Supplementary Information

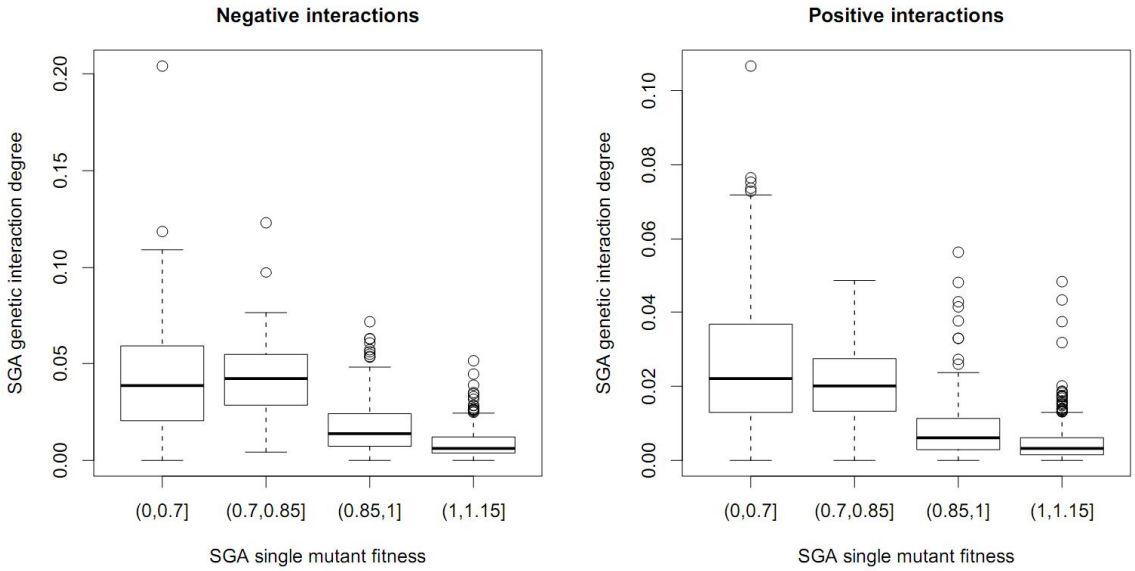
An integrated approach to characterize genetic interaction networks in yeast metabolism

Balázs Szappanos, Károly Kovács, Béla Szamecz, Frantisek Honti, Michael Costanzo, Anastasia Baryshnikova, Gabriel Gelius-Dietrich, Martin J. Lercher, Márk Jelasity, Chad L. Myers, Brenda J. Andrews, Charles Boone, Stephen G. Oliver, Csaba Pál, Balázs Papp

Contents:

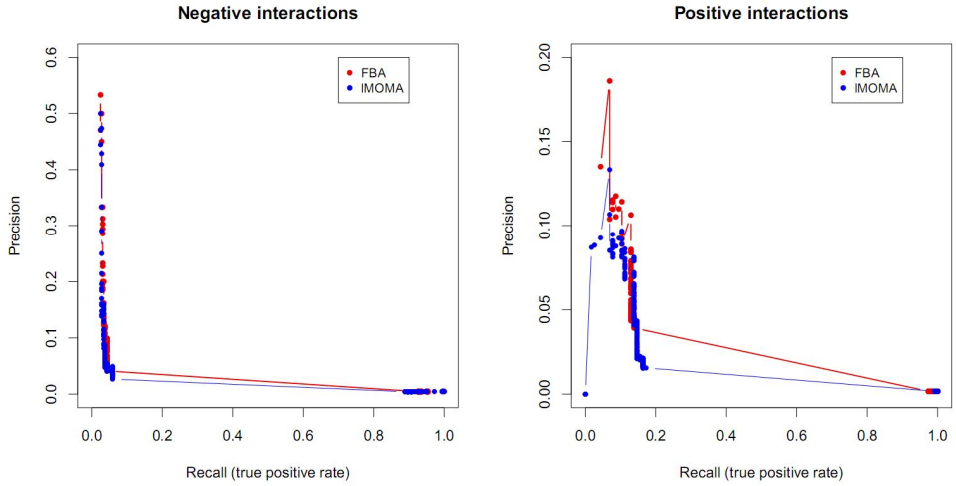
Supplementary Figure 1. Correlation between the fitness effect of a gene's deletion and the number of its genetic interactions.

The empirical genetic interaction degree of a gene negatively correlates with the empirical fitness of its deletion mutant for both negative (Spearman's correlation on raw data: $\rho = -0.47$, $P<10^{-37}$) and positive interactions ($\rho = -0.42$, $P<10^{-28}$). Genetic interaction degree of a gene was defined as the fraction of tested genes that show significant interaction with the gene in question. Genetic interactions and fitness of single mutants were derived from our empirical genetic interaction map. Only null mutants of non-essential genes were included in the analysis.

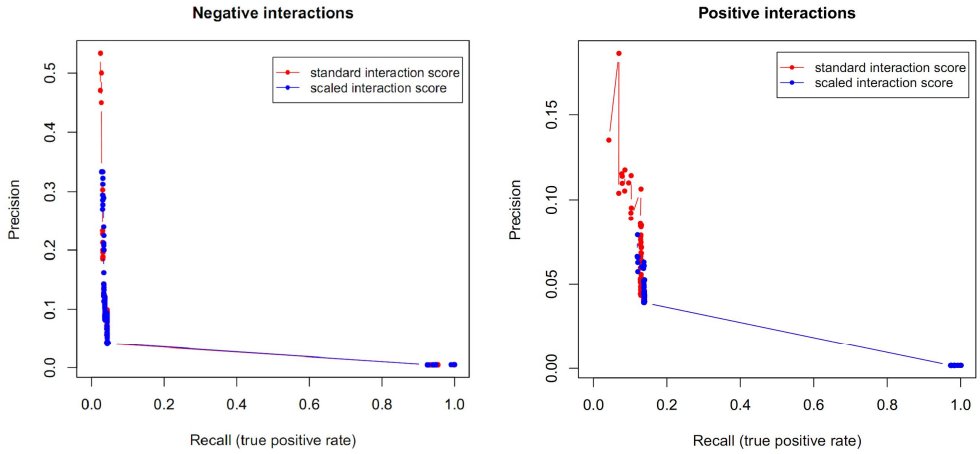


Supplementary Figure 2. Evaluation of computationally predicted genetic interaction scores under various model assumptions and under transcriptional constraints.

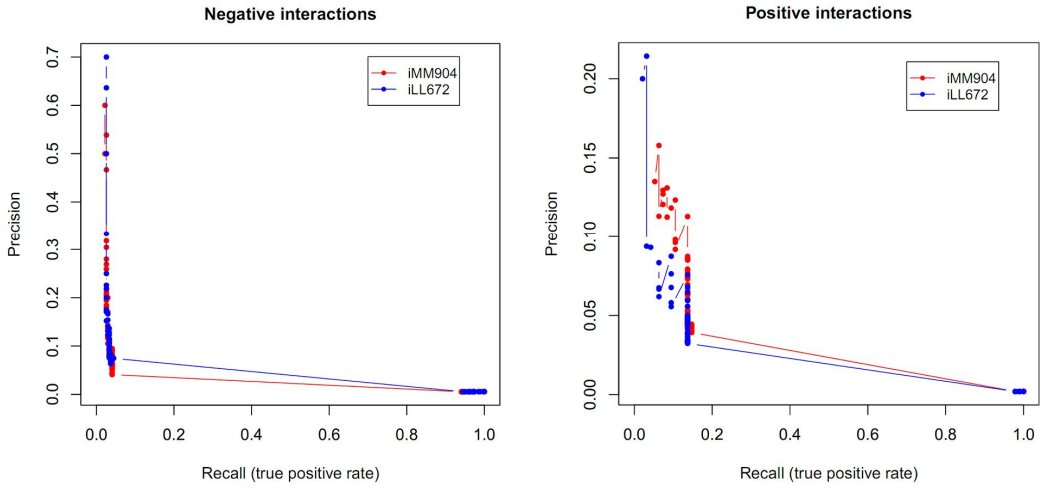
(a) Comparing the predictive performance of FBA and MOMA algorithms. FBA assumes maximal biomass production (fitness), whereas MOMA assumes minimally perturbed flux distributions in mutant strains. The two algorithms show similar performance.



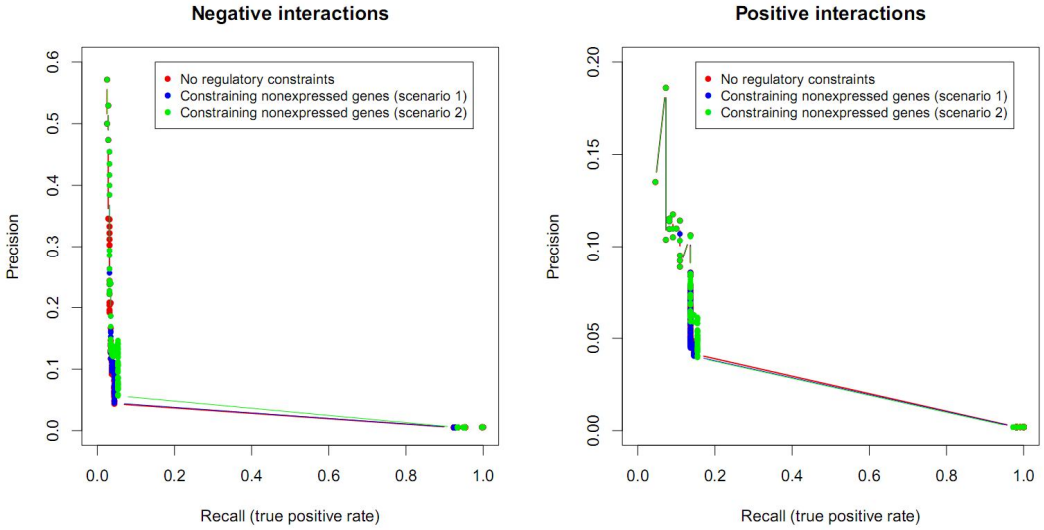
(b) Effect of using an alternative definition of interaction scores to derive *in silico* genetic interactions. Here, we compared the standard definition of genetic interaction score ($\varepsilon = f_{12} - f_1 \cdot f_2$) to a scaled measure that quantifies the relative strength of the interactions[1]. The scaling is based on two natural references. For negative interactions, the reference case is complete synthetic lethality. For positive interactions, the reference is the special case when the mutation with the stronger effect completely masks (alleviates) the effect of the other mutation. Apparently, scaled interaction score results in slightly reduced maximum precision values.



(c) Comparing the predictive performance of different yeast models. iMM904[2] is a more recent version that contains 8 compartments, while iLL672[3] is an earlier reconstruction that contains only 2 compartments. Both models were evaluated on an overlapping set of 46,814 gene pairs. While iLL672 shows slightly higher maximum precisions, there is no difference in recalls.

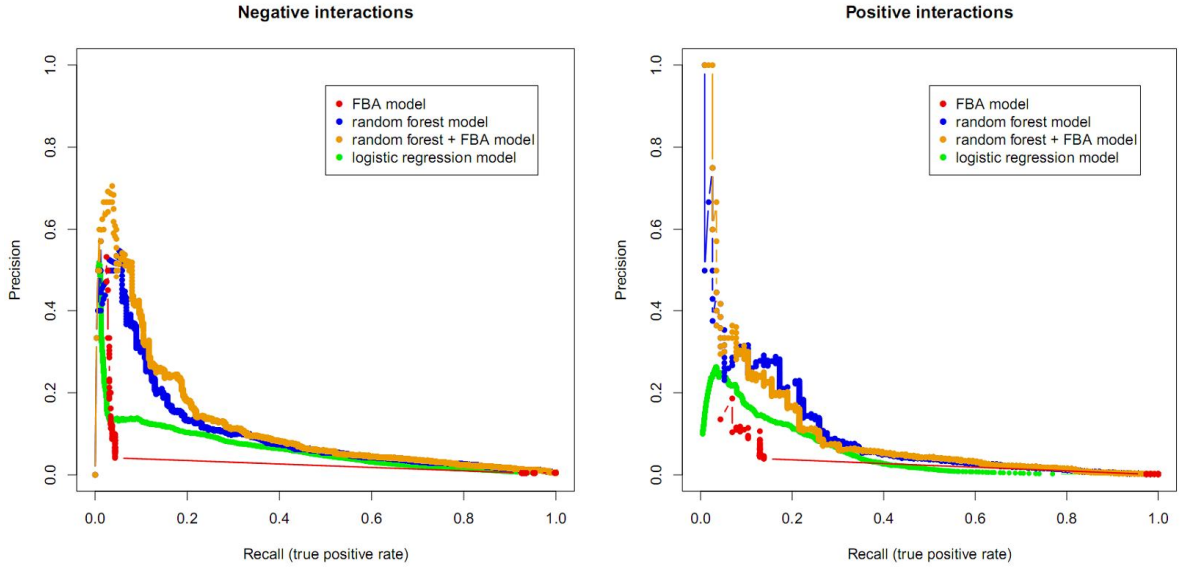


(d) Effect of imposing transcriptional constraints on the predictive performance of FBA. We integrated gene expression data into the FBA framework by inactivating those genes in the model whose expression is reported to be undetectable (tag count=0, Scenario 1) or very low (tag count ≤ 2, Scenario 2) according to a recent quantitative transcriptome map of yeast[4].

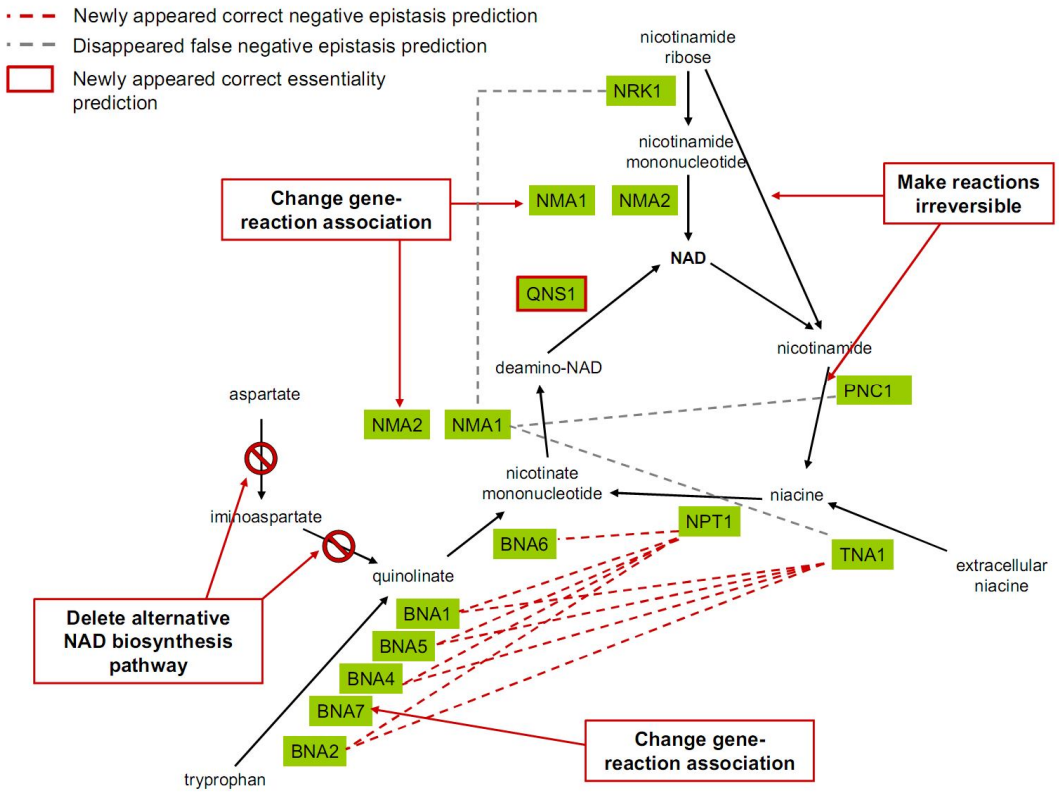


Supplementary Figure 3. Evaluation of statistical modeling approaches to predict genetic interactions in yeast metabolism.

We built classifiers to predict negative and positive genetic interactions based on a large set of gene-pair characteristics (see Supplementary Note) using both logistic regression and random forest[5] models as implemented in the R statistical environment[6,7]. Only high-confidence genetic interactions between null mutants of non-essential genes were included in the analysis (325 negative and 116 positive interactions among 67,517 gene pairs, see Online Methods for definition). Random forest is a recently proposed non-parametric method that uses an ensemble of classification trees created by using bootstrap samples of the training data and random feature selection in tree building[5]. Here, we employed balanced random forest classifiers to deal with the strong imbalance in the number of interacting and non-interacting gene pairs and built 5000 trees. Prediction success was evaluated using out-of-bag samples for random forest and 5-fold cross-validation repeated 10 times for logistic regression analyses, respectively. Precision – recall plots were prepared using the visualization package ROCR[8]. 'FBA model' denotes the predictive performance of FBA-derived genetic interaction scores, while 'random forest + FBA' corresponds to a random forest model that has been built by incorporating FBA-derived fitness and genetic interaction scores as predictors in addition to gene-pair characteristics listed in Supplementary Note.



Supplementary Figure 4. Revised reconstruction of NAD biosynthesis.
Proposed model changes and their cumulative effects on negative genetic interaction and gene essentiality predictions: i) remove the *de novo* aspartate pathway, ii) make purine-nucleoside phosphorylase and nicotinamidase[9] reactions irreversible, iii) change gene-reaction associations by assigning the same set of reactions to *NMA1* and *NMA2*, which are isoenzymes[10], iv) the transformation of formylkynurenine to kynurenine is coded by *BNA7*, not by *BNA3*, based on ref.[11]. Altogether, these modifications introduce 9 correctly predicted cases of negative interactions, remove 3 incorrectly predicted negative interactions and introduce 1 correctly predicted conditional lethal single mutant phenotype. Additionally, some modifications introduced discrepancies that were resolved by other modifications and are therefore not shown. An updated metabolic reconstruction incorporating these modifications can be downloaded in Systems Biology Markup Language[12] format from http://www.utoronto.ca/boonelab/data/szappanos/.



Supplementary Table 1. Monochromaticity of genetic interactions between functional annotation groups.

Results of the monochromaticity analysis as described in the main text. Genes assigned to multiple functional annotation groups were excluded from the analysis. The degree of monochromaticity was measured with a monochromatic score (MC, see Methods of the main text), and a pair of functional groups was considered monochromatic if $|MC_{ij}| > 0.5$. Significance level of the observed monochromaticity was assessed by randomizations (see Online Methods). Because pairs of annotation groups that show only one genetic interaction would always be monochromatic, we restricted our analysis to those functional group pairs that show at least 2 and 3 genetic interactions between each other, respectively.

	Functional group pairs showing at least 2 genetic interactions	**Functional group pairs showing at least 3 genetic interactions**
Total number of functional group pairs investigated	451	339
Number of monochromatic functional group pairs in the real data	177	111
Mean number and standard deviation of monochromatic group pairs expected based on randomized genetic interaction networks	142.87 (8.70)	82.75 (7.15)
Relative excess of monochromatic group pairs in the real genetic interaction network compared to randomized networks	23.89%	34.13%
Significance level of observed monochromaticity	$P < 10^{-4}$	$P < 10^{-4}$
Background ratio of positive to all interactions (*bpr*)	0.341	0.337

Results using the high-confidence genetic interaction dataset:

	Functional group pairs showing at least 2 genetic interactions
Total number of functional group pairs investigated	17
Number of monochromatic functional group pairs in the real data	10
Mean number and standard deviation of monochromatic group pairs expected based on randomized genetic interaction networks	6.01 (1.83)
Relative excess of monochromatic group pairs in the real genetic interaction network compared to randomized networks	66.39%
Significance level of observed monochromaticity	$P = 0.03$
Background ratio of positive to all interactions (*bpr*)	0.415

Supplementary Table 2. Partial correlation analysis of FBA predicted genetic interaction degree, single mutant fitness and pleiotropy.

We used partial correlation analyses to disentangle the effects of predicted single mutant fitness and pleiotropy on *in silico* genetic interaction degrees. To do this, we calculated Spearman's partial rank correlation coefficients between a predictor and a response variable while controlling the effect of a third variable using the pcor.test function implemented in R (http://www.yilab.gatech.edu/pcor.html). To avoid spurious correlations stemming from the fact that for genes associated with silent reactions the model would predict i) zero fitness contribution, ii) lack of genetic interactions and iii) zero pleiotropy, we focused only on those genes whose removal has non-zero fitness effect. Furthermore, some sets of genes would always produce identical phenotypes in the model simulations and cannot be treated as independent data points in statistical analyses (*e.g.* genes encoding flux coupled reactions or subunits of the same protein complex). To avoid such a bias in our analysis, we represented each correlated gene set with one randomly chosen gene. These filtering procedures resulted in 34 genes.

	Negative interaction degree		**Positive interaction degree**	
	Spearman's partial ρ	*P*-value	Spearman's partial ρ	*P*-value
Correlation between interaction degree and pleiotropy while controlling the effect of fitness	**0.447**	**0.005**	-0.112	0.531
Correlation between interaction degree and fitness while controlling the effect of pleiotropy	0.089	0.620	**-0.451**	**0.005**

Supplementary Table 3. Metabolic model modifications suggested by the automated model refinement method.

List of individual, and combinations of related model modifications that were suggested by the automated model refinement algorithm. Few modifications clearly contradict literature data and are therefore not shown. While the algorithm did not evaluate positive genetic interactions during the optimization process, we note that none of the listed model changes decreased the prediction accuracy of positive interactions.

Affected reaction(s) or biomass component[a]	Suggested modification	Improved predictions[b]		Impaired predictions		Description / Explanation	Frequency[c]	MCC increment[d]	
		Negative interaction	Single mutant	Negative interaction	Single mutant			Negative interaction	Single mutant
ASPOcm, QULNS, NMNAT, RNMK	Deletion (ASPOcm or QULNS) and (NMNAT or RNMK)	9 correctly predicted negative interaction appeared (TNA1–BNA1, TNA1–BNA2, TNA1–BNA4, TNA1–BNA5, NPT1–BNA1, NPT1–BNA2, NPT1–BNA4, NPT1–BNA5, NPT1–BNA6) and 3 erroneously predicted negative interaction disappeared (NMA1–TNA1, NMA1–PNC1, NMA1–NRK1)	1 correctly predicted essential phenotype appeared (QNS1)	1 correctly predicted negative interaction disappeared (NMA1–NMA2) and 2 erroneously predicted negative interaction appeared (TNA1–BNA3, NPT1–BNA3)	1 erroneously predicted essential phenotype appeared (NMA1)	Erroneous presence of an alternative *de novo* NAD biosynthesis in the model; corrupt gene-reaction association and reaction reversibility	8	0.069842	0.000775
ADPT, PUNP1, PUNP5	Deletion of ADPT, disallowing reversibility of PUNP1 and PUNP5	5 correctly predicted negative interaction appeared (HPT1–ADE1, HPT1–ADE2, HPT1–ADE4, HPT1–ADE5,7, HPT1–ADE6)	-	1 erroneously predicted negative interaction appeared (ADE6–AAH1)	-	Alternative pathways for the biosynthesis of AMP[e]	1	0.040612	0
ASPOcm, QULNS	Deletion	4 correctly predicted negative interaction appeared (TNA1–BNA4, TNA1–BNA1, TNA1–BNA2, TNA1–BNA5)	-	1 erroneously predicted negative interaction appeared (TNA1–BNA3)	-	Erroneous presence of an alternative *de novo* NAD biosynthesis in the model	8	0.032538	0
Glycogen	Omission from biomass	2 erroneously predicted negative interaction disappeared (GSY1–GSY2, GLG1–GLG2)	-	-	-	Dispensable biomass component (glycogen not essential for growth)	8	0.006114	0
ACALDtm	Disallowing reversibility	1 correctly predicted negative interaction appeared (ALD6–PDB1)	-	2 erroneously predicted negative interaction appeared (ALD6–ACH1, ALD6–PDX1)	-	Alternative pathway for acetate biosynthesis[e]	5	0.003958	0
Ubiquinone-6	Omission from biomass	1 erroneously predicted negative interaction disappeared (ARO3–ARO4)	5 correctly predicted non-essential phenotype appeared (COQ2, COQ3, COQ5, COQ6, PPA2)	-	-	Dispensable biomass component (ubiquinone-6 is essential only when yeast grows on non-fermentable carbon sources)	4	0.002937	0.032093

[a] Reaction abbreviations: ASPOcm, aspartate oxidase; QULNS, quinolinate synthase; NMNAT, nicotinamide-nucleotide adenylyltransferase; RNMK, ribosylnicotinamide kinase; ADPT, adenine phosphoribosyltransferase; PUNP1, purine-nucleoside phosphorylase (Adenosine); PUNP5, purine-nucleoside phosphorylase (Inosine); ACALDtm, acetaldehyde mitochondrial diffusion.

[b] Gene pairs or genes whose genetic interactions or single mutant fitness are affected by the suggested modification(s) are listed in parentheses.

[c] The number of times the modification appeared in the best hypothesis set out of 8 replicate runs (*i.e.* a frequency of 8 indicates that the modification was present in all best sets).
[d] MCC increment measures the impact of the suggested modification(s) on genetic interaction and single mutant phenotype prediction success. It is defined as the difference between the Matthews Correlation Coefficients of the original and the modified model. The MCC values of the original model on genetic interaction and single mutant phenotype predictions are 0.110720 and 0.498761, respectively.
[e] The mechanism suggested by the model is open to alternative explanations based on literature data and therefore no specific explanation is proposed here.

Supplementary Note

Definition of *in silico* growth medium

Medium composition for the flux balance analysis simulations was defined so as to faithfully mimic experimental conditions under which genetic interactions were screened (*i.e.* a synthetic complete media lacking histidine, arginine and lysine, SD/MSD – His/Arg/Lys, see ref. [13]). Upper bounds on nutrient uptake fluxes were defined following ref. [14] and are given in parentheses in mmol/hour/gram dry weight. Nutrients included were as follows: 4-Aminobenzoate (0.000002), Adenine (3.01), L-Alanine (0.36), L-Asparagine (0.36), L-Aspartate (0.36), Biotin (0.00000142), L-Cysteine (0.36), Fe2+ (1000), D-Glucose (22.6), L-Glutamine (0.36), L-Glutamate (3.6), Glycine (0.36), L-Isoleucine (0.36), myo-Inositol (0.11), K+ (4.44), L-Leucine (1.8), L-Methionine (0.36), Na+ (0.75), Nicotinate (0.000002), O_2 (6.3), L-Phenylalanine (0.36), Phosphate (0.89), (R)-Pantothenate (0.0002), L-Proline (0.36), Riboflavin (0.00092), L-Serine (0.36), Sulfate (100), Thiamin (0.0032), L-Threonine (0.36), L-Tryptophan (0.36), L-Tyrosine (0.36), Uracil (3.63), L-Valine (0.36)

Description of gene-pair features employed for probabilistic prediction of genetic interactions.

Gene-pair characteristics were compiled following earlier studies[15,16], however, features incorporating information on genetic interactions were not included. Pair-wise features that are characteristic of the metabolic network reconstruction (*e.g.* metabolic annotation groups) were included. Besides simple pair-wise features, we also included so-called "2hop" features. Each 2hop feature captures a specific relationship between a gene pair and a third gene. For instance, if protein A shows physical interaction with protein C, and gene C has a common upstream regulator with gene B then the gene pair A–B is considered as a "2hop PPI – Regulator".

Gene-pair features investigated were as follows: single deletant fitness (average and absolute difference of the two single deletion fitnesses), paralogy (with 3 levels: no sequence similarity, paralogs with gene family size of 2, paralogs with gene family size of >2), shortest path in the metabolic network, sharing a metabolic annotation group[2], occurrence in a specific metabolic annotation group[2] (30 features defined corresponding to 30 annotation groups with at least four genes from our genetic interaction dataset), local sub-graphs in the metabolic network (5 levels defined based on ref. [17]: non-adjacent, chains, forks, OR funnel, AND funnel), flux coupling (3 levels: uncoupled, directionally or fully coupled), co-occurrence in any manually curated protein complex[19], occurrence in a specific manually curated protein complex (5 features defined based on 5 protein complexes with at least four genes from our genetic interaction dataset), any physical interaction according to BioGrid[20], PPI network degree[20] (average degree and absolute difference between degrees), shortest path in the PPI network[20], mutual clustering coefficient in the PPI network[20], sharing an upstream regulator[22], 2hop PPI – PPI, 2hop PPI – Regulator, 2hop PPI – Paralogy, 2hop Regulator – Paralogy, mRNA expression correlation[23], sharing a MIPS phenotype[24], association with a specific MIPS phenotype[24] (14

features corresponding to 14 MIPS phenotypes with at least four genes from our genetic interaction dataset), quantitative phenotype correlation[25], sharing subcellular localization (based on GO slim)[26,27] and occurrence in a specific subcellular compartment[26,27] (13 features corresponding to 13 compartments with at least four genes from our genetic interaction dataset).

Automated model refinement

We introduce a machine-learning method to automatically generate hypotheses that improve prediction of compensating (negative) interactions between genes. The method is based on a genetic algorithm to improve the fit between empirical data and genetic interaction predictions. Genetic algorithms comprise a heuristic search method inspired by the theory of natural selection to find solutions to optimization problems by evolving a population of candidate solutions ("individuals") toward better solutions[28]. Similar algorithms have been successfully applied in metabolic engineering to identify *in silico* gene deletions that maximize the production of a desired metabolite[29]. Our implementation of the algorithm starts with a population of randomly modified ("mutated") models, evaluates the performance ("fitness") of each model in the population and selects those with high prediction accuracy to form a new population. Because evaluating each model is computationally intensive (*i.e.* a large number of gene deletions should be simulated for each individual model), we employed a two-step procedure to make use of all available phenotypic data while maintaining computational feasibility. In the first step, we searched for models by evaluating a model on only those gene pairs that display either *in vivo* interaction or *in silico* interaction according to the original iMM904 model[2]. Because genetic interactions are very rare both *in vivo* and *in silico*, most gene pairs examined in this study show no interaction and omitting them significantly speeds up the exploration of the hypothesis space. In the second step, we defined a new, very restricted hypothesis space based on the most successful models from the first step, but selected for models that improve overall prediction accuracy as assessed by a comprehensive evaluation of each model in the population (*i.e.* the genetic algorithm maximized a combined measure of prediction success that takes into account true and false positives and negatives of all genetic interaction and single mutant viability predictions). We present details of the method below.

Metabolic model evaluation

To evaluate the performance of each model (in both the first and second steps), we first computed *in silico* genetic interaction scores (ε) using flux balance analysis, converted them into binary values (interaction / no interaction) and compared them to experimental data. We focused on strong negative genetic interaction predictions only and applied a threshold of $\varepsilon < -0.5$ to define *in silico* interacting pairs as the original model showed sufficiently high precision at this cutoff. The performance of genetic interaction classification was assessed by comparing predictions to our high-confidence genetic interaction dataset. The number of true and false positive and negative predictions for the original iMM904 model is presented below.

		Predicted interaction	
		Negative interaction	No interaction
Empirical interaction	Negative interaction	True Positive (TP) 9	False Negative (FN) 316
	No interaction	False Positive (FP) 11	True Negative (TN) 67,181

In the first round of genetic algorithm optimization, we evaluated each model on a small set of gene pairs that were either interacting *in vivo* or showed *in silico* interactions according to the original model. Model changes that either increased the number of true positives or decreased the number of false positives were preferred by the optimization algorithm. The performance of each model was defined by the following fitness function:

$$F_{step1} = TP - FP$$

In the second round of optimization, each model in the population was evaluated using all available phenotype measurements. We employed a combined measure of prediction accuracy that takes into account both true and false positives and negatives in a balanced manner (MCC score, Matthews correlation coefficient, see ref. [30]). Our overall prediction accuracy score was derived by averaging the MCC score of genetic interaction predictions (based on 67,517 gene pairs) and the MCC score of single mutant viability predictions (566 single gene deletion phenotypes on YPD medium) resulting in the following fitness function:

$$F_{step2} = [(MCC_{interactions} + MCC_{single\_mutants}) / 2] - \pi$$

where $MCC_{interactions}$ and $MCC_{single\_mutants}$ denote the Matthews correlation coefficient of the model predictions on the genetic interaction and single deletion phenotype datasets, respectively, and π is a penalty (see below).

Model representation

Two types of metabolic network parameters were allowed to change during the optimization following a related protocol by Kumar & Maranas[31]. First, we altered the list of biomass compounds that are required for *in silico* growth. As a number of compounds are undoubtedly required for growth (e.g., amino acids), we only modified the presence of those biomass components that show variation among different versions of the yeast FBA model (i.e., present in some, but not all of four published models; iFF708[32], iND750[33], iLL672[3], iMM904[2]). The list of these 21 modifiable biomass components is presented below.

List of modifiable biomass compounds

Ergosterol

Glycogen

Trehalose

cAMP

Chitin

Coenzyme A

Flavin adenine dinucleotide

Reduced glutathione

Protoheme

Mannose-(inositol-P)2-ceramide, ceramide-1 (24C), yeast-specific

Mannose-(inositol-P)2-ceramide, ceramide-1 (26C), yeast-specific

Mannose-(inositol-P)2-ceramide, ceramide-2 (24C), yeast-specific

Mannose-(inositol-P)2-ceramide, ceramide-2 (26C), yeast-specific

Mannose-(inositol-P)2-ceramide, ceramide-3 (24C), yeast-specific

Mannose-(inositol-P)2-ceramide, ceramide-3 (26C), yeast-specific

Nicotinamide adenine dinucleotide

Riboflavin

5,6,7,8-Tetrahydrofolate

Thiamin triphosphate

Ubiquinone-6

Sulfate

Second, we allowed changing reaction reversibility (e.g. reversible to irreversible) and deletion of reactions. Addition of reactions was not considered here because the model typically misses *in vivo* negative interactions due to over-prediction of double-mutant fitness, an issue that cannot be remedied by increasing the size, and potentially the redundancy, of the metabolic network.

Reactions essential *in silico*, reactions belonging to experimentally identified essential genes[14,34] and blocked reactions[18] were not modified during the optimization process. Furthermore, to reduce computational cost, we represented fully coupled reaction sets[18] as a single reaction in the parameter space, resulting in 454 modifiable reactions. Because reversible reactions could be set to irreversible in either direction, changes to these reactions were represented by two binary parameters. Overall, the hypothesis space consisted of 21 binary parameters representing biomass modifications and 615 binary parameters representing reaction deletions and reversibility changes.

Details of the genetic algorithm

Our algorithm evolves a population of individuals, with each individual represented by an array of binary parameters and each parameter encoding a certain modification of the original metabolic model as described above. State 1 of a parameter means the modification is applied, while state 0 means the original model is not affected by this parameter. Individuals of the starting population were generated by setting the majority of binary parameters to 0 and the remaining few parameters to 1. We first determined the number of these parameters by

drawing a random number from a Poisson distribution with $\lambda = 1$ for each individual. The drawn number was increased by one to ensure that we have at least one parameter with value 1. We then selected this number of parameters at random and set these to 1, and set the rest to 0. A fitness value was calculated for each individual in each generation by evaluating the encoded modified model on experimental data as described above. Individuals of the next generation were created by applying one of the following operators: elitism, crossing-over, mutation, insertion and deletion. First, individuals with the highest fitness were transferred unaltered to the next generation (elitism). Second, a mating pool was defined in each generation by selecting half of the individuals based on their fitness values. For each individual, the probability of being involved in the mating pool was calculated using a linear ranking procedure[35]. Next, we applied operators on randomly selected parents from the mating pool to create the individuals in the next generation. Each new individual is created using exactly one of these operators. Details of the operators were as follows:

Crossing-over: two parents were selected and for each binary parameter one of the parental states was inherited by the offspring (with 50% – 50% probability).

Mutation: two different parameters, one in state 1 and the other in state 0, were randomly selected in the parent and their values were exchanged (i.e., the total number of 1s and 0s remained unchanged).

Insertion and deletion: the number of parameters in state 1 is increased or decreased, by randomly switching a small number of parameters to the opposite state.

The first round of optimization involved 1000 generations of reproduction with a population size of 100 individuals. Preliminary simulations suggested that a fixed number of 1000 generations is sufficient to reach convergence. The proportions of the individuals created using the operators elitism, crossing-over, mutation, insertion and deletion were 5%, 45%, 40%, 5% and 5%, respectively. At the end of the simulations, a set of candidate modifications was identified by selecting the most frequent modifications (that is, parameters with value 1, observed in more models than the average number a parameter is observed in a model) in the set of individuals of the highest fitness (including those individuals that were present in earlier generations and attained the same fitness value).

These candidate model modifications defined the hypothesis space of the second round of optimization, where a population of 50 individuals was evolved during 20 generations and the fitness of each modified model was evaluated using all available phenotypic data. This resulted in a substantially reduced hypothesis space to explore, with typically less than 50 binary parameters. Notwithstanding the reduced hypothesis space, the second round of optimization still has a heavy computational demand due to the comprehensive evaluation of each model in the population. Therefore, we opted to terminate the genetic algorithm after a fixed number of generations. Our preliminary simulations showed that 20 generations were sufficient to reach convergence.

Individuals of the initial population of the second round of optimization were generated randomly (*i.e.* the state of each parameter was set to 0 or 1 with 50% - 50% probability). In order to minimize the occurrence of neutral modifications in the solutions, we introduced a penalty based on the number of parameters that are in state 1 as follows:

$$\pi = 10^{-7} \text{ (number of parameters in state 1) / (number of parameters)}$$

The penalty parameter was chosen in such a way that it could not overwhelm any beneficial effect gained by adding a new modification to the model. Thus, among equally fit individuals, those encoding fewer model modifications were preferred during selection. Besides, we applied an elevated deletion rate by changing the application rates of the operators to 5%, 32.5%, 0.27.5%, 5% and 30% for elitism, crossing-over, mutation, insertion and deletion, respectively.

At the end of the second round of optimization we simplified the best solutions by removing neutral modifications. The effect of each model modification on overall prediction accuracy was examined in two ways. First, we introduced one modification at a time to measure the impact of a single modification in isolation. Second, all but one modifications were applied simultaneously thereby interrogating the essentiality of each suggested model modifications in the presence of others. Modifications that failed to affect prediction accuracy based on either of these two criteria were excluded from the final set.

We repeated the whole two-round optimization procedure 8 times and found high overlap between the best solutions of replicate runs. The list of recurring model modifications, their effects on prediction accuracy and their possible interpretations are summarized in Supplementary Table 3.

Cross-validation of the model refinement procedure

We performed cross-validation studies to investigate whether our automated procedure for refining the metabolic model improves the prediction of unseen genetic interactions (*i.e.* those not used for model refinement). To do this, we performed a series of 2-fold cross validations as follows:

1. We randomly partitioned both the genetic interaction and the single-mutant viability data into two sub-samples of equal size. The partitioning was carried out in a stratified manner, *i.e.* each sub-sample contained the same proportion of significantly interacting gene pairs and essential genes for the genetic interaction and single-mutant viability datasets, respectively.
2. We ran our two-step genetic algorithm on one partition of the data (training set).
3. We evaluated the optimized model(s) obtained from step 2 using the other partition of the data (test set). If multiple, equivalently performing models have been identified in step 2, then each of them was evaluated on the test set and their average performance was recorded. In addition, the original unmodified model was also evaluated on the test set.
4. We swapped the training and the test partitions and repeated steps 2 and 3.
5. Steps 1-4 were repeated five times.

The above procedure resulted in 10 independent estimates of the genetic interaction prediction accuracies ($MCC_{interactions}$) of both the original and the optimized models. A pair-wise comparison of prediction accuracies before and after applying the model refinement algorithm reveals that our method significantly improves the metabolic model, with the $MCC_{interactions}$ score being increased in all 10 cases ($P < 0.002$, Wilcoxon signed-rank test). Furthermore, the method also significantly improves the recall of genetic interaction predictions, by ~87% on average (as evaluated at an *in silico* genetic interaction score threshold of -0.9, corresponding to a precision of 0.5 for the original model).

Bioinformatic search for homologs of aspartate oxidase and quinolinate synthase in the yeast genome

Enzymes aspartate oxidase (EC 1.4.3.16) and quinolinate synthase (EC 2.5.1.72) form the aspartate *de novo* NAD pathway in *E. coli*[36]. Although the biochemical reactions catalyzed by these enzymes have been included in the iMM904[2] yeast metabolic network reconstruction, there are no ORFs assigned to them. To investigate whether these enzyme encoding genes are genuinely missing from the yeast genome we applied iterative PSI-BLAST searches[37]. PSI-BLAST iteratively builds position-specific score matrices (PSSM) from multiple alignments of the highest scoring hits from the previous BLAST search and is able to detect even weak relationships between sequences. We used the *E. coli* enzyme encoding genes *nadA* (quinolinate synthase) and *nadB* (L-aspartate oxidase) as queries for the initial BLASTP search with an E-value threshold of 10 (default) and the number of iteration was set to 20. We used three different strategies to search for homologues in the yeast genome. In the first strategy, we only included protein sequences from *S. cerevisiae* in all iterations. In the second strategy, protein sequences from all organisms were included during the search and yeast proteins were selected from the final list of hits. In the third strategy, the search was started with protein sequences from all organisms and the position-specific scoring matrix was saved after the first iteration. Next, another PSI-BLAST was performed with the PSSM as a query sequence profile and the search was restricted to *S. cerevisiae* sequences.

Our approaches failed to find any potential yeast homologue when using NadA protein as a query. Protein NadB shares the overall folding topology with other oxidoreductases[38] (succinate dehydrogenase and fumarate reductase) and therefore might show homology with yeast orthologs of these latter enzymes. We report that the most significant *S. cerevisiae* hits of the *NadB* query gene were annotated with succinate dehydrogenase and fumarate oxidase functions (*e.g. SDH1*, *OSM1*, *FRD1*, etc.). In addition, we detected two genes (*CIR2* and *HER1*) showing very low sequence similarities to *NadB* (E>0.1) that have unknown molecular functions. Given that our PSI-BLAST approaches easily retrieved yeast genes encoding oxidoreductases related to L-aspartate oxidase, but not L-aspartate oxidase itself, we conclude that there is no evidence for the presence of an L-aspartate oxidase gene in the yeast genome.

References to Supplementary Information

1. Segrè, D., Deluna, A., Church, G.M. & Kishony, R. Modular epistasis in yeast metabolism. *Nat Genet* **37**, 77-83 (2005).
2. Mo, M.L., Palsson, B.O. & Herrgard, M.J. Connecting extracellular metabolomic measurements to intracellular flux states in yeast. *BMC Syst Biol* **3**, 37 (2009).
3. Kuepfer, L., Sauer, U. & Blank, L.M. Metabolic functions of duplicate genes in Saccharomyces cerevisiae. *Genome Res* **15**, 1421-30 (2005).
4. Nagalakshmi, U. et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**, 1344-9 (2008).
5. Breiman, L. Random forests. *Machine Learning* **45**, 5-32 (2001).
6. R: A Language and Environment for Statistical Computing. (Vienna, Austria, 2007).
7. Liaw, A. & Wiener, M. Classification and Regression by randomForest. *R News* **2**, 18-22 (2002).
8. Sing, T., Sander, O., Beerenwinkel, N. & Lengauer, T. ROCR: visualizing classifier performance in R. *Bioinformatics* **21**, 3940-1 (2005).
9. Yan, C. & Sloan, D.L. Purification and characterization of nicotinamide deamidase from yeast. *J Biol Chem* **262**, 9082-7 (1987).
10. Emanuelli, M. et al. Identification and characterization of a second NMN adenylyltransferase gene in Saccharomyces cerevisiae. *Protein Expr Purif* **27**, 357-64 (2003).
11. Wogulis, M., Chew, E.R., Donohoue, P.D. & Wilson, D.K. Identification of formyl kynurenine formamidase and kynurenine aminotransferase from Saccharomyces cerevisiae using crystallographic, bioinformatic and biochemical evidence. *Biochemistry* **47**, 1608-21 (2008).
12. Hucka, M. et al. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* **19**, 524-31 (2003).
13. Tong, A.H. & Boone, C. Synthetic genetic array analysis in Saccharomyces cerevisiae. *Methods Mol Biol* **313**, 171-92 (2006).
14. Snitkin, E.S. et al. Model-driven analysis of experimentally determined growth phenotypes for 465 yeast gene deletion mutants under 16 different conditions. *Genome Biol* **9**, R140 (2008).
15. Wong, S.L. et al. Combining biological networks to predict genetic interactions. *Proc Natl Acad Sci U S A* **101**, 15682-7 (2004).
16. Ulitsky, I., Krogan, N.J. & Shamir, R. Towards accurate imputation of quantitative genetic interactions. *Genome Biol* **10**, R140 (2009).
17. Chechik, G. et al. Activity motifs reveal principles of timing in transcriptional control of the yeast metabolic network. *Nat Biotechnol* **26**, 1251-9 (2008).
18. Burgard, A.P., Nikolaev, E.V., Schilling, C.H. & Maranas, C.D. Flux coupling analysis of genome-scale metabolic network reconstructions. *Genome Res* **14**, 301-12 (2004).
19. Pu, S., Wong, J., Turner, B., Cho, E. & Wodak, S.J. Up-to-date catalogues of yeast protein complexes. *Nucleic Acids Res* **37**, 825-31 (2009).
20. Breitkreutz, B.J. et al. The BioGRID Interaction Database: 2008 update. *Nucleic Acids Res* **36**, D637-40 (2008).
21. Goldberg, D.S. & Roth, F.P. Assessing experimentally derived interactions in a small world. *Proc Natl Acad Sci U S A* **100**, 4372-6 (2003).
22. Balaji, S., Babu, M.M., Iyer, L.M., Luscombe, N.M. & Aravind, L. Comprehensive analysis of combinatorial regulation using the transcriptional regulatory network of yeast. *J Mol Biol* **360**, 213-27 (2006).
23. Huttenhower, C., Hibbs, M., Myers, C. & Troyanskaya, O.G. A scalable method for integration and functional analysis of multiple microarray datasets. *Bioinformatics* **22**, 2890-7 (2006).
24. Mewes, H.W. et al. MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res* **32**, D41-4 (2004).
25. Brown, J.A. et al. Global analysis of gene function in yeast by quantitative phenotypic profiling. *Mol Syst Biol* **2**, 2006 0001 (2006).

26. Christie, K.R. et al. Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from Saccharomyces cerevisiae and related sequences from other organisms. *Nucleic Acids Res* **32**, D311-4 (2004).
27. Boyle, E.I. et al. GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics* **20**, 3710-5 (2004).
28. Goldberg, D.E. *Genetic algorithms in search, optimization and machine learning*, (Addison-Wesley, 1989).
29. Rocha, M. et al. Natural computation meta-heuristics for the in silico optimization of microbial strains. *BMC Bioinformatics* **9**, 499 (2008).
30. Baldi, P., Brunak, S., Chauvin, Y., Andersen, C.A. & Nielsen, H. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* **16**, 412-24 (2000).
31. Kumar, V.S. & Maranas, C.D. GrowMatch: an automated method for reconciling in silico/in vivo growth predictions. *PLoS Comput Biol* **5**, e1000308 (2009).
32. Forster, J., Famili, I., Fu, P., Palsson, B.O. & Nielsen, J. Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. *Genome Res* **13**, 244-53 (2003).
33. Duarte, N.C., Herrgard, M.J. & Palsson, B.O. Reconstruction and validation of Saccharomyces cerevisiae iND750, a fully compartmentalized genome-scale metabolic model. *Genome Res* **14**, 1298-309 (2004).
34. Giaever, G. et al. Functional profiling of the Saccharomyces cerevisiae genome. *Nature* **418**, 387-91 (2002).
35. Bäck, T. & Hoffmeister, F. Extended Selection Mechanisms in Genetic Algorithms. *Proceedings of the Fourth International Conference on Genetic Algorithms*, 92-9 (1991).
36. Flachmann, R. et al. Molecular biology of pyridine nucleotide biosynthesis in Escherichia coli. Cloning and characterization of quinolinate synthesis genes nadA and nadB. *Eur J Biochem* **175**, 221-8 (1988).
37. Altschul, S.F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-402 (1997).
38. Mattevi, A. et al. Structure of L-aspartate oxidase: implications for the succinate dehydrogenase/fumarate reductase oxidoreductase family. *Structure* **7**, 745-56 (1999).