

Figure 1: **Illustration of Connectivity Pruning.** In the raw metabolic network assembled from the reaction formula, there is a link between every pair of reactants that are on opposite sides (A). In the network assembled [1], connection 2 and 3 are pruned away. Only connection 1 and 4 are present (B).

Table 1: **False positive (FP) and false negative (FN) connectivity pruning in pathway eco00010.** False positives are defined as reactant pairs that are missing in the reference pathway but with $SCL \geq 0.4$. False negatives are defined as reactant pairs that appear in the reference pathway but with $SCL < 0.4$. Abbreviations in ‘explanation’ column: **frag**: fragment; **quant**: improper quantification of chemical content; **annot**: incomplete annotation; see text for more details.

Type	reactant pair (A, D)	reaction	$SCL_{\text{self}}(A D)$	$SCL_{\text{self}}(D A)$	SCL	explanation
FP	C00011-C00022	R00014	1	0.5	0.5	quant
	C00002-C00008	R00200, R00341, R01512, R01600, R01786, R04779	0.87	1	0.87	frag
	C00010-C00024	R00235, R02569	1	0.94	0.94	quant
	C00002-C00020	R00235	0.71	0.96	0.71	frag
	C00003-C00004	R00754, R01061, R07618	1	1	1	frag
	C00615-C04261	R02738, R04394, R05132	1	0.79	0.79	frag
FN	C00022-C05125	R00014	0.5	0.10	0.1	quant
	C00024-C00033	R00235, R00393, R01179, R01323, R07832	0.06	0.75	0.06	quant
	C00024-C16255	R02569	0.06	0.19	0.06	quant
	C05125-C16255	R03270	0.1	0.19	0.1	quant
	C15972-C00068	R03270, R03316	0	0	0	annot

Table 2: **Pruning of *SCL* on KEGG pathways.** The optimal range is found by minimizing the total false connections, i.e., false positive + false negative. Note that this choice of optimality criterion is arbitrary and not necessarily a good one. F: total false connections; FP: false positive connections; FN: false negative connections; T: true connections; all: all connections. The last two rows are from global pathway combinations.

pathwayID	KEGG title	start	end	F	FP	FN	T	all
eco00010	Glycolysis / Gluconeogenesis	0.380	0.500	22	12	10	66	180
eco00020	Citrate cycle (TCA cycle)	0.510	0.600	28	10	18	50	130
eco00030	Pentose phosphate pathway	0.210	0.330	18	12	6	66	140
eco00040	Fructose and glucuronate interconversions	0.380	0.620	8	6	2	58	156
eco00051	Fructose and mannose metabolism	0.220	0.370	16	16	0	54	142
eco00052	Galactose metabolism	0.220	0.440	16	14	2	74	153
eco00053	Ascorbate and aldarate metabolism	0.220	0.460	6	6	0	28	72
eco00061	Fatty acid biosynthesis	0.190	0.310	26	16	10	80	236
eco00071	Fatty acid metabolism	0.020	0.080	8	8	0	80	170
eco00130	Ubiquinone and other terpenoid-quinone biosynthesis	0.380	0.620	8	6	2	14	52
eco00230	Purine metabolism	0.480	0.500	68	32	36	178	620
eco00240	Pyrimidine metabolism	0.410	0.450	54	24	30	114	356
eco00250	Alanine, aspartate and glutamate metabolism	0.750	0.750	32	12	20	38	220
eco00260	Glycine, serine and threonine metabolism	0.530	0.600	32	14	18	70	218
eco00270	Cysteine and methionine metabolism	0.460	0.460	44	34	10	66	268
eco00280	Valine, leucine and isoleucine degradation	0.110	0.850	8	6	2	28	78
eco00281	Geraniol degradation	0.900	0.980	6	2	4	8	30
eco00290	Valine, leucine and isoleucine biosynthesis	0.870	0.870	28	10	18	46	174
eco00300	Lysine biosynthesis	0.420	0.580	14	12	2	24	118
eco00310	Lysine degradation	0.620	0.980	10	6	4	12	40
eco00330	Arginine and proline metabolism	0.380	0.400	36	24	12	88	318
eco00340	Histidine metabolism	0.410	0.710	8	6	2	20	74
eco00350	Tyrosine metabolism	0.130	0.870	6	6	0	10	62
eco00360	Phenylalanine metabolism	0.300	0.330	12	10	2	26	124
eco00361	gamma-Hexachlorocyclohexane degradation	0.290	0.710	0	0	0	6	16
eco00362	Benzoate degradation via hydroxylation	0.120	0.120	2	2	0	10	16
eco00364	Fluorobenzoate degradation	0.100	0.900	0	0	0	4	12
eco00380	Tryptophan metabolism	0.860	0.980	16	6	10	14	86
eco00400	Phenylalanine, tyrosine and tryptophan biosynthesis	0.410	0.430	16	14	2	48	154
eco00401	Novobiocin biosynthesis	0.190	0.920	4	4	0	4	24
eco00410	beta-Alanine metabolism	0.300	0.660	14	10	4	20	90
eco00430	Taurine and hypotaurine metabolism	0.340	1.000	14	0	14	14	56
eco00440	Phosphonate and phosphinate metabolism	0.210	1.000	2	2	0	2	8
eco00450	Selenoamino acid metabolism	0.850	0.850	26	2	24	30	112
eco00460	Cyanoamino acid metabolism	0.710	0.710	6	2	4	8	56
eco00471	D-Glutamine and D-glutamate metabolism	0.130	0.170	2	2	0	12	48
eco00473	D-Alanine metabolism	0.130	1.000	2	2	0	4	14
eco00480	Glutathione metabolism	0.240	0.340	24	16	8	52	156
eco00500	Starch and sucrose metabolism	0.320	0.410	31	20	11	63	178
eco00520	Amino sugar and nucleotide sugar metabolism	0.650	0.710	27	12	15	69	216
eco00521	Streptomycin biosynthesis	0.320	0.750	6	4	2	14	44
eco00523	Polyketide sugar unit biosynthesis	0.320	0.970	4	2	2	8	26
eco00540	Lipopolysaccharide biosynthesis	0.410	0.440	6	6	0	16	60
eco00550	Peptidoglycan biosynthesis	0.210	0.260	18	8	10	44	134
eco00561	Glycerolipid metabolism	0.580	0.600	10	6	4	20	62
eco00562	Inositol phosphate metabolism	0.260	0.750	0	0	0	10	24
eco00564	Glycerophospholipid metabolism	0.410	0.410	18	12	6	46	142
eco00590	Arachidonic acid metabolism	0.510	0.790	0	0	0	6	28
eco00592	alpha-Linolenic acid metabolism	0.870	1.000	8	0	8	8	28
eco00600	Sphingolipid metabolism	0.270	0.760	0	0	0	6	20
eco00620	Pyruvate metabolism	0.810	0.810	52	12	40	66	260
eco00621	Biphenyl degradation	0.120	0.330	0	0	0	6	8
eco00622	Toluene and xylene degradation	0.120	0.330	0	0	0	8	12
eco00624	1- and 2-Methylnaphthalene degradation	0.010	1.000	2	2	0	4	20
eco00627	1,4-Dichlorobenzene degradation	0.130	0.330	0	0	0	12	20
eco00628	Fluorene degradation	0.120	0.330	0	0	0	6	8
eco00629	Carbazole degradation	0.120	0.330	0	0	0	6	8
eco00630	Glyoxylate and dicarboxylate metabolism	0.340	0.450	24	16	8	58	172
eco00632	Benzoate degradation via CoA ligation	0.310	0.980	8	6	2	12	44
eco00633	Trinitrotoluene degradation	0.000	0.910	0	0	0	2	2
eco00640	Propanoate metabolism	0.900	0.940	24	2	22	32	108
eco00641	3-Chloroacrylic acid degradation	0.010	1.000	2	2	0	4	20
eco00642	Ethylbenzene degradation	0.120	0.130	2	2	0	10	16
eco00643	Styrene degradation	0.120	0.330	2	2	0	8	26
eco00650	Butanoate metabolism	0.510	0.690	26	16	10	30	116
eco00660	C5-Branched dibasic acid metabolism	0.880	0.890	4	0	4	6	26
eco00670	One carbon pool by folate	0.920	0.960	10	8	2	22	118
eco00680	Methane metabolism	0.970	1.000	12	4	8	16	60
eco00730	Thiamine metabolism	0.210	0.590	8	6	2	18	54
eco00740	Riboflavin metabolism	0.420	0.580	6	4	2	20	80
eco00750	Vitamin B6 metabolism	0.440	0.560	8	8	0	26	118
eco00760	Nicotinate and nicotinamide metabolism	0.710	0.750	22	8	14	28	126
eco00770	Pantothenate and CoA biosynthesis	0.340	0.560	24	10	14	46	136
eco00780	Biotin metabolism	0.830	0.920	10	2	8	12	84
eco00785	Lipoic acid metabolism	0.680	0.810	4	0	4	12	56
eco00790	Folate biosynthesis	0.410	0.560	15	10	5	39	133
eco00860	Porphyrin and chlorophyll metabolism	0.710	0.780	28	12	16	56	238
eco00900	Terpenoid backbone biosynthesis	0.650	0.730	24	8	16	34	112
eco00903	Limonene and pinene degradation	0.020	0.980	0	0	0	4	6
eco00910	Nitrogen metabolism	0.980	1.000	42	6	36	36	220
eco00920	Sulfur metabolism	0.360	1.000	28	4	24	24	90
eco00930	Caprolactam degradation	0.020	0.980	2	2	0	4	16
eco00970	Aminoacyl-tRNA biosynthesis	0.030	0.030	8	6	2	94	310
eco00980	Metabolism of xenobiotics by cytochrome P450	0.910	1.000	28	2	26	26	102
eco01053	Biosynthesis of siderophore group nonribosomal peptides	0.380	0.680	2	2	0	6	22
eco01100	Metabolic pathways	0.380	0.390	452	250	202	1200	3555
eco01110	Biosynthesis of secondary metabolites	0.390	0.430	132	80	52	420	1272

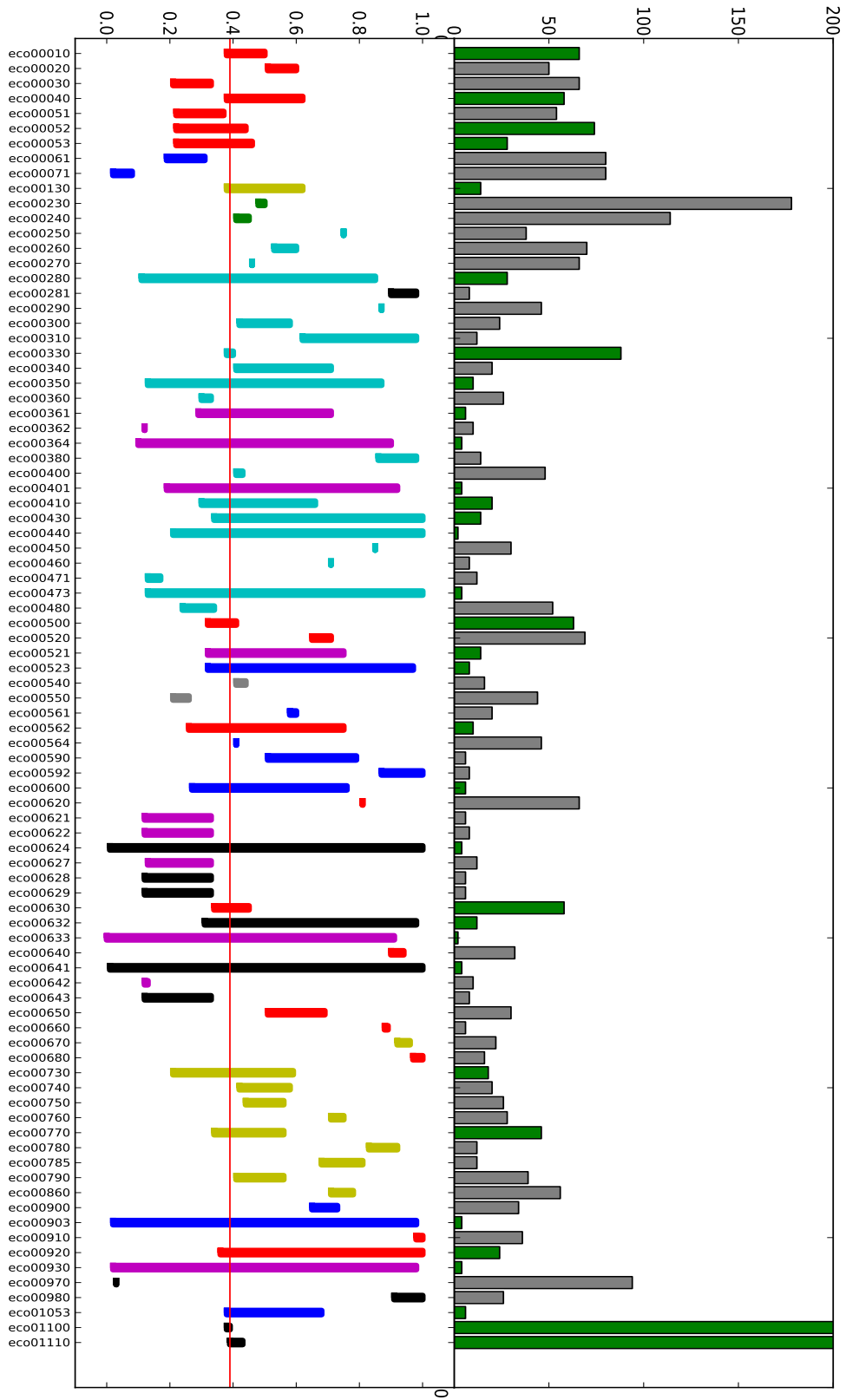


Figure 2: Optimal thresholds for KEGG pathways. Detailed description see text below.

For Fig. 2, **Left panel:** The color code is Lipid Metabolism: blue; Metabolism of Terpenoids and Polyketides: blue; Nucleotide Metabolism: green; Energy Metabolism: red; Carbohydrate Metabolism: red, Amino Acid Metabolism: cyan; Metabolism of Other Amino Acids: cyan; Biosynthesis of Other Secondary Metabolites: magenta; Xenobiotics Biodegradation and Metabolism: magenta; Metabolism of Cofactors and Vitamins: yellow; Glycan Biosynthesis and Metabolism: gray; Others: black. The red vertical line indicates threshold value 0.39. **Right panel:** the number of connections that exist in each of the KEGG pathways. Green facecolor indicates that its range of optimal thresholds (as shown in the left panel) includes interval $[0.38, 0.39]$. Two pathway combinations are also included (the last two row).

Table 3: **Connected components of the network pruned by $SCL \geq t = 0.4$.** The reaction set is restricted to pathway eco00010. The ‘type’ column refers to whether the connections between these metabolites appear in the reference pathway. FP indicates absence and TP indicates presence (all the categories listed here have all their connections classified as either FP or TP.). Many reactant pairs identified as FP in Table 1 are fragments in terms of connectivity. The reference pathway is not completely connected, as is indicated by some fragments labeled as TP.

size	compounds	type
2	C00004, C00003	FP
2	C00084, C00469	TP
2	C00068, C05125	TP
2	C04261, C00615	FP
2	C00024, C00010	FP
3	C16255, C15972, C15973	TP
3	C00020, C00002, C00008	FP
21	C01172, C00221, C00267, C00103, C00668, C00031, C05345, C05378, C00111, C00118, C00236, C00197, C00631, C00074, C00036, C00022, C00011, C06188, C01451, C06187, C06186	TP

strategies	1	2	3	4	5	6
max length	5	5	5	5	8	8
prune1		✓		✓	✓	✓
prune2			✓	✓		✓
total visits	41173	1293	13750	1293	55208	24874
result set	1382	9	467	9	976	351
occurence	122	1	15	1	1	1
rank by length	15	1	15	1	1	1
ties on rank	124	1	124	0	1	1
rank by our criteria	1	1	1	1	1	1
ties on rank	0	0	0	0	0	0

Figure 3: **Comparisons of pruning strategies on Oxidative Pentose Phosphate pathway** Even no prune case is based on rpair existence. rank is the highest among ties. prune 1 is prune on extension by SCL_{self} . prune 2 is prune on exploration by min SCL.

pathway name	source	sink	main metabolites	reactions (in KEGG Rnumber)
Arginine Catabolism	N2-Succinyl-L-arginine	N-Succinyl-L-glutamate	N2-Succinyl-L-ornithine, N-Succinyl-glutamate-5-semialdehyde	R04189, R04217, R05049
Arginine Utilization	Agmatine	Succinate semialdehyde	Putrescine, 4-Aminobutanal, 4-Aminobutanoate	R01157, R01155, R02549, R01648
Chorismate Biosynthesis	3-Dydroquinate	5-O-(1-Carboxyvinyl)-3-phosphoshikimate	3-Dehydroshikimate, Shikimate, Shikimate 3-phosphate	R03084, R02413, R02412, R03460
D-Glucuronate Catabolism	D-Glucuronate	2-Dehydro-3-deoxy-D-gluconate	D-Fructuronate, D-Mannose	R01482, R02454, R05606
Lysine Biosynthesis	4-Phospho-L-aspartate	meso-2,6-Diaminoheptanedioate	L-Aspartate 4-semialdehyde, L-2,3-Dihydrodipicolinate, 2,3,4,5-Tetrahydrodipicolinate, N-Succinyl-2-L-amino-6-oxoheptanedioate, N-Succinyl-LL-2,6-diaminoheptanedioate, LL-2,6-Diaminoheptanedioate, meso-2,6-Diaminoheptanedioate	R02291, R02292, R04198, R04365, R04475, R02734, R02735
Threonine Biosynthesis	4-Phospho-L-aspartate	O-Phospho-L-homoserine	L-Aspartate 4-semialdehyde, L-Homoserine	R02291, R01775, R01771
Oxidative Pentose Phosphate Pathway	D-Glucose-6-phosphate	D-Ribose-5-Phosphate	D-Glucono-1,5-lactone-6-phosphate, 6-Phospho-D-gluconate, D-Ribulose 5-phosphate	R00835, R02035, R01528, R01056
Glycolysis	beta-D-Fructose 6-phosphate	Phosphoenolpyruvate	beta-D-Fructose-1,6-bisphosphate, (2R)-2-Hydroxy-3-(phosphonoxy)propanal, 3-Phospho-D-glyceroyl phosphate, 3-Phospho-D-glycerate, 2-Phospho-D-glycerate	R04779, R01070, R01061, R01512, R01518, R00658
Methionine Biosynthesis	4-Phospho-L-aspartate	L-Homocysteine	L-Aspartate-4-semialdehyde, L-Homoserine, O-Succinyl-L-homoserine, Cystathionine	R02291, R01775, R01777, R02508, R01285
TCA cycle	NA	NA	Citrate, cis-Aconitate, Isocitrate, 2-Oxoglutarate, Succinyl-CoA, Succinate, Fumarate, L-Aspartate, (S)-Malate, Oxaloacetate	R01325, R01324, R01900, R00267, R00479, R05320, R01197, R00405, R00408, R00490, R01082, R00342, R00355, R00351

Figure 4: **Reference pathway details.** The reference pathways are extracted from aMAZE database [2] except Oxidative Pentose Phosphate Pathway and TCA cycle which are manually extracted from KEGG pathway map [3].

strategies	1	2	3	4
max length	6	6	6	6
prune1		✓		✓
prune2			✓	✓
total visits	> 1.3E5	11479	25862	11479
result set	> 4000	235	876	235
occurence	NA	74	569	65
ranked in length	NA	65	568	65
ties on rank	NA	170	308	170
ranked in <i>SCL</i>	NA	2	2	2
ties on rank	NA	0	0	0

Figure 5: **Comparisons of pruning strategies on Glycolysis pathway** Even no prune case is based on rpair existence. rank is the highest among ties. prune 1 is prune on extension by SCL_{self} . prune 2 is prune on exploration by min SCL.

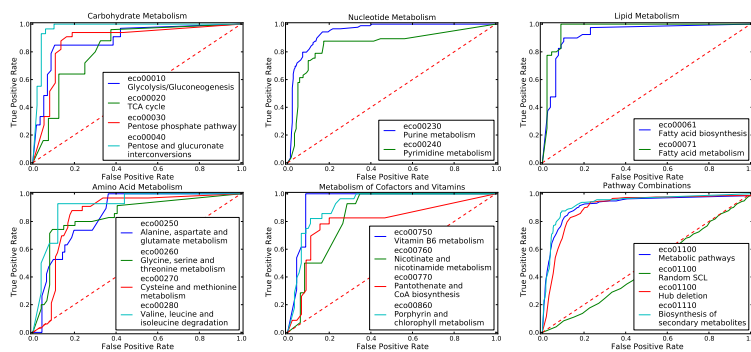


Figure 6: ROC curves based on thresholding the *SCL* values. Each curve is based on the complete set of edges between any two metabolites sitting on opposite sides of any reaction in that particular pathway and on 50 threshold values evenly distributed in the range $[0, 1]$.

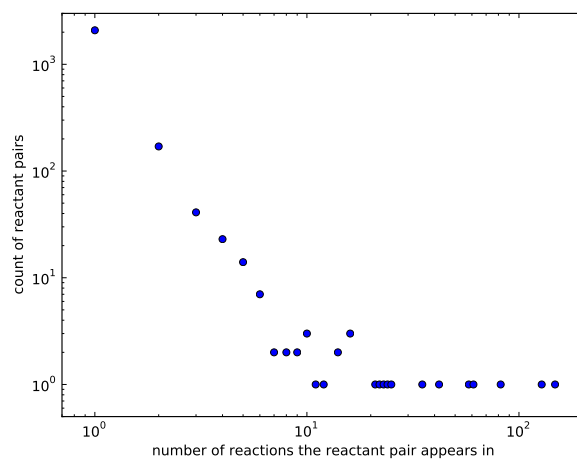


Figure 7: **Usage of reactant pairs as carriers.** The log-log plot of the count of reactant pairs that participate in a certain number of reactions. The reaction set is restricted to the entire *E.coli* network. Reactant pairs are only those with clear RPair annotation in the KEGG RPair database.

Table 4: Common Biochemical Carriers

class	KEGG ID	names
Proton/Electron Flux	(C00003,C00004)	(NADH,NAD ⁺)
	(C00005,C00006)	(NADPH,NADP ⁺)
	(C00016,C01352)	(FAD,FADH ₂)
	(C01359,C00113)	(PQQH ₂ ,PQQ)
	(C01847,C00061)	(reduced FMN,FMN)
	(C00026,C00042)	(2-Oxoglutarate,Succinate)
	(C00125,C00126)	(Ferricytochrome c,Ferrocycytochrome c)
	(C01119,C00265)	(Oxidized dithiothreitol,Dithiothreitol)
	(C00139,C00138)	(Oxidized ferredoxin,Reduced ferredoxin)
	(C00662,C00667)	(Reduced adrenal ferredoxin,Oxidized adrenal ferredoxin)
	(C00272,C00268)	(Tetrahydrobiopterin,Dihydrobiopterin)
	(C03024,C03161)	(Reduced flavoprotein,Oxidized flavoprotein)
	(C00340,C00435)	(Reduced rubredoxin,Oxidized rubredoxin)
	(C04253,C04570)	(e ⁻ transferr flavoprotein,Reduced e ⁻ transfer flavoprotein)
	(C01070,C01071)	(Ferricytochrome c-553, Ferrocycytochrome c-553)
(C06259,C06260)	(Ferrocycytochrome b,Ferricytochrome b)	
(C00999,C00996)	(Ferrocycytochrome b5,Ferricytochrome b5)	
One Carbon Unit Flux	(C00019,C00021)	(S-Adenosyl-L-methionine,S-Adenosyl-L-homocysteine)
	(C00155,C00073)	(L-Homocysteine,L-Methionine)
	(C00234,C00101)	(10-Formyltetrahydrofolate, Tetrahydrofolate)
	(C00101,C00143)	(Tetrahydrofolate,5,10-Methylenetetrahydrofolate)
	(C00445,C00101)	(5,10-Methenyltetrahydrofolate,Tetrahydrofolate)
(C00068,C13378)	(Thiamin diphosphate,alpha,beta-Dihydroxyethyl-TPP)	
Two Carbon Unit Flux	(C00024,C00010)	(Acetyl-CoA, CoA)
	(C00227,C00009)	(Acetyl phosphate, orthophosphate)
Acyl Flux	(C00173,C00229)	(Acyl-[acp],Acyl-carrier protein)
Sulphur/Sulfide Flux	(C00053,C00054)	(3'-Phosphoadenylyl sulfate,Adenosine 3',5'-bisphosphate)
	(C00343,C00342)	(Thioredoxin disulfide,Thioredoxin)
	(C00051,C00127)	(Glutathione,Glutathione disulfide)
Amino Group Flux	(C00022,C00041)	(Pyruvate,L-Alanine)
	(C00026,C00025)	(2-Oxoglutarate,L-Glutamate)
	(C00026,C00302)	(2-Oxoglutarate,Glutamate)
	(C00064,C00025)	(L-Glutamine,L-Glutamate)
	(C00647,C00018)	(Pyridoxamine phosphate,Pyridoxal phosphate)
Phosphate/Pyrophos Flux	(C00002,C00008)	(ATP,ADP)
	(C00044,C00035)	(GTP,GDP)
	(C00002,C00020)	(ATP,AMP)
	(C00063,C00055)	(CTP,CMP)
	(C00081,C00104)	(ITP,IDP)
	(C00013,C00009)	(Pyrophosphate,orthophosphate)
	(C00131,C00206)	(dATP,dADP)
Hydroxy Group Flux	(C00272,C15522)	(tetrahydrobiopterin,hydroxytetrahydrobiopterin)
CoA Mediated Acyl Chain Flux	(C00040,C00010)	(Acyl-CoA, CoA)
	(C00083,C00010)	(Malonyl-CoA,CoA)
	(C00223,C00010)	(p-Coumaroyl-CoA,CoA)
NDP Mediated Sugar Flux	(C02199,C00015)	(UDP-L-rhamnose,UDP)
	(C00167,C00015)	(UDP-glucuronate,UDP)
	(C00052,C00015)	(UDP-D-galactose,UDP)
	(C00043,C00015)	(UDP-N-acetyl-D-glucosamine,UDP)
	(C00203,C00015)	(UDP-N-acetyl-D-galactosamine,UDP)
	(C00307,C00055)	(CDP-choline,CMP)
	(C00190,C00015)	(UDP-D-xylose,UDP)
	(C00570,C00055)	(CDP-ehtanolamine,CMP)
	(C00052,C00105)	(UDP-D-galactose,UMP)

References

- [1] H. W. Ma and A. P. Zeng, "Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms," *Bioinformatics*, vol. 19, pp. 270–277, 2003. [Online]. Available: <http://bioinformatics.oxfordjournals.org/cgi/content/abstract/19/2/270>
- [2] C. Lemer, E. Antezana, F. Couche, F. Fays, X. Santolaria, R. Janky, Y. Deville, J. Richelle, and S. J. Wodak, "The aMAZE LightBench: a web interface to a relational database of cellular processes," *Nucl. Acids Res.*, vol. 32, no. suppl, pp. D443–448, 2004. [Online]. Available: http://nar.oxfordjournals.org/cgi/content/abstract/32/suppl_1/D443
- [3] M. Kanehisa and S. Goto, "KEGG: Kyoto Encyclopaedia of Genes and Genomes," *Nucleic Acids Res*, vol. 28, pp. 27–30, 2000.