***	**#proteins.**	**#assoc.**		
"stress" co-expression	98,479	471	4,879	110,682	1,021	0.727 (0.089) [0.723]
"oxidiation" co-expression	64,167	499	4,923	111,480	1,022	0.72 (0.086) [0.714]
similar_size network_1	28,800	298	1,865	32,336	426	**0.684** (0.101) [0.677]
similar_size network_2	27,488	255	2,899	58,358	681	**0.682** (0.089) [0.687]
very small network	7,073	112	661	8,862	203	0.635 (0.113) [0.614]

Table 16: Impact of the nature of the network on th eprediction performance using yeast data.