**Additional files for**

**Identification of Cancer–associated Metabolic Vulnerabilities by Modeling Multi-objective Optimality in Metabolism**

Ziwei Dai1, Shiyu Yang4, Liyan Xu4, Hongrong Hu4, Kun Liao4, Jianghuang Wang4, Qian Wang2, Shuaishi Gao1, Bo Li4\*, Luhua Lai1,2,3\*

1. Center for Quantitative Biology, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, China

2. Beijing National Laboratory for Molecular Sciences, State Key Laboratory for Structural Chemistry of Unstable and Stable Species, College of Chemistry and Molecular Engineering, Peking University, Beijing 100871, China

3. Peking-Tsinghua Center for Life Sciences, Peking University, Beijing 100871, China

4. Program of Cancer Research, Affiliated Guangzhou Women and Children's Hospital, Zhongshan School of Medicine, Sun Yat-Sen University, Guangzhou 510080, China

A screenshot of a cell phone

Description automatically generated

**Figure S1 (Related to Figure 2). Comparison between actual growth rates of NCI-60 cell lines and growth rates predicted by alternative models.**

(a-d) Comparisons between actual growth rates and growth rates predicted by multi-objective optimization models with three objectives: (a) maximizing ATP hydrolysis, minimizing enzyme abundance, minimizing carbon uptake; (b) maximizing biomass synthesis, minimizing enzyme abundance, minimizing carbon uptake; (c) maximizing biomass synthesis, maximizing ATP hydrolysis, minimizing carbon uptake; (d) maximizing biomass synthesis, maximizing ATP hydrolysis, minimizing enzyme abundance.

(e) Comparison between actual growth rates and growth rates predicted by cell line-specific models constructed using proteomics data as the only input dataset.

(f) Comparison between actual growth rates and growth rates predicted by cell line-specific models constructed using exchange flux data as the only input dataset.

A screenshot of a cell phone screen with text

Description automatically generated

**Figure S2 (Related to Figure 2). Comparisons between actual growth rates of NCI-60 cell lines and growth rates predicted by other methods.**

(a) Comparison between actual growth rates of NCI-60 cell lines and predictions by e-Flux using raw protein expression levels as the input dataset.

(b) Same as in (a) but using log-transformed protein expression levels as input.

(c) Comparison between actual growth rates of NCI-60 cell lines and predictions by PRIME using raw protein expression levels as the input dataset.

(d) Same as in (c) but using log-transformed protein expression levels as input.



**Figure S3 (Related to Figure 2). Comparison between model-predicted essential and nonessential metabolic genes.**

(a) Numerically computed PDS values for NCI-60 cell lines without perturbations.

(b) Quantile-quantile (Q-Q) plot comparing distributions of maximal Pareto fluxes (i.e. maximal flux carried by the corresponding reaction across all Pareto solutions) associated with essential and nonessential metabolic genes. The p-value was computed using one-sided Kolmogorov-Smirnov test.

(c) Top 10 KEGG metabolic pathways enriched in model-predicted essential metabolic genes.

(d) Top 10 KEGG metabolic pathways enriched in model-predicted nonessential metabolic genes.



**Figure S4 (Related to Figure 2). Distributions of Spearman’s rank correlation coefficients between experimentally measured sensitivities of NCI-60 cell lines to metabolic gene knockdowns and sensitivities predicted by different computational methods.** P-values were computed using one-sided Wilcoxon’s signed rank test with the null hypothesis of median value = 0. Left panel: CRISPR-based gene ablation dataset; right panel: RNAi-based gene ablation dataset.



**Figure S5 (Related to Figure 3). Influences of monotonousness score cutoff used in Pareto surface analysis on the associations between model-predicted targets and cancer patient survival.**

(a) Relationship between monotonousness score cutoff used in the Pareto surface analysis and resulting numbers of model-predicted tumor-suppressive metabolic genes associated with decreased, increased or unchanged breast cancer patient survival.

(b) Same as in (a) but for pro-oncogenic metabolic genes.

(c) Same as in (a) but for ambiguous genes.

(d) Relationship between monotonousness score cutoff used in the Pareto surface analysis and chi-squared p-values for enrichment of metabolic genes associated with decreased, increased or unchanged breast cancer patient survival in the resulting pro-oncogenic, tumor-suppressive or ambiguous genes.



**Figure S6 (Related to Figure 3). Comparison between targets identified by Pareto surface analysis and other methods.**

(a) Venn diagram comparing tumor-suppressive metabolic genes identified by Pareto surface analysis and other methods.

(b) Venn diagram comparing pro-oncogenic metabolic genes identified by Pareto surface analysis and other methods.

(c) Numbers of genes associated with decreased, increased or unchanged breast cancer patient survival in pro-oncogenic genes predicted by Pareto surface analysis and other methods.

(d) Numbers of genes associated with decreased, increased or unchanged breast cancer patient survival in tumor-suppressive genes predicted by Pareto surface analysis and other methods.



**Figure S7 (Related to Figures 4-6). Validation of efficiencies of gene knockdowns and over-expressions.**

(a) mRNA levels of FTCD, HAL and UROC1 with or without knockdown in HeLa cells.

(b) Protein abundances of AASS, AADAT, PYCR1, PYCR2, CTPS1, CTPS2 and MDH2 with or without over-expressions of the corresponding genes in HeLa cells.

A picture containing indoor, monitor

Description automatically generated

**Figure S8 (Related to Figure 6). Mitochondrial respiration and ECAR profiles of SW620, A549, BT549, HeLa, RCC10 and U87 cells with or without over-expression of MDH2, CTPS1, CTPS2, PYCR1 or PYCR2.**

**A screenshot of a cell phone

Description automatically generated**

**Figure S9 (Related to Figure 6). Relative number of cells after 4 days in the control group (PCDH) or upon over-expression of MDH2, CTPS1, CTPS2, PYRC1 or PYRC2 in the tested cell lines.**

**Supplementary Methods**

**Genome-scale metabolic model and definition of metabolic objectives**

We considered four metabolic objectives including maximization of biomass production flux (fBM), maximization of ATP turnover (fATP), minimization of carbon uptake (CU), and minimization of total metabolic enzyme abundance (EA). The genome-scale metabolic model used here, Recon 1, already contains a biomass producing flux whose coefficients are determined by the molecular composition of mammalian cells. We employed a curated and decomposed (i.e. all reversible reactions are decomposed into forward and backward reactions) Recon 1 model by Shlomi *et al* to simplify the following procedures. Model files in the SBML format were downloaded from the BioModels Database (<https://www.ebi.ac.uk/biomodels/>, Model ID: MODEL1105100000). We then translated the SBML files into MATLAB .m files using the COBRA Toolbox1, and supplemented the original model with additional biomass components for the following analysis. An ATP hydrolysis flux was directly included in the model. The carbon uptake (CU) flux was calculated as follows:

n is number of fluxes in the model, is the number of carbon atoms imported into intracellular compartments by the i-th flux and is its flux rate. If this flux does not lead to any carbon uptake, the value of is zero.

The enzyme abundance (EA) was determined by:

Since the coefficients have already been evaluated in the curated model in an artificial flux corresponding to the solvent capacity constraint, these values were directly utilized in our analysis. The artificial solvent capacity flux was removed from the model. Upper limits of nutrient influxes were set according to the composition of RPMI-1640 medium. Details about the model is included in Supplementary Table 1.

**Sampling the Pareto surface**

Pareto solutions of optimization problems with *N* objectives were generated using -Constraint Method, in which a single objective selected from the *N* objectives is optimized while the remaining *N*-1 of the objectives are transformed into soft constraints. In general, a multi-objective optimization problem on a genome-scale metabolic model can be written in the form below:

In which ***S*** is the stoichiometric matrix for the metabolic network, ***x*** is the flux vector in which each element is the rate of a reaction (*i.e.* a flux), ***ub*** is the vector of maximal rates of reactions. Assume that is the objective to be optimized and all other objectives are treated as constraints, the single-objective optimization problem below, if feasible, gives a solution with Pareto optimality for the original multi-objective optimization problem:

By selecting different combinations of objectives to be transformed to constraints, adjusting the ε-values (*e1, …, eN*), and repeatedly solving the corresponding single objective optimization problem, we can obtain a set of solutions with Pareto optimality. In our model all objectives and constraints are linear, thus enabling us to solve the single objective optimization problems with efficient algorithms (e.g. simplex method, interior point method, etc) for linear programming (LP). The LP problems were solved using Mosek (<http://www.mosek.com/>).

To generate feasible linear programming problems more efficiently, we applied a two-step pre-processing strategy to select combinations of ε-values that yield feasible single-objective linear programming problems. First, we calculated the maximal values for the biomass production flux () and the ATP hydrolysis flux (). We then sampled 10000 combinations of in the region with Latin Hypercube Sampling. For each combination we first checked the feasibility of the linear programming problem with the two additional constraints and , and then calculated for each feasible the feasible ranges of CU and EA ( and ) compatible with constraints in the GSMM, and . Calculating the feasible range of and according to sampled prior to generation of the single-objective optimization problems helps avoid generating a large number of infeasible problems. We then applied stratified sampling to select values within and values within and used these randomly sampled values as the upper limits for CU or EA to generate LP problems among which problems minimize EA and minimize CU. The values of and were determined by the range of and with 10 as the maximal value. Solutions of all feasible linear programming problems (42930 in total) were summed up as the sampled Pareto surface.

**Retrieving and processing the omics datasets**

Proteomics data of the NCI-60 cell lines are available at the NCI-60 Proteome Resource: <http://wzw.tum.de/proteomics/nci60>. The exchange flux rates of these cell lines were obtained from a study at 20122. Protein abundances and exchange fluxes were normalized to cell sizes by dividing cell volumes calculated based on cell diameters available at: <http://www.nexcelom.com/Applications/Cancer-Cells.html>, with the assumption that single cells are perfect spheres. Expression levels of metabolic enzymes were evaluated based on proteomics data and gene-protein-reaction rules (GPR rules) included in the GSMM. GPR rules are logical expressions connecting genes, enzymes and reactions. These expressions include logic operators AND, OR and brackets. Expression levels of enzymes were estimated by incorporating the expression levels of genes to the GPR formulas and then evaluating the formulas with AND operators treated as minimum operators and OR operators treated as addition operators. Below is one example of GPR rules:

In which *R* is the GPR rule connecting the enzyme *E* and , and which are three transcripts related to the enzyme in the following way: E has two subunits, one is translated from , the other can be translated from either or . Let , and denote abundance of the three transcripts, the relative abundance of E, e, is evaluated by:

CRISPR-based gene dependency dataset in the Achilles database (“gene\_dependency\_corrected.csv”) was retrieved from <https://figshare.com/articles/DepMap_Achilles_19Q1_Public/7655150>. RNAi-based gene dependency dataset (“D2\_Achilles\_gene\_dep\_scored.csv”) in the Achilles database was retrieved from <https://figshare.com/articles/DEMETER2_data/6025238/5>. Cell growth rates were evaluated based on doubling times available at: <http://discover.nci.nih.gov/cellminer/>.

**Construction of cell line-specific models**

We defined a similarity score to evaluate to which extent the distribution of flux configurations predicted by the individualized models can reproduce the corresponding omics data. The similarity metric (S) is comprised of two parts: . is the summation of correlation coefficients between model-predicted and experimentally-determined exchange fluxes:

In which n is the total number of cell lines, is the total number of experimentally measured exchange fluxes, is the experimentally measured value of the i-th exchange flux in the k-th cell line, is the model-predicted value of the i-th exchange flux in the k-th cell line.

is the summation of correlation coefficients between model-predicted average flux rates and average expression levels of metabolic pathways:

In which is the total number of metabolic pathways in the KEGG database, is the average protein expression level of metabolic enzymes included in the i-th pathway for the k-th cell line, is the model-predicted average flux rate for reactions included in the i-th pathway for the k-th cell line. Spearman’s rank correlation coefficient was used to quantify the correlation. Single enzymes were mapped to metabolic pathways according to the BRENDA Database based on their EC Numbers. The cell line-specific models (i.e. flux configurations for the cell lines) were constructed by searching for combination of Pareto solutions maximizing the similarity metric S using simulated annealing.

**Calculation of the Pareto deviation score**

We simulated the effects of metabolic gene knockdowns using minimization of metabolic adjustments (MOMA)3. Metabolic gene knockdowns were simulated by reducing upper bounds to zero for fluxes carried by enzymes associated with the ablated genes. For a flux configuration , the new flux configuration after knockdown of the i-th gene was calculated by solving the quadratic programming problem below:

Inhibition of glycolysis was simulated by reducing the upper bound of flux carried by the reaction “hexokinase (D-glucose:ATP)” to 10% of the current flux value. Inhibition of OXPHOS was simulated by reducing the upper bound of flux carried by the reaction “ATP synthase (four protons for one ATP)” to 10% of the current flux value. Flux configuration after inhibition of glycolysis or OXPHOS or both pathways was computed using MOMA as described above.

Let denote the simulated new flux configuration after metabolic gene knockdown or metabolic enzyme inhibition, the Pareto deviation score (PDS) was then computed as below:

In which PS denotes the set containing all sampled Pareto solutions, , , and are the biomass synthesis flux, ATP hydrolysis flux, carbon uptake and total metabolic enzyme abundance corresponding to flux configuration . PDS values below 0.01 were converted to zero to correct for numerical errors (Supplementary Figure 3a).

**Pareto surface analysis**

The upper and lower bounds of projected Pareto surface were approximated by mathematical discretization. Ranges of cell growth rate or the Warburg effect metric (defined as the ratio of lactate secretion flux to glucose uptake flux) of all Pareto solutions were divided into 50 bins with identical sizes. Let denote the variable describing the *i*-th flux, the lower bounds (i.e. minimal values) and upper bounds (i.e. maximal values) of in all Pareto solutions whose growth rates or the Warburg effect metric fell in each of the 50 bins were calculated as and . The monotonous scores for upper and lower bounds of this flux was quantified by Spearman’s rank correlation coefficients between the vector or and the vector [1, 2, ..., 50]. Bounds were considered as monotonously decreasing if the monotonous scores were less than -0.9, and monotonously increasing if larger than 0.9.

**Patient survival and pathway enrichment analysis**

Kaplan-Meier survival analyses were performed using the Kaplan-Meier Plotter (<http://kmplot.com/analysis/>)4 using the probe set option ‘only JetSet best probe set’, patients split option ‘Auto select best cutoff’. Default parameters were used for all other options not mentioned. For genes without JetSet best probe sets, the probe set with maximal number of patients was selected. Pathway enrichment analysis was performed using the module ‘Investigate Gene Sets’ available at the Molecular Signatures Database (MSigDB, <http://software.broadinstitute.org/gsea/msigdb/index.jsp>).

**Comparison with alternative methods**

For construction of cell line-specific models, PRIME5 and E-Flux6 methods were implemented in MATLAB codes using proteomics data for NCI-60 cell lines as the input. Raw intensity values were either directly used (Supplementary Fig. 2a, c) or log-transformed (Supplementary Fig. 2b, d) using the expression before being used as input of the PRIME and E-Flux algorithms. Doubling times of the NCI-60 cell lines were also used as the input for PRIME.

For identification of putative metabolic targets, three alternative methods including correlation between mRNA abundances and growth rates, correlation between protein expression levels and growth rates and minimization of metabolic adjustments (MOMA) were applied. For the correlation-based methods, correlations between NCI-60 cancer cell proliferation and mRNA or protein expression levels of metabolic enzymes in these cell lines were quantified using Spearman’s rank correlation coefficients. Metabolic enzymes with mRNA or protein expression level significantly correlated with growth rates of NCI-60 cell lines (i.e. Benjamini-Hochberg adjusted p-value<0.05) were identified as tumor-suppressive (Spearman’s correlation coefficient<0) or tumor-promoting (Spearman’s correlation coefficient>0) targets. For the MOMA-based method, MOMA was applied to simulate the effects of inhibition or activation of metabolic enzymes on growth rates of NCI-60 cell lines. Inhibition of a metabolic enzyme in a cell line-specific model was simulated by reducing the upper bound of the flux catalyzed by this enzyme to 10% of the current value of this flux in the corresponding model, while activation of a metabolic enzyme was simulated by adjusting the lower bound of the flux catalyzed by this enzyme to two-folds of the current flux rate. MOMA was then applied to compute the new metabolic flux configuration upon the inhibition or activation. Average relative reduction of the biomass synthesis flux in the NCI-60 cell line-specific models was used as the metric for target identification. Metabolic enzymes whose inhibition or activation resulted in more than 50% relative reduction of biomass synthesis flux averaged over the NCI-60 cell line models were identified as tumor-promoting or tumor-suppressing targets.

**References**

1 Schellenberger, J. *et al.* Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nat Protoc* **6**, 1290-1307, doi:10.1038/nprot.2011.308 (2011).

2 Jain, M. *et al.* Metabolite profiling identifies a key role for glycine in rapid cancer cell proliferation. *Science* **336**, 1040-1044, doi:10.1126/science.1218595 (2012).

3 Segre, D., Vitkup, D. & Church, G. M. Analysis of optimality in natural and perturbed metabolic networks. *Proc Natl Acad Sci U S A* **99**, 15112-15117, doi:10.1073/pnas.232349399 (2002).

4 Szasz, A. M. *et al.* Cross-validation of survival associated biomarkers in gastric cancer using transcriptomic data of 1,065 patients. *Oncotarget* **7**, 49322-49333, doi:10.18632/oncotarget.10337 (2016).

5 Yizhak, K. *et al.* Phenotype-based cell-specific metabolic modeling reveals metabolic liabilities of cancer. *Elife* **3**, doi:10.7554/eLife.03641 (2014).

6 Colijn, C. *et al.* Interpreting expression data with metabolic flux models: predicting Mycobacterium tuberculosis mycolic acid production. *PLoS Comput Biol* **5**, e1000489, doi:10.1371/journal.pcbi.1000489 (2009).