*Gene Function Prediction*

# Exploiting Ontology Graph for Predicting Sparsely Annotated Gene Function

Sheng Wang[1,†], Hyunghoon Cho[2,†], ChengXiang Zhai[1], Bonnie Berger[2,3]and Jian Peng[1,*]

[1] Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL, USA.

[2] Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA, USA

[3] Department of Mathematics, MIT, Cambridge, MA, USA

## SUPPLEMENTARY INFORMATION

## 1 Two cluster structures discussed in Section 3.3

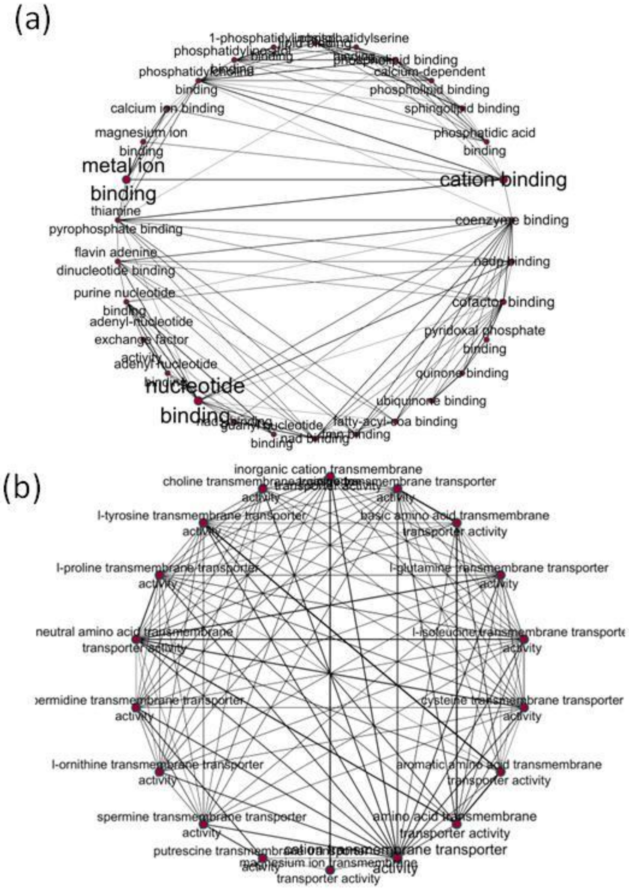
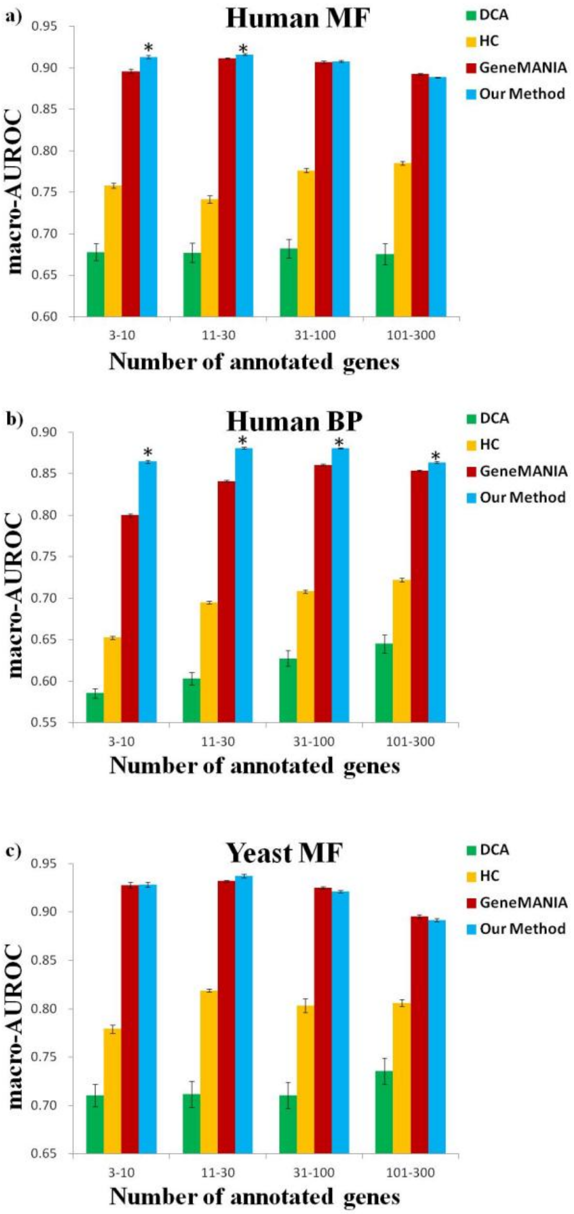

(a)
1-phosphatidylinositol
phosphatidylserine
lipid binding
binding
phosphatidylinositol binding
phospholipid binding
phosphatidylcholine binding
calcium-dependent phospholipid binding
calcium ion binding
sphingolipid binding
magnesium ion binding
phosphatidic acid binding
metal ion binding
cation binding
thiamine pyrophosphate binding
coenzyme binding
flavin adenine dinucleotide binding
nadp binding
purine nucleotide binding
cofactor binding
adenyl-nucleotide exchange factor activity
pyridoxal phosphate binding
adenyl nucleotide binding
quinone binding
nucleotide binding
ubiquinone binding
nad binding
guanyl nucleotide binding
fatty-acyl-coa binding
fmn binding
nad binding
(b)
inorganic cation transmembrane transporter activity
choline transmembrane transporter activity
transmembrane transporter activity
l-tyrosine transmembrane transporter activity
basic amino acid transmembrane transporter activity
l-proline transmembrane transporter activity
l-glutamine transmembrane transporter activity
neutral amino acid transmembrane transporter activity
l-isoleucine transmembrane transporte activity
ɔermidine transmembrane transporter activity
cysteine transmembrane transporter activity
l-ornithine transmembrane transporter activity
aromatic amino acid transmembrane transporter activity
spermine transmembrane transporter activity
amino acid transmembrane transporter activity
putrescine transmembrane transporter activity
cation transmembrane transporter activity
magnesium ion transmembrane transporter activity

**Fig. S1.** (a) The cluster structure of several binding functions. (b) The cluster structure of GO labels related to transmembrane transporting.

## 2 Comparison of performance in terms of macro-AUROC



[†]These authors equally contribute to this work.

[*]To whom correspondence should be addressed. *jianpeng@illinois.edu*

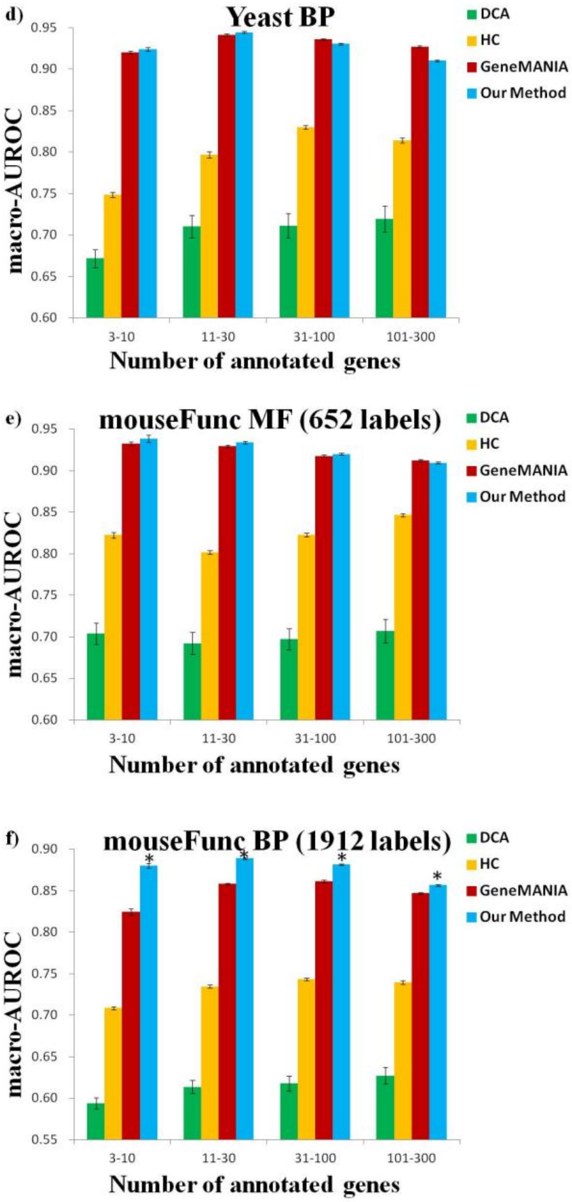


**Fig. S2.** Comparison of our approach with other methods in terms of macro-AUROC. * indicates that our approach is statistically significant in comparison with GeneMANIA. Performance is evaluated for different subsets of GO labels with varying sparsity levels as shown on the x-axis.

## 3 Comparison of number of dimensions discussed in Section 3.7

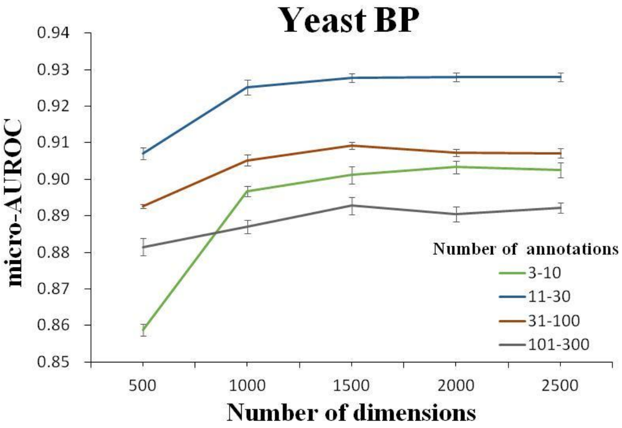
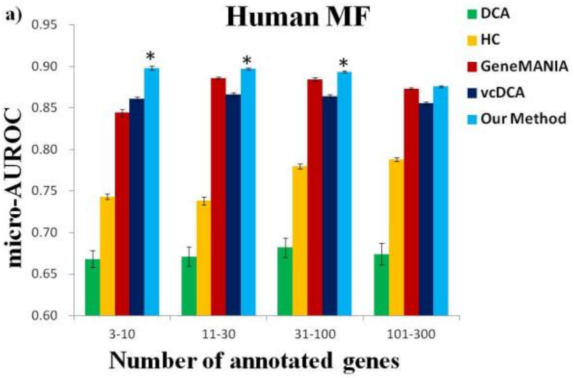


**Fig. S3.** Comparison of number of dimensions in terms of micro-AUROC of biological process in yeast.

## 4 Performance of clustering based on learned label vectors

Besides clustering GO labels based on the sparseness, we explored to cluster the GO labels based on learned label vectors. We denoted this method as vcDCA. We show the performance of this method in **Fig. S4** and **Fig. S5.**



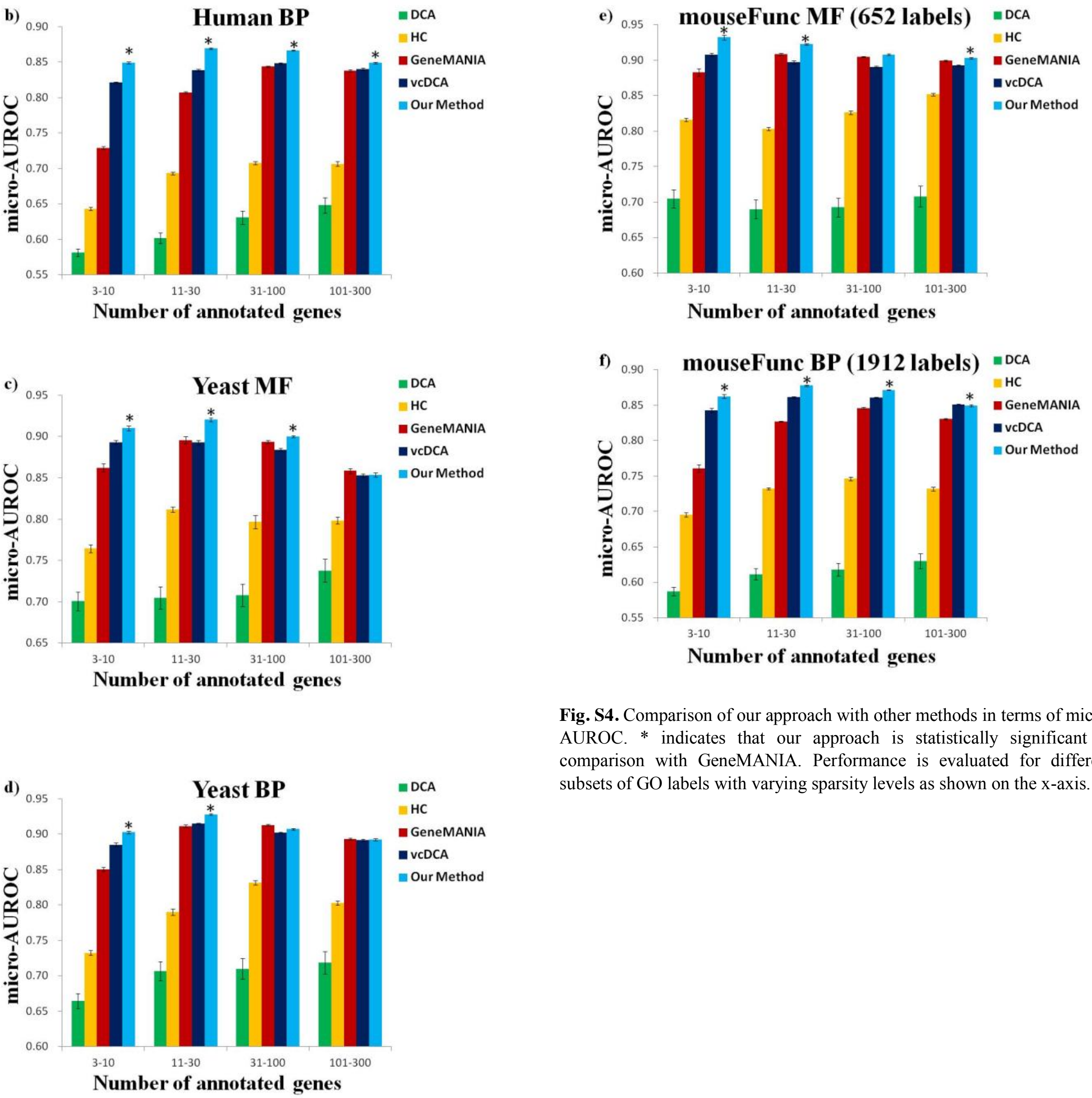
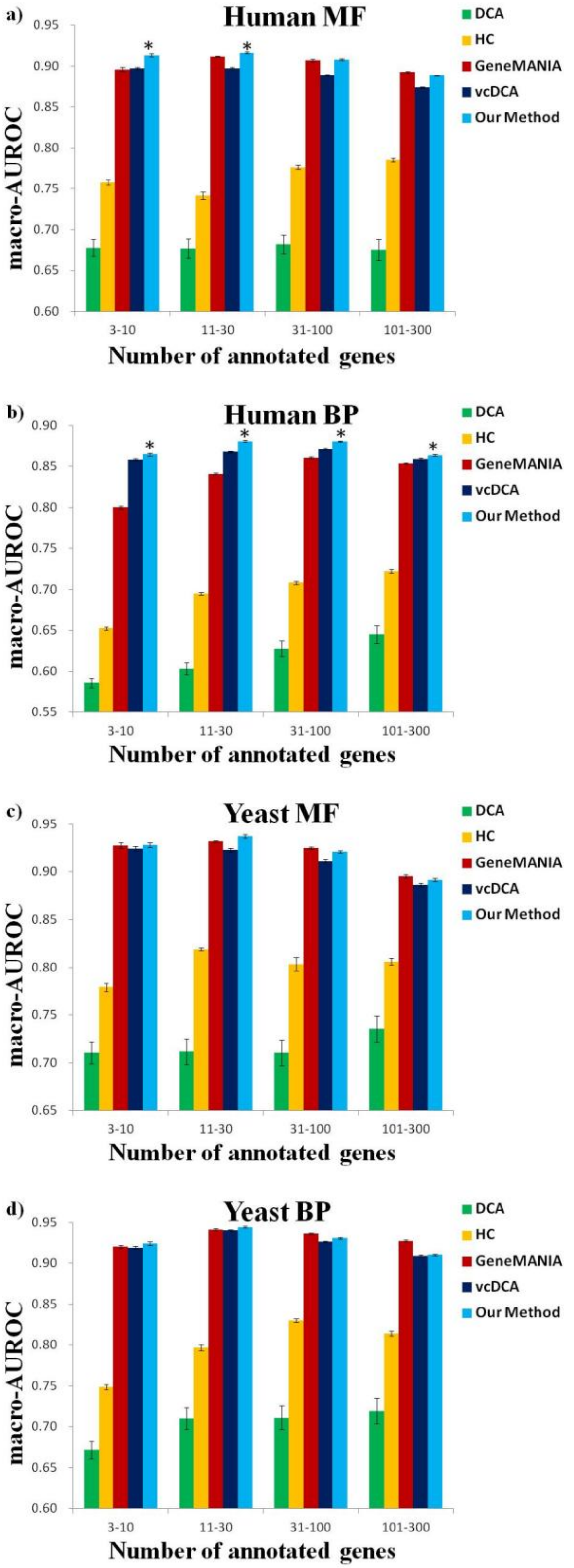


**Fig. S4.** Comparison of our approach with other methods in terms of micro-AUROC. * indicates that our approach is statistically significant in comparison with GeneMANIA. Performance is evaluated for different subsets of GO labels with varying sparsity levels as shown on the x-axis.



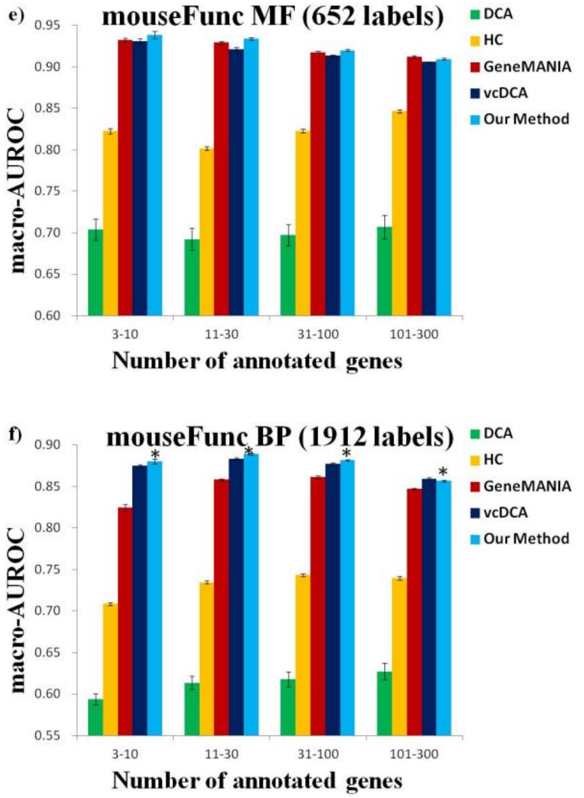


**Fig. S5.** Comparison of our approach with other methods in terms of macro-AUROC. * indicates that our approach is statistically significant in comparison with GeneMANIA. Performance is evaluated for different subsets of GO labels with varying sparsity levels as shown on the x-axis.

## 5 Comparison of performance in terms of AP@10 and macro-APRUC

We show the comparison of performance in terms of AP@10 and macro-APRUC in **Table S1** and **Table S2.** In human, our method achieved 0.1849 AUPRC on BP labels with 101-300 annotations, which is higher than 0.1764 AUPRC for GeneMANIA. In mouse, our method achieved 0.4238 AP@10 on MF labels with 31-100 annotations, which is higher than 0.3706 AP@10 for GeneMANIA.

**Table S1.** Comparison of our approach with other methods in terms of AP@10. * indicates that our approach is statistically significant in comparison with GeneMANIA.

| | #annotated genes | HC | GeneMANIA | clusDCA |
|---|---|---|---|---|
| Human MF | 3-10 | 0.1471 | 0.2123 | **0.2500** * |
| | 11-30 | 0.1810 | 0.2338 | **0.2770** * |
| | 31-100 | 0.2596 | 0.3502 | **0.3794** * |
| | 101-300 | 0.3139 | 0.4883 | **0.5541** * |
| Human BP | 3-10 | 0.0566 | 0.0975 | **0.1224** * |
| | 11-30 | 0.0769 | 0.1309 | **0.1448** * |
| | 31-100 | 0.1305 | 0.2628 | **0.2845** * |
| | 101-300 | 0.1802 | 0.4197 | **0.4938** * |

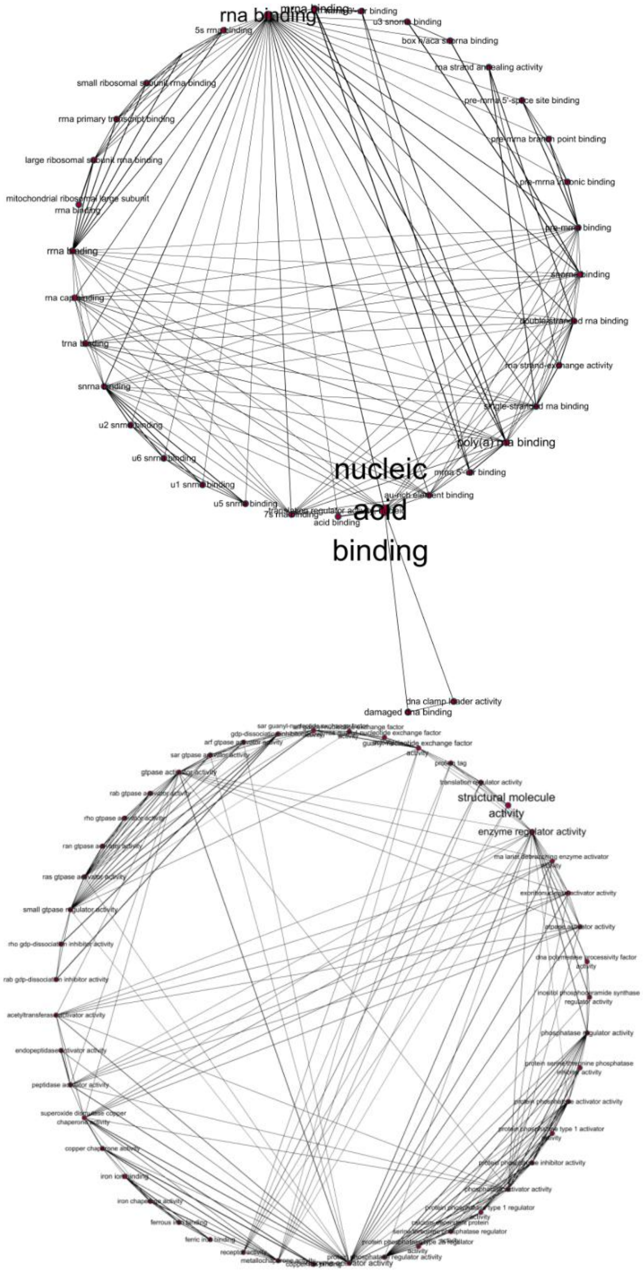
| | | | | |
|---|---|---|---|---|
| Yeast MF | 3-10 | 0.1851 | 0.3060 | **0.4152 *** |
| | 11-30 | 0.2320 | 0.3499 | **0.3953 *** |
| | 31-100 | 0.2921 | 0.5325 | **0.5559 *** |
| | 101-300 | 0.3197 | 0.6160 | **0.7109 *** |
| Yeast BP | 3-10 | 0.1658 | 0.2639 | **0.3453 *** |
| | 11-30 | 0.2265 | 0.3482 | **0.3987 *** |
| | 31-100 | 0.3434 | **0.5424** | 0.5292 |
| | 101-300 | 0.3875 | 0.7225 | **0.7249** |
| mouseFunc MF | 3-10 | 0.0961 | 0.2254 | **0.3514 *** |
| | 11-30 | 0.1978 | 0.2931 | **0.3450 *** |
| | 31-100 | 0.2877 | 0.3706 | **0.4238 *** |
| | 101-300 | 0.3250 | 0.5278 | **0.6627 *** |
| mouseFunc BP | 3-10 | 0.0559 | 0.1010 | **0.1523 *** |
| | 11-30 | 0.1020 | 0.1380 | **0.1806 *** |
| | 31-100 | 0.1365 | 0.2359 | **0.3019 *** |
| | 101-300 | 0.1748 | 0.3437 | **0.4268 *** |

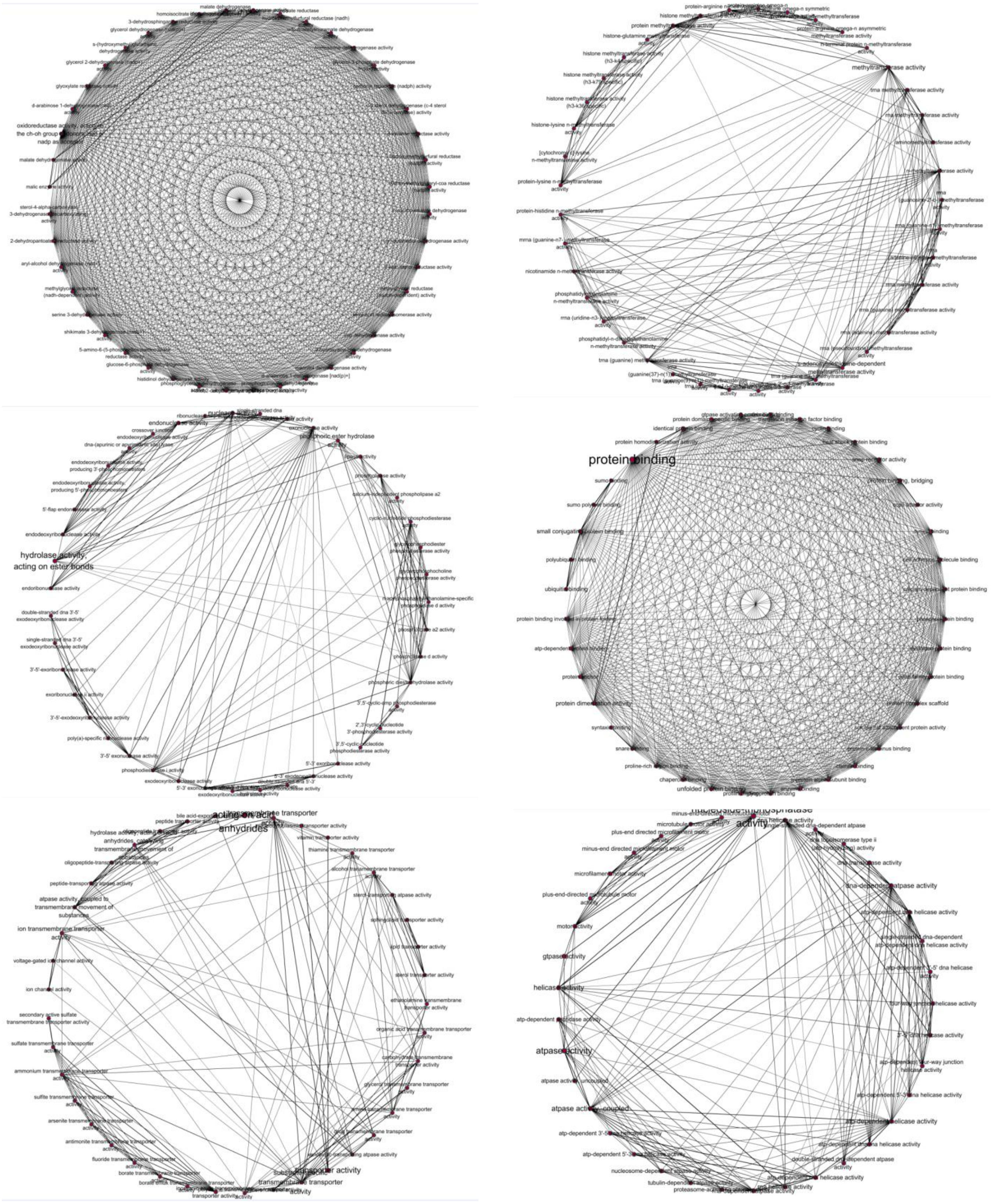
**Table S2.** Comparison of our approach with other methods in terms of macro-AUPRC. * indicates that our approach is statistically significant in comparison with GeneMANIA.

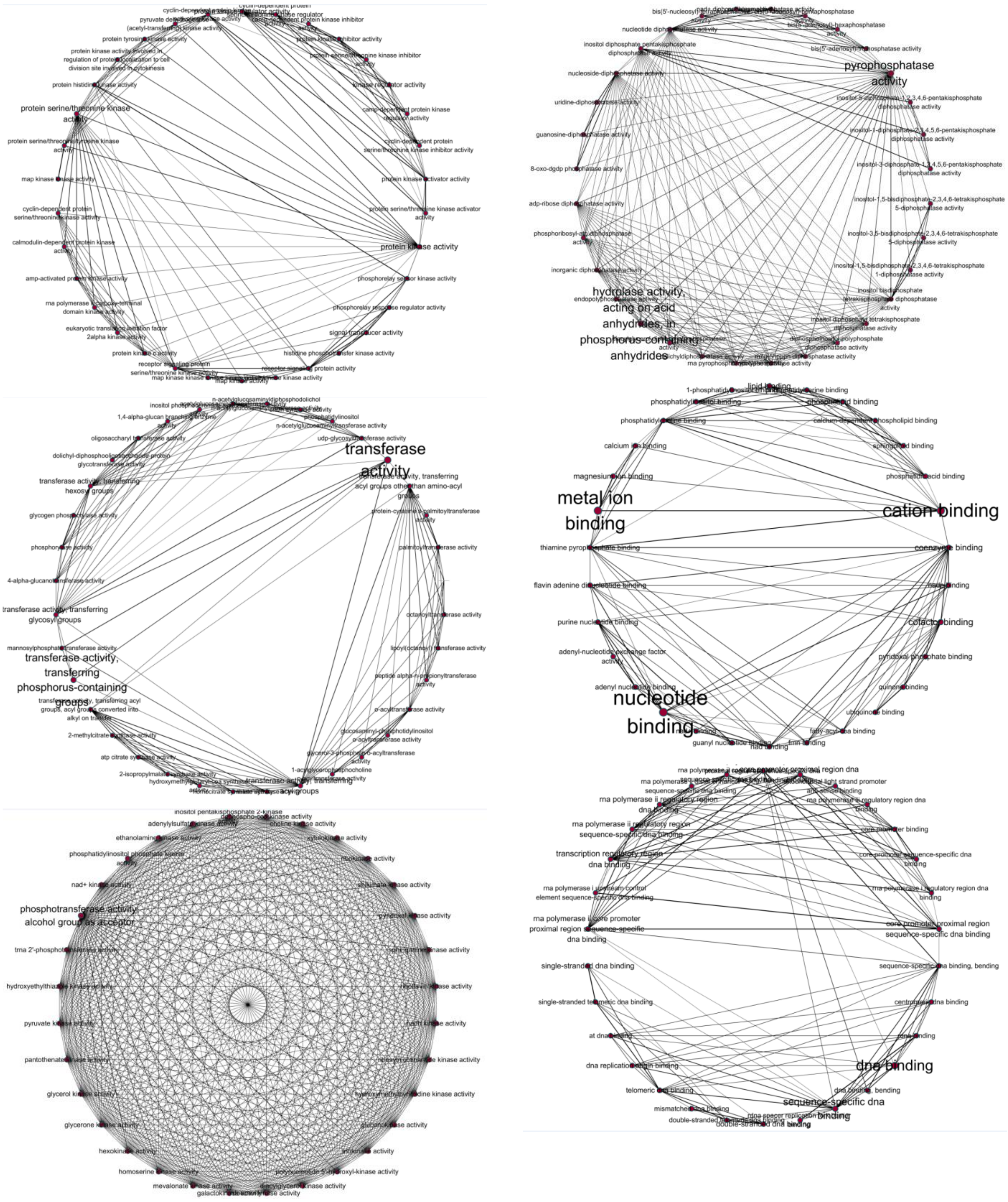
| | #annotated genes | HC | GeneMANIA | clusDCA |
|---|---|---|---|---|
| Human MF | 3-10 | 0.0153 | 0.0755 | **0.0818 *** |
| | 11-30 | 0.0486 | 0.1344 | **0.1574 *** |
| | 31-100 | 0.0677 | **0.1670** | 0.1625 |
| | 101-300 | 0.0637 | **0.2377** | 0.2151 |
| Human BP | 3-10 | 0.0098 | 0.0368 | **0.0429 *** |
| | 11-30 | 0.0282 | 0.0673 | **0.0703 *** |
| | 31-100 | 0.0384 | 0.1201 | **0.1228** |
| | 101-300 | 0.0475 | 0.1764 | **0.1849 *** |
| Yeast MF | 3-10 | 0.0190 | 0.1075 | **0.1360 *** |
| | 11-30 | 0.0776 | 0.1985 | **0.2334 *** |
| | 31-100 | 0.0808 | **0.2751** | 0.2686 |
| | 101-300 | 0.0844 | 0.3348 | **0.3445** |
| Yeast BP | 3-10 | 0.0299 | 0.0954 | **0.1098 *** |
| | 11-30 | 0.0857 | 0.1949 | **0.2217 *** |
| | 31-100 | 0.1155 | **0.2939** | 0.2737 |
| | 101-300 | 0.1195 | **0.4321** | 0.3926 |
| mouseFunc MF | 3-10 | 0.0062 | 0.0965 | **0.1367 *** |
| | 11-30 | 0.0451 | 0.1719 | **0.1950 *** |
| | 31-100 | 0.0681 | 0.1866 | **0.1991 *** |
| | 101-300 | 0.0651 | **0.2875** | 0.2814 |
| mouseFunc BP | 3-10 | 0.0074 | 0.0389 | **0.0516 *** |
| | 11-30 | 0.0361 | 0.0680 | **0.0825 *** |
| | 31-100 | 0.0394 | 0.0985 | **0.1088 *** |
| | 101-300 | 0.0474 | 0.1350 | **0.1367** |

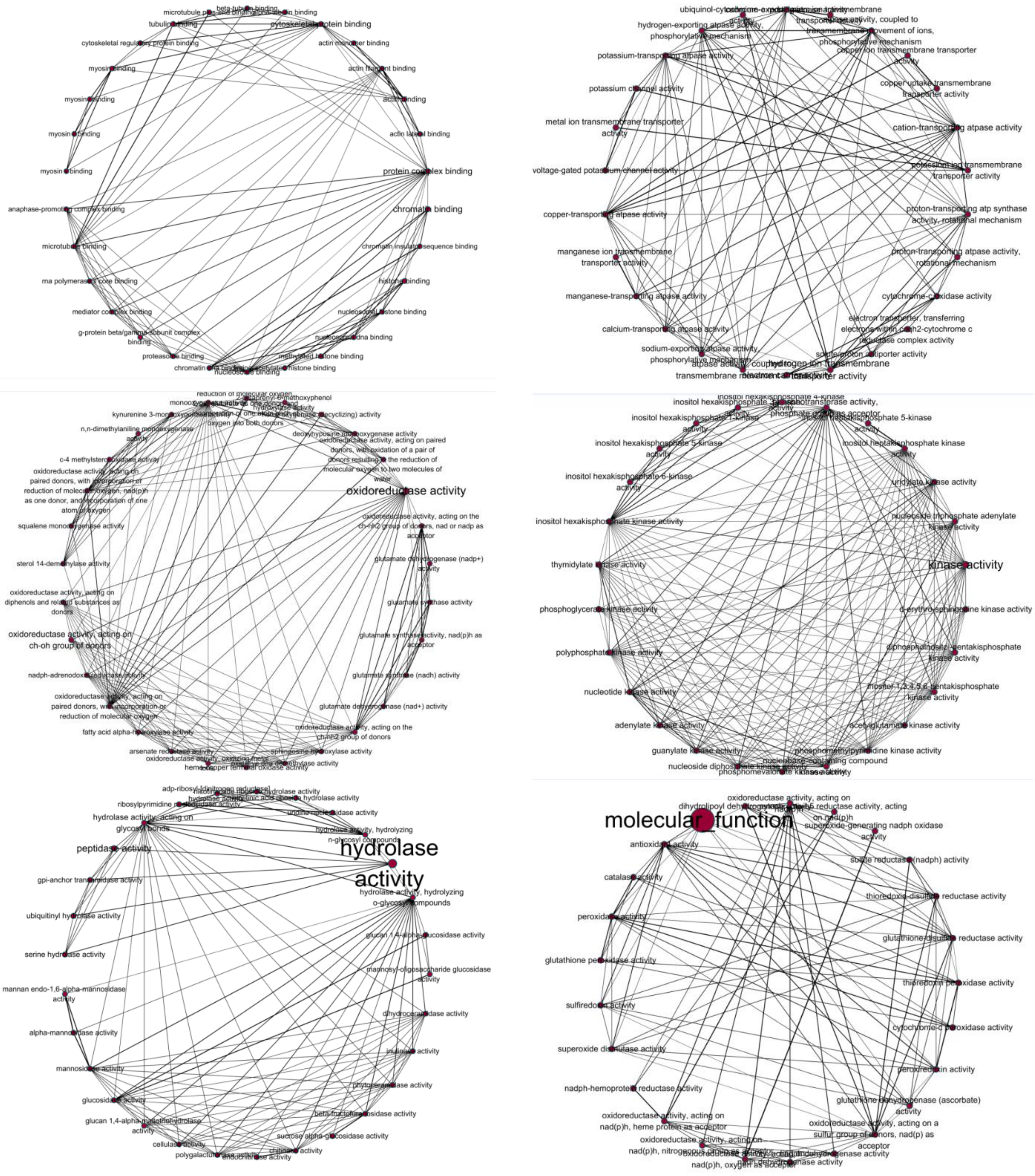
## 6 The complete list of clusters

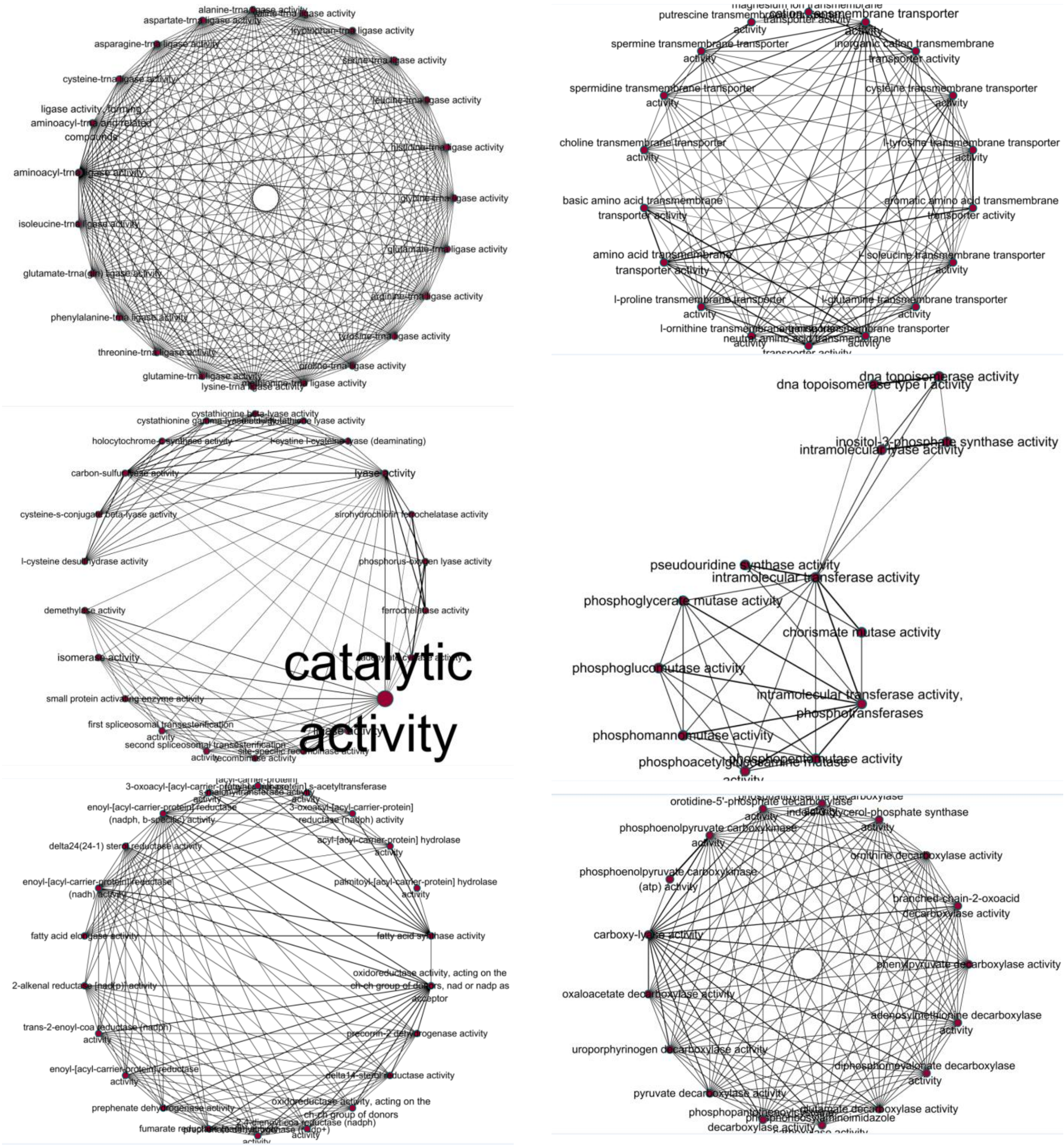
We show the complete list of clusters here.

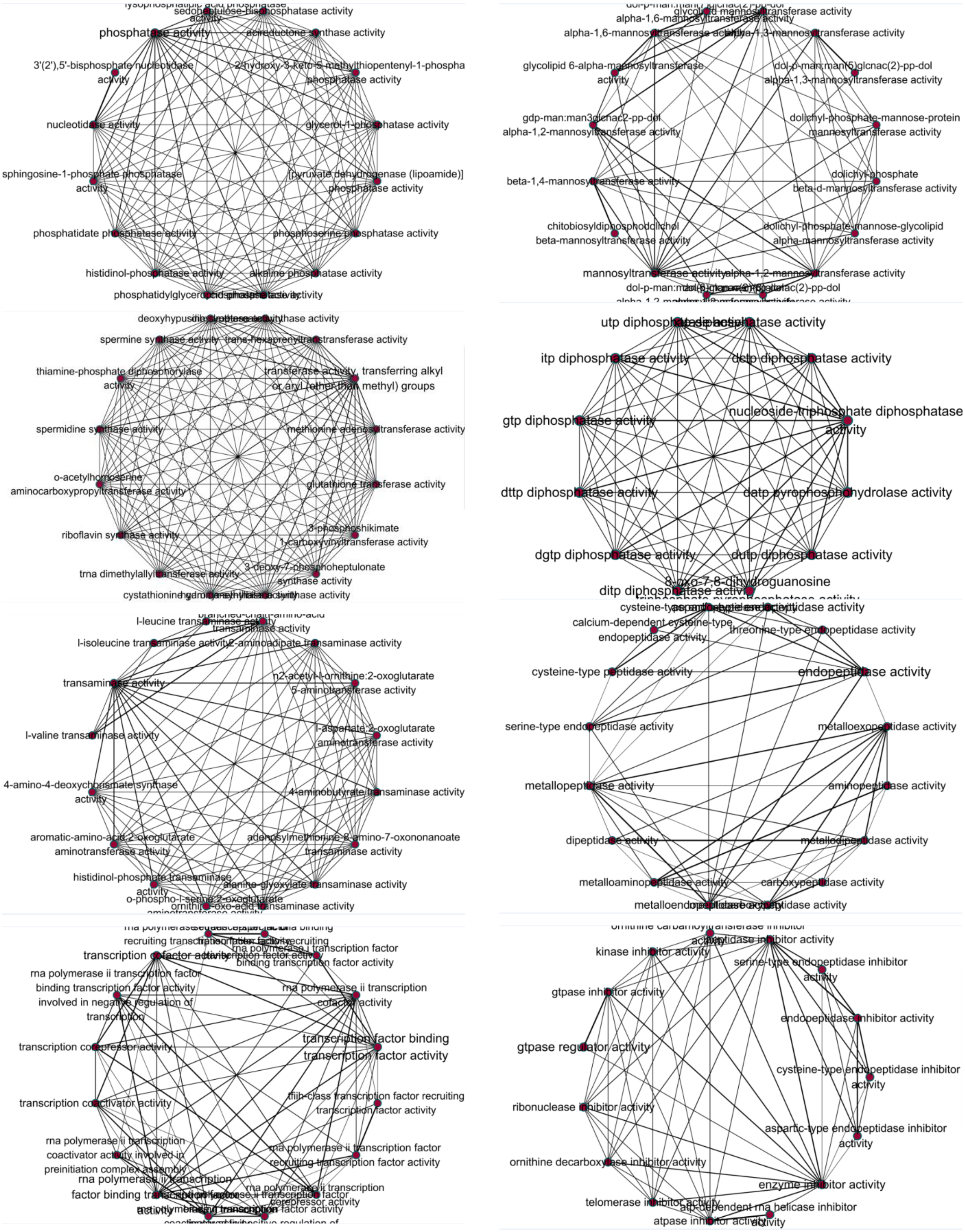

rna binding
mrna binding
u3 snorna binding
box h/aca snorna binding
rna strand annealing activity
pre-mrna 5'-splice site binding
pre-mrna branch point binding
pre-mrna intronic binding
pre-mrna binding
snorna binding
double-stranded rna binding
rna strand-exchange activity
single-stranded rna binding
poly(a) rna binding
mrna 5'-utr binding
au-rich element binding
nucleic acid binding
acid binding
7s rna binding
u5 snrna binding
u1 snrna binding
u6 snrna binding
u2 snrna binding
snrna binding
trna binding
rna cap binding
rrna binding
mitochondrial ribosomal large subunit rrna binding
large ribosomal subunit rrna binding
rrna primary transcript binding
small ribosomal subunit rrna binding
5s rrna binding
dna clamp loader activity
damaged dna binding
guanyl-nucleotide exchange factor activity
protein tag
translation regulator activity
structural molecule activity
enzyme regulator activity
exoribonuclease activator activity
dna polymerase processivity factor activity
inositol phosphoceramide synthase regulator activity
phosphatase regulator activity
protein serine/threonine phosphatase inhibitor activity
protein phosphatase activator activity
protein phosphatase type 1 activator activity
protein phosphatase inhibitor activity
phosphatase activator activity
protein phosphatase type 1 regulator activity
calcium-dependent protein serine/threonine phosphatase regulator activity
protein phosphatase type 2a regulator activity
protein phosphatase regulator activity
enzyme activator activity
copper chaperone activity
metallochaperone activity
receptor activity
ferric iron binding
ferrous iron binding
iron chaperone activity
iron ion binding
superoxide dismutase copper chaperone activity
peptidase activator activity
endopeptidase activator activity
acetyltransferase activator activity
rab gdp-dissociation inhibitor activity
rho gdp-dissociation inhibitor activity
small gtpase regulator activity
ras gtpase activator activity
ran gtpase activator activity
rho gtpase activator activity
rab gtpase activator activity
gtpase activator activity
sar gtpase activator activity
arf gtpase activator activity
gdp-dissociation inhibitor activity
sar guanyl-nucleotide exchange factor

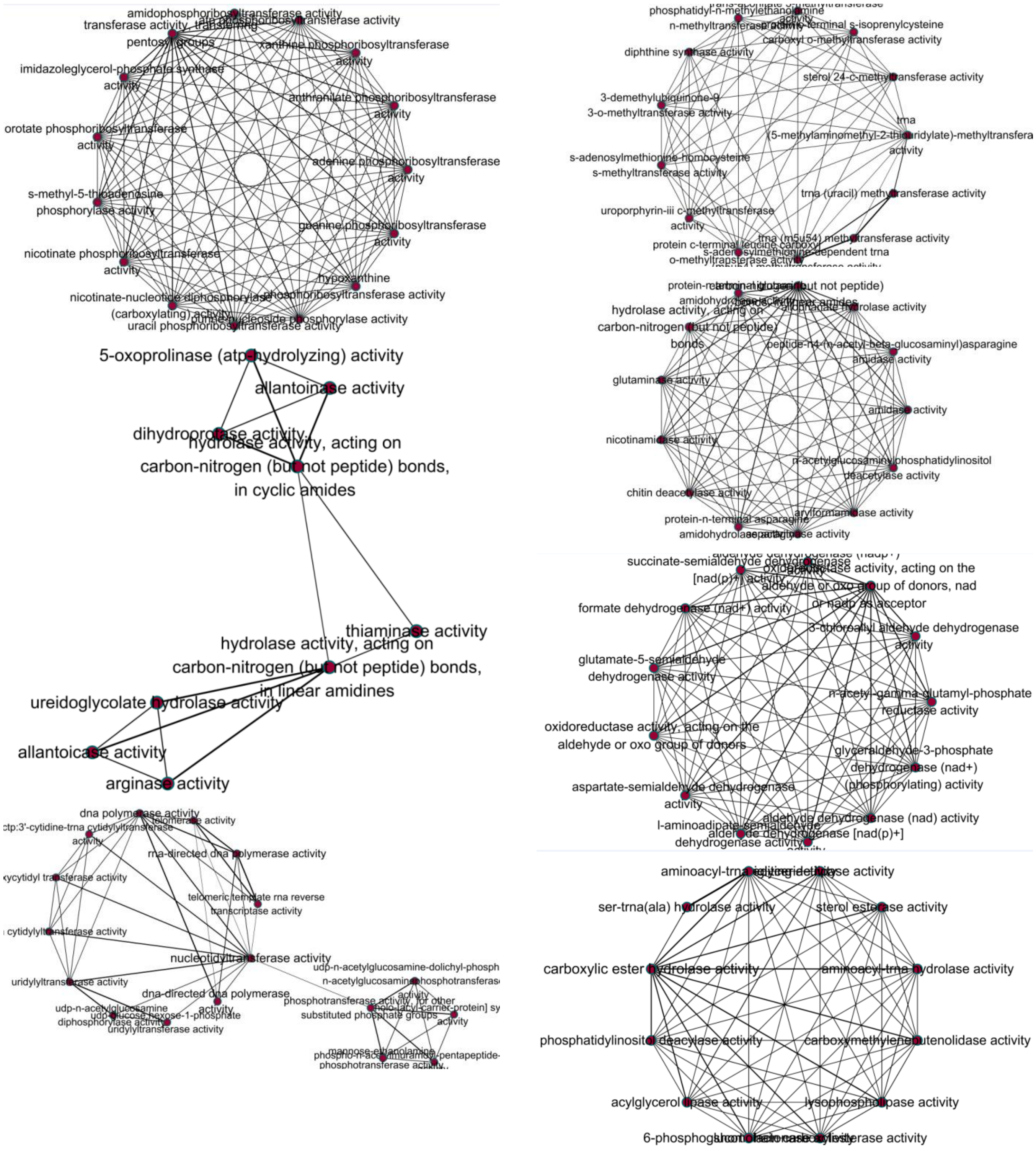

malate dehydrogenase
glycerol dehydrogenase
glyoxylate reductase
oxidoreductase activity, acting on the ch-oh group of donors, nad or nadp as acceptor
malic enzyme activity
2-dehydropantoate 2-reductase activity
aryl-alcohol dehydrogenase (nad+) activity
methylglyoxal reductase (nadh-dependent) activity
serine 3-dehydrogenase activity
shikimate 3-dehydrogenase (nadp+) activity
glucose-6-phosphate dehydrogenase activity
histidinol dehydrogenase activity
methyltransferase activity
protein-arginine n-methyltransferase activity
histone methyltransferase activity
histone-glutamine methyltransferase activity
histone methyltransferase activity (h3-k4 specific)
histone methyltransferase activity (h3-k79 specific)
histone methyltransferase activity (h3-k36 specific)
histone-lysine n-methyltransferase activity
[cytochrome c]-lysine n-methyltransferase activity
protein-lysine n-methyltransferase activity
protein-histidine n-methyltransferase activity
mrna (guanine-n7-)-methyltransferase activity
nicotinamide n-methyltransferase activity
phosphatidylethanolamine n-methyltransferase activity
rrna (uridine-n3-)-methyltransferase activity
phosphatidyl-n-dimethylethanolamine n-methyltransferase activity
trna (guanine) methyltransferase activity
trna methyltransferase activity
rrna methyltransferase activity
aminomethyltransferase activity
n-methyltransferase activity
rrna (adenine) methyltransferase activity
rrna (pseudouridine) methyltransferase activity
s-adenosylmethionine-dependent methyltransferase activity
nuclease activity
endonuclease activity
exonuclease activity
phosphoric ester hydrolase activity
hydrolase activity, acting on ester bonds
endodeoxyribonuclease activity
endoribonuclease activity
5'-flap endonuclease activity
double-stranded dna 3'-5' exodeoxyribonuclease activity
single-stranded dna 3'-5' exodeoxyribonuclease activity
3'-5'-exoribonuclease activity
exoribonuclease ii activity
3'-5'-exodeoxyribonuclease activity
poly(a)-specific ribonuclease activity
3'-5' exonuclease activity
phosphodiesterase i activity
exodeoxyribonuclease activity
5'-3' exoribonuclease activity
5'-3' exodeoxyribonuclease activity
lipase activity
phospholipase activity
calcium-independent phospholipase a2 activity
cyclic-nucleotide phosphodiesterase activity
phospholipase a2 activity
phospholipase d activity
phosphoric diester hydrolase activity
3',5'-cyclic-amp phosphodiesterase activity
2',3'-cyclic-nucleotide 3'-phosphodiesterase activity
protein binding
protein homodimerization activity
identical protein binding
protein domain specific binding
sumo binding
sumo polymer binding
small conjugating protein binding
polyubiquitin binding
ubiquitin binding
protein binding involved in protein folding
atp-dependent protein binding
protein dimerization activity
syntaxin binding
snare binding
proline-rich region binding
chaperone binding
unfolded protein binding
cyclin binding
heat shock protein binding
protein binding, bridging
dynein binding
protein complex scaffold
enzyme binding
transporter activity
substrate-specific transmembrane transporter activity
hydrolase activity, acting on acid anhydrides
peptide transporter activity
oligopeptide-transporting atpase activity
peptide-transporting atpase activity
atpase activity, coupled to transmembrane movement of substances
ion transmembrane transporter activity
voltage-gated ion channel activity
ion channel activity
secondary active sulfate transmembrane transporter activity
sulfate transmembrane transporter activity
ammonium transmembrane transporter activity
sulfite transmembrane transporter activity
arsenite transmembrane transporter activity
antimonite transmembrane transporter activity
fluoride transmembrane transporter activity
borate transmembrane transporter activity
vitamin transporter activity
thiamine transmembrane transporter activity
alcohol transmembrane transporter activity
sterol-transporting atpase activity
lipid transporter activity
sterol transporter activity
ethanolamine transmembrane transporter activity
organic acid transmembrane transporter activity
carbohydrate transmembrane transporter activity
glycerol transmembrane transporter activity
nucleoside-triphosphatase activity
microtubule motor activity
minus-end-directed microtubule motor activity
plus-end directed microfilament motor activity
microfilament motor activity
plus-end-directed microtubule motor activity
motor activity
gtpase activity
helicase activity
atp-dependent peptidase activity
atpase activity
atpase activity, uncoupled
atpase activity, coupled
atp-dependent 3'-5' dna helicase activity
atp-dependent 5'-3' rna helicase activity
nucleosome-dependent atpase activity
tubulin-dependent atpase activity
proteasome-activating atpase activity
single-stranded dna-dependent atpase activity
dna topoisomerase type ii (atp-hydrolyzing) activity
dna translocase activity
dna-dependent atpase activity
atp-dependent dna helicase activity
3'-5' dna helicase activity
four-way junction helicase activity
atp-dependent helicase activity
atp-dependent rna helicase activity
double-stranded dna-dependent atpase activity
rna helicase activity


cyclin-dependent protein
pyruvate dehydrogenase kinase activity
(acetyl-transferring) kinase activity
protein tyrosine kinase activity
protein kinase activity involved in regulation of protein localization to cell division site involved in cytokinesis
protein histidine kinase activity
protein serine/threonine kinase activity
protein serine/threonine/tyrosine kinase activity
map kinase kinase activity
cyclin-dependent protein serine/threonine kinase activity
calmodulin-dependent protein kinase activity
amp-activated protein kinase activity
rna polymerase ii carboxy-terminal domain kinase activity
eukaryotic translation initiation factor 2alpha kinase activity
protein kinase c activity
receptor signaling protein serine/threonine kinase activity
map kinase activity
kinase regulator activity
camp-dependent protein kinase inhibitor activity
protein kinase inhibitor activity
protein serine/threonine kinase inhibitor activity
kinase regulator activity
camp-dependent protein kinase regulator activity
cyclin-dependent protein serine/threonine kinase inhibitor activity
protein kinase activator activity
protein serine/threonine kinase activator activity
protein kinase activity
phosphorelay sensor kinase activity
phosphorelay response regulator activity
signal transducer activity
histidine phosphotransfer kinase activity
receptor signaling protein activity
bis(5'-nucleosyl)-tetraphosphatase activity
nucleotide diphosphatase activity
inositol diphosphate pentakisphosphate diphosphatase activity
nucleoside-diphosphatase activity
uridine-diphosphatase activity
guanosine-diphosphatase activity
8-oxo-dgdp phosphatase activity
adp-ribose diphosphatase activity
phosphoribosyl-atp diphosphatase activity
inorganic diphosphatase activity
endopolyphosphatase activity
hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides
rna pyrophosphohydrolase activity
bis(5'-adenosyl)-hexaphosphatase activity
bis(5'-adenosyl)-triphosphatase activity
pyrophosphatase activity
inositol-5-diphosphate-1,2,3,4,6-pentakisphosphate diphosphatase activity
inositol-1-diphosphate-2,3,4,5,6-pentakisphosphate diphosphatase activity
inositol-3-diphosphate-1,2,4,5,6-pentakisphosphate diphosphatase activity
inositol-1,5-bisdiphosphate-2,3,4,6-tetrakisphosphate 5-diphosphatase activity
inositol-3,5-bisdiphosphate-2,3,4,6-tetrakisphosphate 5-diphosphatase activity
inositol-1,5-bisdiphosphate-2,3,4,6-tetrakisphosphate 1-diphosphatase activity
inositol bisdiphosphate tetrakisphosphate diphosphatase activity
inositol diphosphate tetrakisphosphate diphosphatase activity
diphosphoinositol polyphosphate diphosphatase activity
n-acetylglucosaminyldiphosphodolichol
phosphatidylinositol
n-acetylglucosaminyltransferase activity
udp-glycosyltransferase activity
transferase activity
transferase activity, transferring acyl groups other than amino-acyl groups
protein-cysteine s-palmitoyltransferase activity
palmitoyltransferase activity
octanoyltransferase activity
lipoyl(octanoyl) transferase activity
peptide alpha-n-propionyltransferase activity
o-acyltransferase activity
glucosaminyl-phosphotidylinositol o-acyltransferase activity
glycerol-3-phosphate o-acyltransferase activity
1-acylglycerophosphocholine o-acyltransferase activity
transferase activity, transferring acyl groups
homocitrate synthase activity
hydroxymethylglutaryl-coa synthase activity
2-isopropylmalate synthase activity
atp citrate synthase activity
2-methylcitrate synthase activity
transferase activity, transferring acyl groups, acyl groups converted into alkyl on transfer
transferase activity, transferring phosphorus-containing groups
mannosylphosphate transferase activity
transferase activity, transferring glycosyl groups
4-alpha-glucanotransferase activity
phosphorylase activity
glycogen phosphorylase activity
transferase activity, transferring hexosyl groups
dolichyl-diphosphooligosaccharide-protein glycotransferase activity
oligosaccharyl transferase activity
1,4-alpha-glucan branching enzyme activity
1-phosphatidylinositol binding
lipid binding
phosphatidylserine binding
phosphatidylinositol binding
phospholipid binding
phosphatidylcholine binding
calcium-dependent phospholipid binding
calcium ion binding
sphingolipid binding
magnesium ion binding
phosphatidic acid binding
metal ion binding
cation binding
thiamine pyrophosphate binding
coenzyme binding
flavin adenine dinucleotide binding
nadp binding
purine nucleotide binding
cofactor binding
adenyl-nucleotide exchange factor activity
pyridoxal phosphate binding
adenyl nucleotide binding
quinone binding
nucleotide binding
ubiquinone binding
fatty-acyl-coa binding
guanyl nucleotide binding
nad binding
fmn binding
inositol pentakisphosphate 2-kinase activity
adenylylsulfate kinase activity
choline kinase activity
ethanolamine kinase activity
xylulokinase activity
phosphatidylinositol phosphate kinase activity
ribokinase activity
nad+ kinase activity
shikimate kinase activity
phosphotransferase activity, alcohol group as acceptor
pyridoxal kinase activity
trna 2'-phosphotransferase activity
sphinganine kinase activity
hydroxyethylthiazole kinase activity
riboflavin kinase activity
pyruvate kinase activity
nadh kinase activity
pantothenate kinase activity
ribosylnicotinamide kinase activity
glycerol kinase activity
hydroxymethylpyrimidine kinase activity
glycerone kinase activity
gluconokinase activity
hexokinase activity
homoserine kinase activity
polynucleotide 5'-hydroxyl-kinase activity
mevalonate kinase activity
diacylglycerol kinase activity
galactokinase activity
rna polymerase ii core promoter proximal region dna
sequence-specific dna binding
mitochondrial light strand promoter anti-sense binding
rna polymerase ii regulatory region dna binding
rna polymerase ii regulatory region sequence-specific dna binding
core promoter binding
transcription regulatory region dna binding
core promoter sequence-specific dna binding
rna polymerase i regulatory region dna binding
rna polymerase i upstream control element sequence-specific dna binding
rna polymerase ii core promoter proximal region sequence-specific dna binding
core promoter proximal region sequence-specific dna binding
single-stranded dna binding
sequence-specific dna binding, bending
single-stranded telomeric dna binding
centromeric dna binding
at dna binding
rdna binding
dna replication origin binding
dna binding
telomeric dna binding
dna binding, bending
mismatched dna binding
sequence-specific dna binding
rdna spacer replication fork barrier binding
double-stranded dna binding


beta-tubulin binding
microtubule plus-end binding
tubulin binding
cytoskeletal protein binding
cytoskeletal regulatory protein binding
actin monomer binding
myosin binding
actin filament binding
myosin binding
actin binding
myosin binding
actin lateral binding
myosin binding
protein complex binding
anaphase-promoting complex binding
chromatin binding
microtubule binding
chromatin insulator sequence binding
rna polymerase ii core binding
histone binding
mediator complex binding
nucleosomal histone binding
g-protein beta/gamma-subunit complex binding
nucleosomal dna binding
proteasome binding
methylated histone binding
chromatin dna binding
nucleosome binding
acetylated histone binding
potassium-transporting atpase activity
ubiquinol-cytochrome-c reductase activity
hydrogen-exporting atpase activity, phosphorylative mechanism
atpase activity, coupled to transmembrane movement of ions, phosphorylative mechanism
copper ion transmembrane transporter activity
potassium channel activity
copper uptake transmembrane transporter activity
metal ion transmembrane transporter activity
cation-transporting atpase activity
voltage-gated potassium channel activity
potassium ion transmembrane transporter activity
copper-transporting atpase activity
proton-transporting atp synthase activity, rotational mechanism
manganese ion transmembrane transporter activity
proton-transporting atpase activity, rotational mechanism
manganese-transporting atpase activity
cytochrome-c oxidase activity
calcium-transporting atpase activity
electron transporter, transferring electrons within coqh2-cytochrome c reductase complex activity
sodium-exporting atpase activity, phosphorylative mechanism
solute:proton antiporter activity
atpase activity, coupled to transmembrane movement of substances
hydrogen ion transmembrane transporter activity
electron carrier activity
reduction of molecular oxygen
2-octaprenyl-6-methoxyphenol hydroxylase activity
monooxygenase activity
kynurenine 3-monooxygenase activity
incorporation of one atom each of oxygen into both donors
heme oxygenase (decyclizing) activity
n,n-dimethylaniline monooxygenase activity
deoxyhypusine monooxygenase activity
oxidoreductase activity, acting on paired donors, with oxidation of a pair of donors resulting in the reduction of molecular oxygen to two molecules of water
c-4 methylsterol oxidase activity
oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, nad(p)h as one donor, and incorporation of one atom of oxygen
oxidoreductase activity
squalene monooxygenase activity
oxidoreductase activity, acting on the ch-nh2 group of donors, nad or nadp as acceptor
sterol 14-demethylase activity
glutamate dehydrogenase (nadp+) activity
oxidoreductase activity, acting on diphenols and related substances as donors
glutamate synthase activity
oxidoreductase activity, acting on ch-oh group of donors
glutamate synthase activity, nad(p)h as acceptor
nadph-adrenodoxin reductase activity
glutamate synthase (nadh) activity
oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen
glutamate dehydrogenase (nad+) activity
fatty acid alpha-hydroxylase activity
oxidoreductase activity, acting on the ch-nh2 group of donors
arsenate reductase activity
sphingosine hydroxylase activity
oxidoreductase activity, oxidizing metal ions
heme-copper terminal oxidase activity
oxidative dna demethylase activity
inositol hexakisphosphate 4-kinase activity
inositol hexakisphosphate 3-kinase activity
phosphotransferase activity, phosphate group as acceptor
inositol hexakisphosphate 1-kinase activity
inositol heptakisphosphate 5-kinase activity
inositol hexakisphosphate 5-kinase activity
inositol heptakisphosphate kinase activity
inositol hexakisphosphate 6-kinase activity
uridylate kinase activity
inositol hexakisphosphate kinase activity
nucleoside triphosphate adenylate kinase activity
thymidylate kinase activity
kinase activity
phosphoglycerate kinase activity
d-erythro-sphingosine kinase activity
polyphosphate kinase activity
diphosphoinositol-pentakisphosphate kinase activity
nucleotide kinase activity
inositol-1,3,4,5,6-pentakisphosphate kinase activity
adenylate kinase activity
acetylglutamate kinase activity
guanylate kinase activity
phosphomethylpyrimidine kinase activity
nucleoside diphosphate kinase activity
nucleobase-containing compound kinase activity
phosphomevalonate kinase activity
adp-ribosyl-[dinitrogen reductase] hydrolase activity
nicotinamide riboside hydrolase activity
nicotinic acid riboside hydrolase activity
ribosylpyrimidine nucleosidase activity
uridine nucleosidase activity
hydrolase activity, acting on glycosyl bonds
hydrolase activity, hydrolyzing n-glycosyl compounds
peptidase activity
hydrolase activity
gpi-anchor transamidase activity
hydrolase activity, hydrolyzing o-glycosyl compounds
ubiquitinyl hydrolase activity
glucan 1,4-alpha-glucosidase activity
serine hydrolase activity
mannosyl-oligosaccharide glucosidase activity
mannan endo-1,6-alpha-mannosidase activity
dihydroceramidase activity
alpha-mannosidase activity
inulinase activity
mannosidase activity
phytoceramidase activity
glucosidase activity
beta-fructofuranosidase activity
glucan 1,4-alpha-maltotriohydrolase activity
sucrose alpha-glucosidase activity
cellulase activity
chitinase activity
polygalacturonase activity
endochitinase activity
oxidoreductase activity, acting on nad(p)h
dihydrolipoyl dehydrogenase activity
cytochrome-b5 reductase activity, acting on nad(p)h
molecular_function
superoxide-generating nadph oxidase activity
antioxidant activity
sulfite reductase (nadph) activity
catalase activity
thioredoxin-disulfide reductase activity
peroxidase activity
glutathione-disulfide reductase activity
glutathione peroxidase activity
thioredoxin peroxidase activity
sulfiredoxin activity
cytochrome-c peroxidase activity
superoxide dismutase activity
peroxiredoxin activity
nadph-hemoprotein reductase activity
glutathione dehydrogenase (ascorbate) activity
oxidoreductase activity, acting on nad(p)h, heme protein as acceptor
oxidoreductase activity, acting on a sulfur group of donors, nad(p) as acceptor
oxidoreductase activity, acting on nad(p)h, nitrogenous group as acceptor
oxidoreductase activity, acting on nad(p)h, oxygen as acceptor
nadh dehydrogenase activity


alanine-trna ligase activity
aspartate-trna ligase activity
valine-trna ligase activity
asparagine-trna ligase activity
tryptophan-trna ligase activity
serine-trna ligase activity
cysteine-trna ligase activity
leucine-trna ligase activity
ligase activity, forming aminoacyl-trna and related compounds
histidine-trna ligase activity
aminoacyl-trna ligase activity
glycine-trna ligase activity
isoleucine-trna ligase activity
glutamate-trna ligase activity
glutamate-trna(gln) ligase activity
arginine-trna ligase activity
phenylalanine-trna ligase activity
tyrosine-trna ligase activity
threonine-trna ligase activity
proline-trna ligase activity
glutamine-trna ligase activity
lysine-trna ligase activity
methionine-trna ligase activity
magnesium ion transmembrane transporter activity
putrescine transmembrane transporter activity
cation transmembrane transporter activity
spermine transmembrane transporter activity
inorganic cation transmembrane transporter activity
spermidine transmembrane transporter activity
cysteine transmembrane transporter activity
choline transmembrane transporter activity
l-tyrosine transmembrane transporter activity
basic amino acid transmembrane transporter activity
aromatic amino acid transmembrane transporter activity
amino acid transmembrane transporter activity
l-isoleucine transmembrane transporter activity
l-proline transmembrane transporter activity
l-glutamine transmembrane transporter activity
l-ornithine transmembrane transporter activity
neutral amino acid transmembrane transporter activity
cystathionine beta-lyase activity
cystathionine gamma-lyase activity
holocytochrome-c synthase activity
l-cystine l-cysteine-lyase (deaminating)
carbon-sulfur lyase activity
lyase activity
cysteine-s-conjugate beta-lyase activity
sirohydrochlorin ferrochelatase activity
l-cysteine desulfhydrase activity
phosphorus-oxygen lyase activity
demethylase activity
ferrochelatase activity
isomerase activity
catalytic activity
small protein activating enzyme activity
first spliceosomal transesterification activity
second spliceosomal transesterification activity
site-specific recombinase activity
recombinase activity
dna topoisomerase activity
dna topoisomerase type I activity
inositol-3-phosphate synthase activity
intramolecular lyase activity
pseudouridine synthase activity
intramolecular transferase activity
phosphoglycerate mutase activity
chorismate mutase activity
phosphoglucomutase activity
intramolecular transferase activity, phosphotransferases
phosphomannomutase activity
phosphopentomutase activity
phosphoacetylglucosamine mutase activity
[acyl-carrier-protein] s-malonyltransferase activity
3-oxoacyl-[acyl-carrier-protein] synthase activity
[acyl-carrier-protein] s-acetyltransferase activity
enoyl-[acyl-carrier-protein] reductase (nadph, b-specific) activity
3-oxoacyl-[acyl-carrier-protein] reductase (nadph) activity
delta24(24-1) sterol reductase activity
acyl-[acyl-carrier-protein] hydrolase activity
enoyl-[acyl-carrier-protein] reductase (nadh) activity
palmitoyl-[acyl-carrier-protein] hydrolase activity
fatty acid elongase activity
fatty acid synthase activity
oxidoreductase activity, acting on the ch-ch group of donors, nad or nadp as acceptor
2-alkenal reductase [nad(p)] activity
trans-2-enoyl-coa reductase (nadph) activity
precorrin-2 dehydrogenase activity
enoyl-[acyl-carrier-protein] reductase activity
delta14-sterol reductase activity
prephenate dehydrogenase activity
oxidoreductase activity, acting on the ch-ch group of donors
fumarate reductase (nadh) activity
2,4-dienoyl-coa reductase (nadph) activity
prephenate dehydrogenase (nadp+) activity
phosphatidylserine decarboxylase
orotidine-5'-phosphate decarboxylase activity
indole-3-glycerol-phosphate synthase activity
phosphoenolpyruvate carboxykinase activity
ornithine decarboxylase activity
phosphoenolpyruvate carboxykinase (atp) activity
branched-chain-2-oxoacid decarboxylase activity
carboxy-lyase activity
phenylpyruvate decarboxylase activity
oxaloacetate decarboxylase activity
adenosylmethionine decarboxylase activity
uroporphyrinogen decarboxylase activity
diphosphomevalonate decarboxylase activity
pyruvate decarboxylase activity
phosphopantothenoylcysteine decarboxylase activity
glutamate decarboxylase activity
phosphoribosylaminoimidazole carboxylase activity


lysophosphatidic acid phosphatase
sedoheptulose-bisphosphatase activity
phosphatase activity
acireductone synthase activity
3'(2'),5'-bisphosphate nucleotidase activity
2-hydroxy-3-keto-5-methylthiopentenyl-1-phospha phosphatase activity
nucleotidase activity
glycerol-1-phosphatase activity
sphingosine-1-phosphate phosphatase activity
[pyruvate dehydrogenase (lipoamide)] phosphatase activity
phosphatidate phosphatase activity
phosphoserine phosphatase activity
histidinol-phosphatase activity
alkaline phosphatase activity
phosphatidylglycerophosphatase activity
dol-p-man:man(7)glcnac(2)-pp-dol
glycolipid mannosyltransferase activity
alpha-1,6-mannosyltransferase activity
alpha-1,6-mannosyltransferase activity
alpha-1,3-mannosyltransferase activity
glycolipid 6-alpha-mannosyltransferase activity
dol-p-man:man(5)glcnac(2)-pp-dol alpha-1,3-mannosyltransferase activity
gdp-man:man3glcnac2-pp-dol alpha-1,2-mannosyltransferase activity
dolichyl-phosphate-mannose-protein mannosyltransferase activity
beta-1,4-mannosyltransferase activity
dolichyl-phosphate beta-d-mannosyltransferase activity
chitobiosyldiphosphodolichol beta-mannosyltransferase activity
dolichyl-phosphate-mannose-glycolipid alpha-mannosyltransferase activity
mannosyltransferase activity
alpha-1,2-mannosyltransferase activity
dol-p-man:man(6)glcnac(2)-pp-dol alpha-1,2-mannosyltransferase activity
deoxyhypusine synthase activity
spermine synthase activity
trans-hexaprenyltranstransferase activity
thiamine-phosphate diphosphorylase activity
transferase activity, transferring alkyl or aryl (other than methyl) groups
spermidine synthase activity
methionine adenosyltransferase activity
o-acetylhomoserine aminocarboxypropyltransferase activity
glutathione transferase activity
riboflavin synthase activity
3-phosphoshikimate 1-carboxyvinyltransferase activity
trna dimethylallyltransferase activity
3-deoxy-7-phosphoheptulonate synthase activity
cystathionine gamma-synthase activity
homospermidine synthase activity
utp diphosphatase activity
atp diphosphatase activity
itp diphosphatase activity
dctp diphosphatase activity
gtp diphosphatase activity
nucleoside-triphosphate diphosphatase activity
dttp diphosphatase activity
datp pyrophosphohydrolase activity
dgtp diphosphatase activity
dutp diphosphatase activity
ditp diphosphatase activity
8-oxo-7,8-dihydroguanosine triphosphate pyrophosphatase activity
branched-chain-amino-acid transaminase activity
l-leucine transaminase activity
l-isoleucine transaminase activity
2-aminoadipate transaminase activity
transaminase activity
n2-acetyl-l-ornithine:2-oxoglutarate 5-aminotransferase activity
l-valine transaminase activity
l-aspartate:2-oxoglutarate aminotransferase activity
4-amino-4-deoxychorismate synthase activity
4-aminobutyrate transaminase activity
aromatic-amino-acid:2-oxoglutarate aminotransferase activity
adenosylmethionine-8-amino-7-oxononanoate transaminase activity
histidinol-phosphate transaminase activity
alanine-glyoxylate transaminase activity
o-phospho-l-serine:2-oxoglutarate aminotransferase activity
ornithine-oxo-acid transaminase activity
cysteine-type endopeptidase activity
aspartic-type endopeptidase activity
calcium-dependent cysteine-type endopeptidase activity
threonine-type endopeptidase activity
cysteine-type peptidase activity
endopeptidase activity
serine-type endopeptidase activity
metalloexopeptidase activity
metallopeptidase activity
aminopeptidase activity
dipeptidase activity
metallodipeptidase activity
metalloaminopeptidase activity
carboxypeptidase activity
metalloendopeptidase activity
rna polymerase ii transcription factor binding
recruiting transcription factor activity
transcription cofactor activity
rna polymerase i transcription factor binding transcription factor activity
rna polymerase ii transcription factor binding transcription factor activity involved in negative regulation of transcription
rna polymerase ii transcription cofactor activity
transcription corepressor activity
transcription factor binding transcription factor activity
transcription coactivator activity
tfiih-class transcription factor recruiting transcription factor activity
rna polymerase ii transcription coactivator activity involved in preinitiation complex assembly
rna polymerase ii transcription factor recruiting transcription factor activity
rna polymerase ii transcription factor binding transcription factor activity
rna polymerase ii transcription corepressor activity
ornithine carbamoyltransferase inhibitor activity
peptidase inhibitor activity
kinase inhibitor activity
serine-type endopeptidase inhibitor activity
gtpase inhibitor activity
endopeptidase inhibitor activity
gtpase regulator activity
cysteine-type endopeptidase inhibitor activity
ribonuclease inhibitor activity
aspartic-type endopeptidase inhibitor activity
ornithine decarboxylase inhibitor activity
enzyme inhibitor activity
telomerase inhibitor activity
atp-dependent rna helicase inhibitor activity
atpase inhibitor activity


amidophosphoribosyltransferase activity
atp phosphoribosyltransferase activity
transferase activity, transferring pentosyl groups
xanthine phosphoribosyltransferase activity
imidazoleglycerol-phosphate synthase activity
anthranilate phosphoribosyltransferase activity
orotate phosphoribosyltransferase activity
adenine phosphoribosyltransferase activity
s-methyl-5-thioadenosine phosphorylase activity
guanine phosphoribosyltransferase activity
nicotinate phosphoribosyltransferase activity
hypoxanthine phosphoribosyltransferase activity
nicotinate-nucleotide diphosphorylase (carboxylating) activity
purine-nucleoside phosphorylase activity
uracil phosphoribosyltransferase activity
5-oxoprolinase (atp-hydrolyzing) activity
allantoinase activity
dihydroorotase activity
hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds, in cyclic amides
thiaminase activity
hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds, in linear amidines
ureidoglycolate hydrolase activity
allantoicase activity
arginase activity
dna polymerase activity
telomerase activity
ctp:3'-cytidine-trna cytidylyltransferase activity
rna-directed dna polymerase activity
xycytidyl transferase activity
telomeric template rna reverse transcriptase activity
cytidylyltransferase activity
nucleotidyltransferase activity
uridylyltransferase activity
dna-directed dna polymerase activity
udp-n-acetylglucosamine diphosphorylase activity
udp-glucose:hexose-1-phosphate uridylyltransferase activity
udp-n-acetylglucosamine-dolichyl-phosph
n-acetylglucosaminephosphotransferas
activity
phosphotransferase activity, for other substituted phosphate groups
holo-[acyl-carrier-protein] sy
activity
mannose-ethanolamine phosphotransferase activity
phospho-n-acetylmuramoyl-pentapeptide-
activity
trans-aconitate 3-methyltransferase activity
phosphatidyl-n-methylethanolamine n-methyltransferase activity
protein-c-terminal s-isoprenylcysteine carboxyl o-methyltransferase activity
diphthine synthase activity
sterol 24-c-methyltransferase activity
3-demethylubiquinone-9 3-o-methyltransferase activity
trna (5-methylaminomethyl-2-thiouridylate)-methyltransfera activity
s-adenosylmethionine-homocysteine s-methyltransferase activity
trna (uracil) methyltransferase activity
uroporphyrin-iii c-methyltransferase activity
trna (m5u54) methyltransferase activity
protein c-terminal leucine carboxyl o-methyltransferase activity
s-adenosylmethionine-dependent trna (m5u54) methyltransferase activity
protein-n-terminal glutamine amidohydrolase activity
hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds, in linear amides
allophanate hydrolase activity
hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds
peptide-n4-(n-acetyl-beta-glucosaminyl)asparagine amidase activity
glutaminase activity
amidase activity
nicotinamidase activity
n-acetylglucosaminylphosphatidylinositol deacetylase activity
chitin deacetylase activity
arylformamidase activity
protein-n-terminal asparagine amidohydrolase activity
asparaginase activity
aldehyde dehydrogenase (nadp+) activity
succinate-semialdehyde dehydrogenase [nad(p)+] activity
oxidoreductase activity, acting on the aldehyde or oxo group of donors, nad or nadp as acceptor
formate dehydrogenase (nad+) activity
3-chloroallyl aldehyde dehydrogenase activity
glutamate-5-semialdehyde dehydrogenase activity
n-acetyl-gamma-glutamyl-phosphate reductase activity
oxidoreductase activity, acting on the aldehyde or oxo group of donors
glyceraldehyde-3-phosphate dehydrogenase (nad+) (phosphorylating) activity
aspartate-semialdehyde dehydrogenase activity
aldehyde dehydrogenase (nad) activity
l-aminoadipate-semialdehyde dehydrogenase activity
aldehyde dehydrogenase [nad(p)+] activity
aminoacyl-trna editing activity
triglyceride lipase activity
ser-trna(ala) hydrolase activity
sterol esterase activity
carboxylic ester hydrolase activity
aminoacyl-trna hydrolase activity
phosphatidylinositol deacylase activity
carboxymethylenebutenolidase activity
acylglycerol lipase activity
lysophospholipase activity
6-phosphogluconolactonase activity
acetylcholinesterase activity
carboxylesterase activity

# Diffusion Component Analysis: Unraveling Functional Topology in Biological Networks

Hyunghoon Cho[1], Bonnie Berger[1,2,*] and Jian Peng[1,2,3,*]

[1] Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA, USA
[2] Department of Mathematics, MIT, Cambridge, MA, USA
[3] Department of Computer Science, University of Illinois at Urbana-Champaign, Champaign, IL, USA
* Corresponding authors: bab@mit.edu, jianpeng@illinois.edu

## 1 Introduction

Complex biological systems have been successfully modeled by biochemical and genetic interaction networks, typically gathered from high-throughput (HTP) data. These networks can be used to infer functional relationships between genes or proteins. Using the intuition that the topological role of a gene in a network relates to its biological function, local or diffusion-based "guilt-by-association" and graph-theoretic methods have had success in inferring gene functions [1, 2, 3]. Here we seek to improve function prediction by integrating diffusion-based methods with a novel dimensionality reduction technique to overcome the incomplete and noisy nature of network data.

A type of diffusion algorithm, also known as random walk with restart (RWR), has been extensively studied in the context of biological networks and effectively applied to protein function prediction (e.g., [1]). The key idea is to propagate information along the network, in order to exploit both direct and indirect linkages between genes. Typically, a distribution of topological similarity is computed for each gene, in relation to other genes in the network, so that researchers can select the most related genes in the resulting distribution or, rather, select genes that share the most similar distributions. Though successful, these approaches are susceptible to noise in the input networks due to the high dimensionality of the computed distributions.

## 2 Methods

We propose Diffusion Component Analysis (DCA), a novel analytical framework that combines diffusion-based methods and sophisticated dimensionality reduction to better extract topological network information in order to facilitate more accurate functional annotation of genes or proteins. The key idea behind DCA is to obtain informative, but low-dimensional features, which better encode the inherent topological properties of each node in the network. We first run a diffusion algorithm on a molecular network to obtain a distribution for each node that captures its relation to all other nodes in the network. We then approximate each of these distributions by constructing a multinomial logistic model, parameterized by low-dimensional feature vector(s), for each node. Feature vectors of all nodes are jointly learned by minimizing the Kullback-Leibler (KL) divergence (relative entropy) between the diffusion and parameterized-multinomial logistic distributions. A key differentiating factor of our novel dimensionality reduction from a more conventional approach, such as Principal Component Analysis (PCA), is the use of multinomial logistic models, which

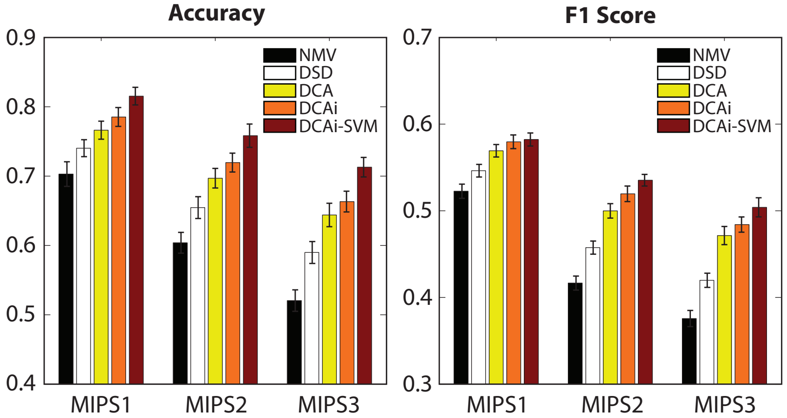


Figure 1: **Protein function prediction performance on yeast STRING networks in terms of both accuracy and F1 score–the harmonic mean of both precision and recall–with different levels of functional categories from MIPS.** Neighbor majority vote (NMV), Diffusion state distance (DSD), DCA with kNN (DCA), DCA combined with novel network integration with kNN (DCAi) or SVM (DCAi-SVM).

more naturally explain the input probability distributions from the diffusion. Moreover, DCA can be naturally extended to integrate multiple heterogeneous networks by performing diffusion on separate networks and jointly optimizing feature vectors. Given the low-dimensional vector representations of nodes, k-nearest neighbor (kNN) voting schemes or support vector machines (SVM) can be used for function prediction.

## 3 Results

We evaluated the ability of our DCA framework to uncover functional relationships in the interactome of yeast. By combining noise reduction via dimensionality reduction, improved integration of multiple heterogeneous networks (e.g., physical interaction, conserved co-expression), and the use of support vector machines, our DCA framework is able to achieve 71.29% accuracy with five-fold cross-validation on the STRING networks with third level functional annotations from MIPS, which is remarkably 12.31% higher than the previous state-of-the-art diffusion state distance (DSD) [1] method (Figure 1). We also observe improved performance over DSD in a different yeast PPI network, constructed from only physical interactions in the BioGRID database. In addition, we found that conventional approaches to dimensionality reduction, such as principal component analysis or non-negative matrix factorization, fail to achieve similar performance improvements. Our results demonstrate the potential of low-dimensional feature vectors learned by DCA to be plugged into other existing machine learning algorithms to decipher functional properties of and obtain novel insights into interactomes.

## References

1. Cao, Mengfei, et al. “New directions for diffusion-based network prediction of protein function: incorporating pathways with confidence.” Bioinformatics 30.12 (2014): i219-i227.

2. Mostafavi, Sara, et al. “GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function.” Genome Biol 9.Suppl 1 (2008): S4.

3. Milenkovi, Tijana, and Nataa Prulj. “Uncovering biological network function via graphlet degree signatures.” Cancer informatics 6 (2008): 257.