Recon8D: A metabolic regulome network from oct-omics and machine learning

Ryan Schildcrout[1], Kirk Smith[1, 2, 3], Rupa Bhowmick[1], Suraj Menon[1], Minali Kapadia[1], Emily Kurtz[4], Anya Coffeen-Vandeven[5], Sriram Chandrasekaran[1, 6, 7]

[1] Department of Biomedical Engineering, University of Michigan, Ann Arbor.

[2] Programs in Metabolism and Medical & Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA

[3] Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA

[4] Department of Chemical Engineering, University of Michigan, Ann Arbor.

[5] Department of Biological Chemistry, University of Michigan, Ann Arbor

[6] Program in Chemical Biology, University of Michigan, Ann Arbor.

[7] Rogel Cancer Center, University of Michigan Medical School, Ann Arbor, MI, 48109, USA.

* Correspondence: csriram@umich.edu

Abstract

The complexity and incompleteness of metabolic-regulatory networks make it challenging to predict metabolomes from other omics. Using machine learning, we predicted metabolomic variation across ~1000 different cancer cell lines from matched oct-omics data: genomics, epigenomics (histone post-translational modifications (PTMs) and DNA-methylation), transcriptomics, RNA splicing, miRNA-omics, proteomics, and phosphoproteomics. Overall, the metabolome is tightly associated with the transcriptome, while miRNAs, phosphoproteins and histone PTMs have the highest metabolic information per feature. Metabolites in peripheral metabolism are predictable via levels of corresponding enzymes, while those in central metabolism require combinatorial predictors in signaling and redox pathways, and may not reflect corresponding pathway expression. We reconstruct multiomic interaction subnetworks for highly predictable metabolites, and YAP1 signaling emerged as a top global predictor across 4 omic layers. We prioritize predictive multiomic features for single-cell and spatial metabolomics assays. Top predictors were enriched for synthetic-lethal interactions and synergistic combination therapies that target compensatory metabolic modulators.

Introduction

Variation in metabolite levels between cells may arise due to differences in regulation at the transcriptional, post-transcriptional, translational, and post-translational levels[1,2]. Predicting intercellular metabolic variation and understanding metabolic regulation in a holistic manner remains challenging due to the complex interconnected nature of metabolic networks, comprising thousands of reactions, that are controlled by an equally complex web of regulatory interactions[3]. In addition to thermodynamics and nutrient availability, intracellular metabolite levels are also driven by enzyme activity, which is the basis for the metabolome's relationship with proteogenomic networks. Because of their wide availability, transcriptomics or proteomics data have been used to predict the metabolome with limited success[4–7]. However, these studies focus on a limited set of omics data and do not examine the relative influence and predictive power of each omics class on the metabolome[8–18]. Further, it is unclear if certain metabolites are more predictable than others.

Here, we leverage natural metabolic variation among hundreds of cancer cell lines to infer a multiomic predictor network of the metabolome (Recon8D) using machine learning (ML). Our approach utilizes omics inputs from 8 biomolecular classes from the Cancer Cell Line Encyclopedia (CCLE), allowing for direct comparison of each class on predicting metabolite variation. We hypothesized that extent of predictability is indicative of the strength of relationship with the metabolome. Similar assumptions have been used in numerous studies to assess relationships between transcription factors and target genes[19–22]. Although many metabolic regulatory relationships have been documented, they represent a small fraction of the number of interactions that likely exist in biological systems. Machine learning algorithms can fill this gap by learning relationships from experimental data. Our approach uncovered numerous mechanistic relationships and identified significant multiomic predictors of metabolomic variation that could serve as markers for either individual metabolites or the overall metabolome for *in vivo*, single-cell or spatial omics studies that require limited measurements[23–26]. Finally, our discovery of synthetic lethal interactions among top metabolite predictors can help design combination therapies by targeting compensatory mechanisms.

Results

Recon8D model construction using machine learning

Analogous to gene regulatory inference with ML, wherein the expression value of each gene is taken as the learning objective, here we use metabolite level variation as the prediction task in a regression problem using multiomic data as input[21,27]. The output is an 'influence' network model that implicitly captures regulatory events at the transcriptomic, proteomic or post-translational levels. Our approach used omics from the CCLE, which include biomolecular assays from up to ~1,000 cancer cell types[28], to predict the corresponding metabolome. We utilized genomics, epigenomics (DNA methylation and histone PTMs), transcriptomics, RNA splicing, miRNA-omics, proteomics, and phosphoproteomics to train ML models capable of predicting metabolomic profiles across matched cell lines (Figure 1, S. Table 1).

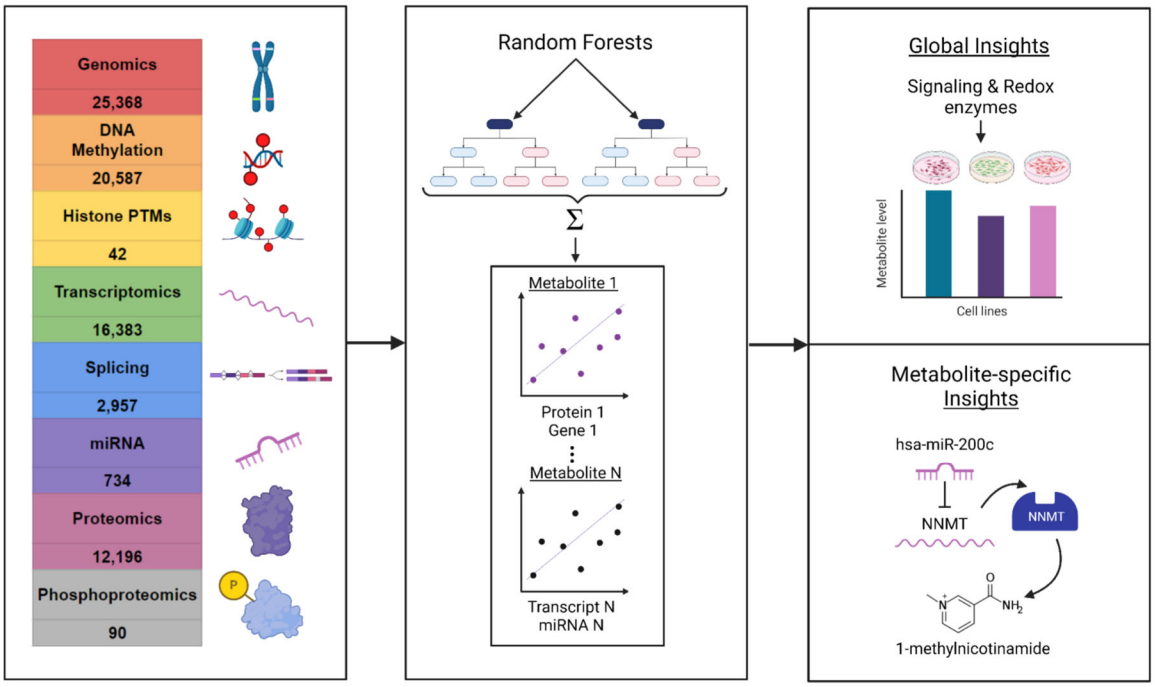


Figure 1: Graphical overview of our approach to predict the metabolome from multiomics and inferring top global and metabolite-specific predictors. Each omics class is utilized in isolation to train a predictive model for each metabolite in a metabolomic panel, and the accuracies and feature importances thereof are used to reconstruct Recon8D metabolic regulatory networks.

Random forest regression ML algorithms were used to generate metabolomic predictions as they outperformed other ML methods based on cross-validation in our benchmarking (S. Figure 1). Random forests have also outperformed other methods in the inference of gene regulatory networks in prior studies[21,27]. We further make use of inherent characteristics of random forests to obtain the importance of each feature in prediction tasks. Higher feature importances indicate a greater possibility that a gene or protein regulates the target metabolite. Separate random forest models were trained iteratively for each of 225 metabolites in the CCLE metabolome using each of the omics features in an input set to create a total of 1,800 models (Figure 2A, 2B). Model performance was evaluated using cross validation, and significantly predictable metabolites were identified based on a Bonferroni-corrected P value of 0.05 to account for false discovery rate (Figure 2C).

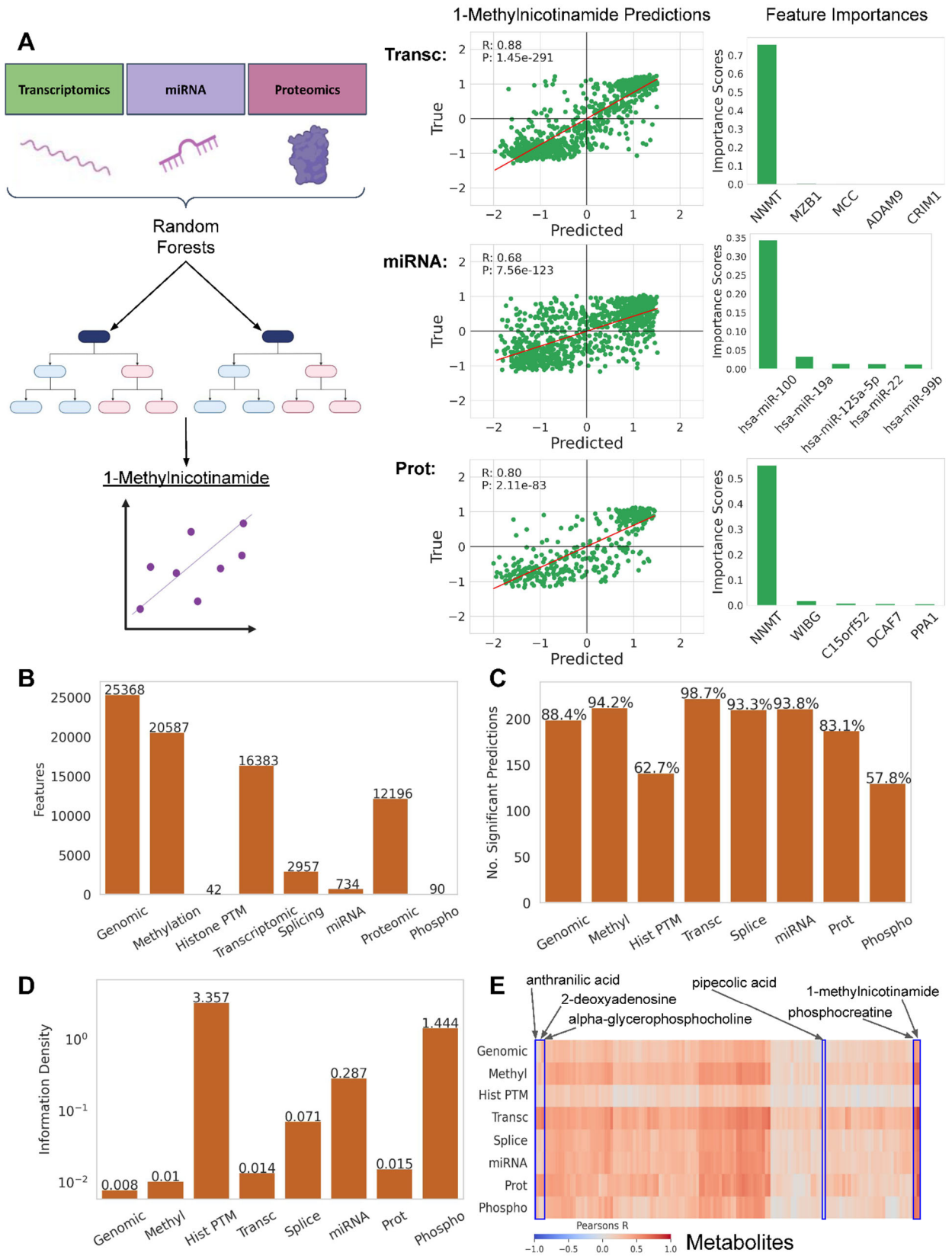


Figure 2. Random forest model performance indicates transcriptomics, DNA methylation, and miRNA-omics are accurate predictors of metabolomic variation. A. Representative examples of 1-methylnicotinamide

predictions using transcriptomic, miRNA, and proteomic inputs. Each omics class is used individually to predict relative 1-methylnicotinamide levels, which yields a predicted metabolite variation array which can be compared to the true values. Scatter plots of predicted vs true values in cross validation help visualize the accuracy of each model. Finally, feature importance analysis identifies the most influential features on the model output. In this case, the model identified NNMT, the enzyme involved in 1-methylnicotinamide synthesis, as the top feature predictive of 1-methylnicotinamide. **B.** Number of features in each input dataset. **C.** Number of significantly predicted (FDR p-value < 0.05) metabolites per input omics class for each of 225 metabolites measured in the CCLE panel. **D.** Log2-scaled information density of each omics class (total significant predictions divided by total features). **E.** Clustered heatmap showing correlation values for all non-lipid metabolites (columns) with select metabolites labeled.

ML model performance indicates metabolic variation reflects transcriptional differences between cell lines

Across all ML algorithms and validation procedures, transcriptomics consistently generated the most accurate predictions (Figure 2, S. Figures 1, 3-5). Model performance does not seem to directly coincide with the number of features present in the training data (Figure 2B & 2C). This is further illustrated in Figure 2D, where information density quantifies the relevance of each feature as a function of its performance. While histone PTMs and phosphoproteomics were not as accurate as other omics classes, their information density is high, showcasing that small feature sets can still impart meaningful predictions. The ability of each omics class to make accurate predictions independent of feature set size instills confidence in the hypothesis that the ML models are picking up mechanistic relationships and are not overfitting to the data. For example, with the third-least number of features out of all omics classes, miRNA inputs produced the third-most significant predictions. In contrast, genomics had the most features, yet only the fifth-highest predictive performance. We further validated this assumption with benchmarking models trained on the top 40 features per dataset identified using holdout validation with a training set to testing set ratio of 75 to 25 (S. Figure 3). Both the transcriptomic and miRNA data achieved high accuracy (96% of metabolites predicted with FDR p-value < 0.05) while only using 40 features. The top 40 features can thus potentially serve as a reduced set of molecular markers that can be measured for reconstructing the metabolome (S. Table 2).

Clustered heatmaps of Pearson's correlations (model predictions versus true relative metabolite levels) illustrate the differences in predictive power across omics inputs (Figure 2E). This visualization is useful for identifying metabolites that are differentially predictable using certain omics inputs. While some metabolites are highly predictable across omics layers, such as 1-methylnicotinamide and phosphocreatine, others have stark differences in predictability. For example, anthranilic acid, 2-deoxyadenosine, and alpha-glycerophosphocholine predictions were more accurate with transcriptomic and proteomic inputs. This suggests that individual metabolites can be influenced preferentially at particular omics levels. Conversely, some metabolites, such as pipecolic acid, do not have high correlations with any model, suggesting that their mode of regulation might not be sufficiently captured in these data.

Control and validation experiments show consistency and robustness in predictive capability

To account for confounding factors that may drive spurious correlations, such as tissue type, culture conditions, and tumor sample origin, several control experiments were conducted. We trained models using omics subsets separated by growth type (adherent and suspension), cancer lineage (lung and hematopoietic/lymphoid tissue), and by tumor sample origin (primary and

147 metastatic) (S. Figure 4). While the raw accuracies of the models in these experiments are lower
148 due to loss of observations per dataset, thereby limiting training data, similar trends in the
149 predictive power of each omic subset are observed. We also assessed the robustness of the ML
150 models by identifying the top predictive features across each control experiment. Models
151 controlled for tumor origin show a high level of concordance in accuracy and feature importance.
152 Overall, transcriptomics had the highest predictive accuracy across all omics models in all
153 control experiments.
154
155 To assess the robustness of the top features in independent datasets, validation experiments were
156 conducted where top features from CCLE-trained models were compared to corresponding
157 relationships between features and metabolites in the NCI60 cancer cell line panel[29]. Random
158 forest models were trained on NCI60 genomics, transcriptomics, miRNA, proteomics, and
159 phosphoproteomics data, and top features from the CCLE models were compared to the top
160 features from the NCI60 models (Methods). CCLE-identified features showed significantly
161 higher feature importance scores in the NCI60 data as compared to the whole feature set (S.
162 Tables 3-4). To further assess the generalizability of this framework to single-cell datasets,
163 wherein the metabolome is challenging to measure, we performed a similar analysis described
164 above with single-cell transcriptomics of the cell cycle and compared with matched metabolome
165 from synchronized cells[30,31]. CCLE-identified features show significantly higher feature
166 importance scores in the single-cell transcriptomics data than expected out of random chance (S.
167 Table 5). These experiments demonstrate generalizability of the CCLE-trained metabolite
168 prediction models across cell line panels.
169
170 ***Model Feature Importance Analysis reveals global and metabolite-specific predictors***
171
172 For each metabolite, we looked at top predictors from each omics-based model based on feature
173 importance scores (Methods). The transcriptomics model not only had high predictive power, but
174 also recalled several known mechanistic metabolite-gene relationships. For some metabolites, the
175 mRNA transcript of the corresponding enzyme that synthesizes the metabolite was the top
176 predictor. These include choline O-acetyltransferase gene for the metabolite acetylcholine,
177 choline dehydrogenase for betaine, cystathionine beta-synthase for cystathionine, and glycerol 3-
178 phosphate dehydrogenase for alpha-glycerophosphate. For other metabolites, the consumption or
179 catabolic degradation was more predictive of their levels. For example, cytidine deaminase for
180 cytidine, and monoamine oxidase A for serotonin. Several metabolites' corresponding
181 transporters also featured as top predictors, including SLC6A6 for taurine and SLC6A8 for
182 phosphocreatine. Nicotinamide and tryptophan metabolism shared several top predictors,
183 suggesting shared regulation. Nicotinamide phosphoribosyltransferase (NAMPT) was a top
184 predictor of NAD. Indoleamine 2,3-dioxygenase (IDO) and tryptophan 2,3-dioxygenase were
185 most predictive of kynurenine. Anthranilic acid was predicted by kynureninase, and both
186 methylnicotinamide and niacinamide were predicted by nicotinamide N-methyltransferase
187 (NNMT).
188
189 The proteomics model also recapitulates several known relationships between metabolites and
190 specific enzymes that impact their levels. For example, the top predictor in the proteomics model
191 for succinate was succinate dehydrogenase, and the top predictor for phosphoenol pyruvate was
192 pyruvate kinase (PKM). Thymine and uracil's top predictor was dihydropyrimidine

193 dehydrogenase (DPYD) - the initial and rate-limiting enzyme in uracil and thymidine catabolism.
194 Similarly, malic enzyme (ME2), part of the malate-aspartate shuttle, was a top predictor of
195 aspartate levels, while glutaminase (GLS) was a top predictor of glutamine.
196
197 While many metabolites were predicted preferentially by a single top feature, others were
198 predicted by multiple features suggesting that they might be involved in multiple pathways and
199 may be regulated combinatorially and will therefore be harder to predict using transcriptomic
200 variation of a single gene. For example, citrate is an intermediate in multiple core metabolic
201 pathways, and its concentration remains low in cancer due to its rapid conversion into acetyl-
202 CoA and oxaloacetate[32]. As opposed to metabolites such as taurine where there is one clear
203 regulator, citrate's feature importances show a more balanced regulatory influence among
204 several top features (Figure 3D). The top predictor for taurine was SLC6A6, the taurine
205 transporter, and was shown by our models to have an extremely high influence in predicting
206 relative taurine levels as compared to even the second most important feature by Shapley
207 analysis (Figure 3D). However, citrate's feature importance plot does not show one clear top
208 predictor, but rather an array of features that appear to impact relative citrate levels similarly. In
209 this way, ML can uncover complex metabolic control by multiple inputs in addition to strong
210 associations between single features and metabolites. Interestingly, metabolites that had a
211 distinct single predictor like anthranilic acid, N-carbamoyl-beta-alanine, and adenosine were also
212 more generalizable to NCI60 (S. Table 3). This may also be related to the limited ability of ML
213 to tease out complex combinatorial relationships using small data such as the NCI60 panel.
214
215 The ML models also have the ability to identify potential novel associations between features
216 and metabolites. While the top feature for 7 out of 10 of the most accurately predicted
217 metabolites from transcriptomics was found to be mechanistically connected to its respective
218 metabolite, the remaining 3 metabolites–oxalate, kynurenic acid, and alpha-hydroxybutyrate–had
219 YAP1, a signaling protein, as the top predictor. Oxalate and kynurenic acid have been found to
220 be associated with YAP1[33,34], but the association between alpha-hydroxybutyrate and YAP1 has
221 yet to be established.
222
223 The ML models also revealed long range metabolite-enzyme interactions. The top features in the
224 proteomics model for glutathione, a key redox metabolite, were SLC7A11 and ALDH3A1.
225 SLC7A11 is a cysteine and glutamate transporter (xCT) and its activity has been shown to
226 influence glutathione levels[35]. Aldehyde dehydrogenase protects cells from oxidative stress and
227 consumes redox cofactors. Interestingly, SLC7A11 and ALDH3A1 also form a synthetic lethal
228 pair, suggesting that they may compensate for each other's activity in maintaining redox
229 balance[36].
230
231 Interestingly, the top predictors for a majority of the metabolome across various omics data
232 involved either signaling proteins like YAP1 or redox enzymes such as GPX8 (Figure 3B and C).
233 YAP1 was among the top 2 predictors in the transcriptomics model for 45 metabolites including
234 numerous amino acids and nucleotides, while GPX8 was among the top 2 predictors for 23
235 metabolites including various lipids. This suggests that the combination of growth signaling and
236 redox homeostasis influences a significant part of the metabolome.
237
238 ***Top predictors are close topologically to the corresponding metabolites***

To further substantiate the ability of ML to identify metabolic regulatory interactions, we mapped features from genomic, DNA methylation, transcriptomic, and proteomic models to reactions present in the Recon3D model (Figure 3A, Methods). Recon3D is a metabolic network model built on biochemical reactions curated from literature[37]. It is the most comprehensive collection of known metabolic reactions in human cells, making it a prime candidate for systematically validating ML models and uncovering topological relationships between top features and their respective predicted metabolites. To ensure overlap with Recon3D, ML models were re-trained on the Recon3D metabolic gene subset. The resulting metabolic subset ML models showed similar accuracy to whole-data models (S. Figure 5B and C). Feature importances from metabolic gene-only models and whole-data models were also conserved significantly (S. Figure 5D and E).

The top 5 features for each metabolite from all the metabolic ML models displayed a substantial number of overlaps (92-96%) with either the corresponding reactions directly associated with each metabolite or their first degree neighbors (S. Tables 6, 7). This suggests an arms-length or 1-step influence strategy wherein the top predictors for ~95% of the metabolites are within 1 step of the reactions involving those metabolites.

While the top features were on average topologically closer to the corresponding metabolite, consistent with prior studies[6,38], we found distinct differences for metabolites in peripheral or central metabolism (Figure 3C). For example, in the transcriptomics ML models, the top features for only 50% of the metabolites in central carbon metabolism overlapped with either the corresponding reaction or first degree neighboring reactions. In contrast the top features for 84% of the metabolites in peripheral metabolism were direct or first degree neighbors, suggesting that peripheral metabolites are more influenced by direct enzymatic regulation, as opposed to those in central carbon metabolism, which are likely influenced more by signaling (Figure 3C).

To further explore if signaling is a larger player in central carbon metabolism than peripheral metabolism, we determined if top features in each category were signaling enzymes for all non-lipid metabolites. We found that 31% of metabolites in central carbon metabolism had signaling enzymes as the top feature compared to only 12.5% for peripheral metabolites (Figure 3C). These findings suggest that central carbon metabolism is influenced to a greater extent by signaling than peripheral metabolism.

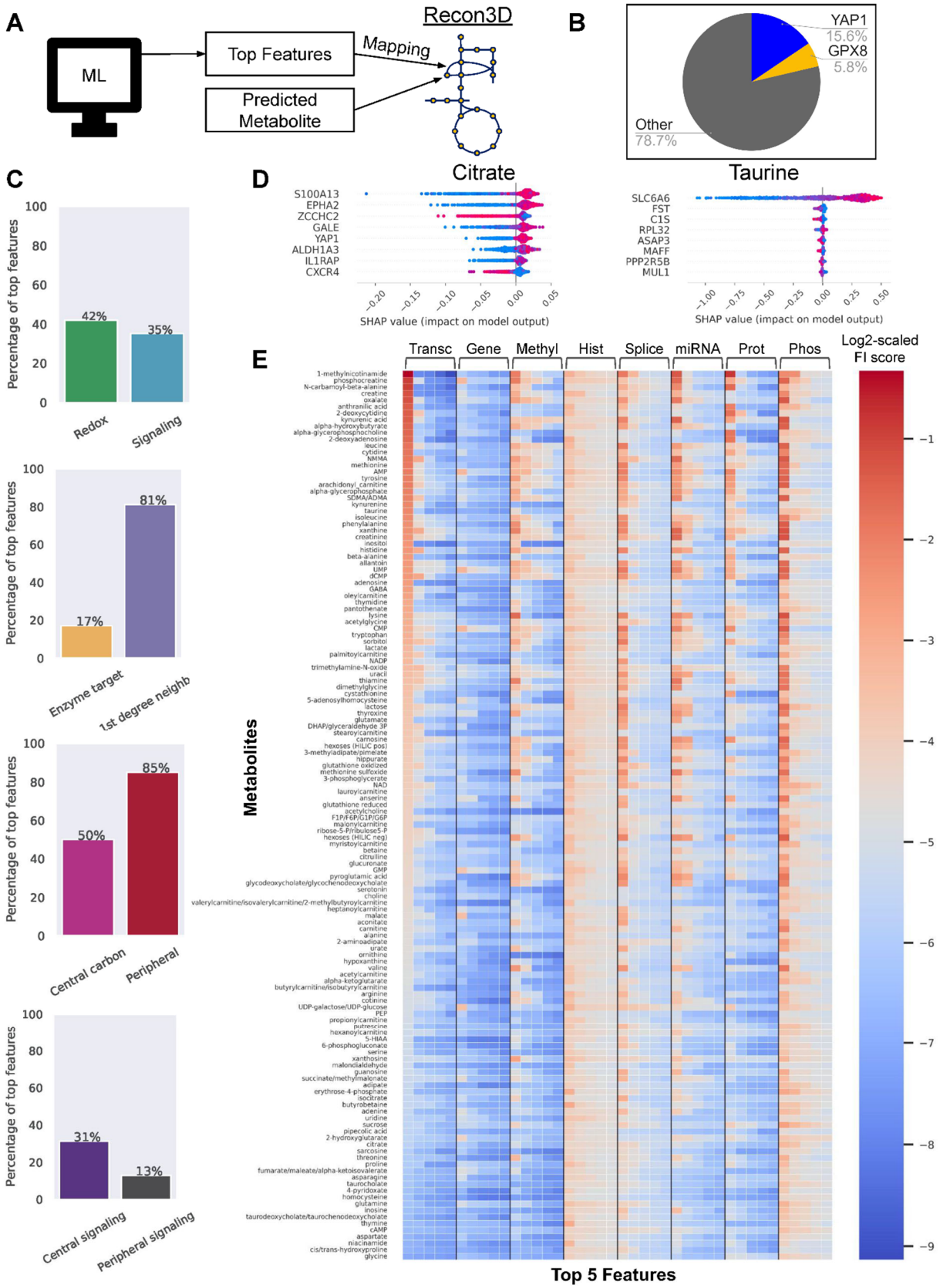


Figure 3. Global insights across all 225 ML models reveal distinct patterns of predictability. A. Workflow for topologically mapping features to Recon3D. Top 1, 2, and 5 features are identified from ML, and are assessed for

overlap with genes present in the reactions for its predicted metabolite (both the enzyme target and the 1st degree neighbor). **B.** Pie chart showing the percentage (of 225) of top features from transcriptomic inputs. YAP1 and GPX8 account for 15.6% and 5.8% of top 1 features, respectively. **C.** Percentage of top 1 features from biologically relevant feature subsets using transcriptomics inputs. 42% (94) of top features (out of 225 metabolite models, metabolic gene-only) are redox enzymes, and 35% (79) of top features (out of 225 metabolite models, whole-data) are signaling enzymes. 17% (16) of top features (out of 95 metabolites that overlapped between CCLE and Recon3D) were involved in the primary reaction for their respective metabolite, and 81% (77) of top features (out of 95 metabolites) were involved within the 1st neighboring reaction for their respective metabolite. 50% (5/10) top features from the central carbon metabolites and 85% (72/85) of peripheral metabolites (from 95 total metabolites that overlapped between CCLE and Recon3D) were involved within the 1st neighboring reaction for their respective metabolites. Lastly, out of 136 non-lipid metabolite models, 31% (5/16) of top features were signaling enzymes for central carbon metabolites, and 13% (15/120) of top features were signaling enzymes for peripheral metabolites. **D.** Contrasting Shapley importance features for citrate and taurine from transcriptomics models. While the taurine transporter SLC6A6 is clearly shown to be the most important feature in predicting relative levels of taurine, citrate does not have one clear top predictor. In this way, feature importance analysis of citrate shows high order, complex influence by multiple features. **E.** Heatmap of log2-scaled feature importance scores for 136 non-lipid metabolites. The top 5 features for each omics class are shown for every metabolite in order of transcriptomics feature importance (descending). This shows a set of metabolites that have 1 strong predictor (top of heatmap) and those metabolites, primarily related to central metabolism, that have multiple predictors (bottom of the heatmap).

Construction of multiomic metabolite regulome networks for top predictive metabolites

To understand how top omics features interact with one another as well as with the metabolite of interest, metabolite regulome networks networks were constructed using top features identified by Shapley importance analysis (Methods) (Figure 4). The Recon8D network containing the top 10 predictors from each of the 8 different omic layers for all 225 metabolites along with their cross validation accuracy is available on the GitHub page. Shapley importance analysis for the top 3 most accurately predicted metabolites for each omics class is shown in S. Figure 2. This analysis sheds light on how top features potentially influence metabolite variation between cells.

We further focused on 1-methylnicotinamide and phosphocreatine, which were the most highly predictable metabolites (lowest P values and largest correlations) across omics classes. Relationships between miRNA and RNA were identified using TarBase[39], and between RNA and proteins using BioGrid[40]. The interconnectedness of these networks highlights the model's ability to identify rational biochemical relationships. For example, nicotinamide N-methyltransferase (NNMT), the enzyme involved in the methylation of nicotinamide, was the top predictor of 1-methylnicotinamide levels from both the transcriptomic and proteomic models. Similarly, solute carrier family 6 member 8 (SLC6A8), a creatine transporter, and creatine kinase (CKMT1A), the enzyme responsible for phosphorylating creatine to create phosphocreatine, were identified as top predictors of phosphocreatine. Shapley analysis also identifies the directionality of those relationships. For example, NNMT is positively correlated with 1-methylnicotinamide concentrations (Figure 4A, right). Similarly, hsa-miR-200c, which was identified as a regulator of NNMT translation via TarBase, is negatively correlated with 1-methylnicotinamide levels. The interconnectedness of the top predictive features across models suggests that the relative levels of these metabolites involve high-order interactions between each of these omics layers.

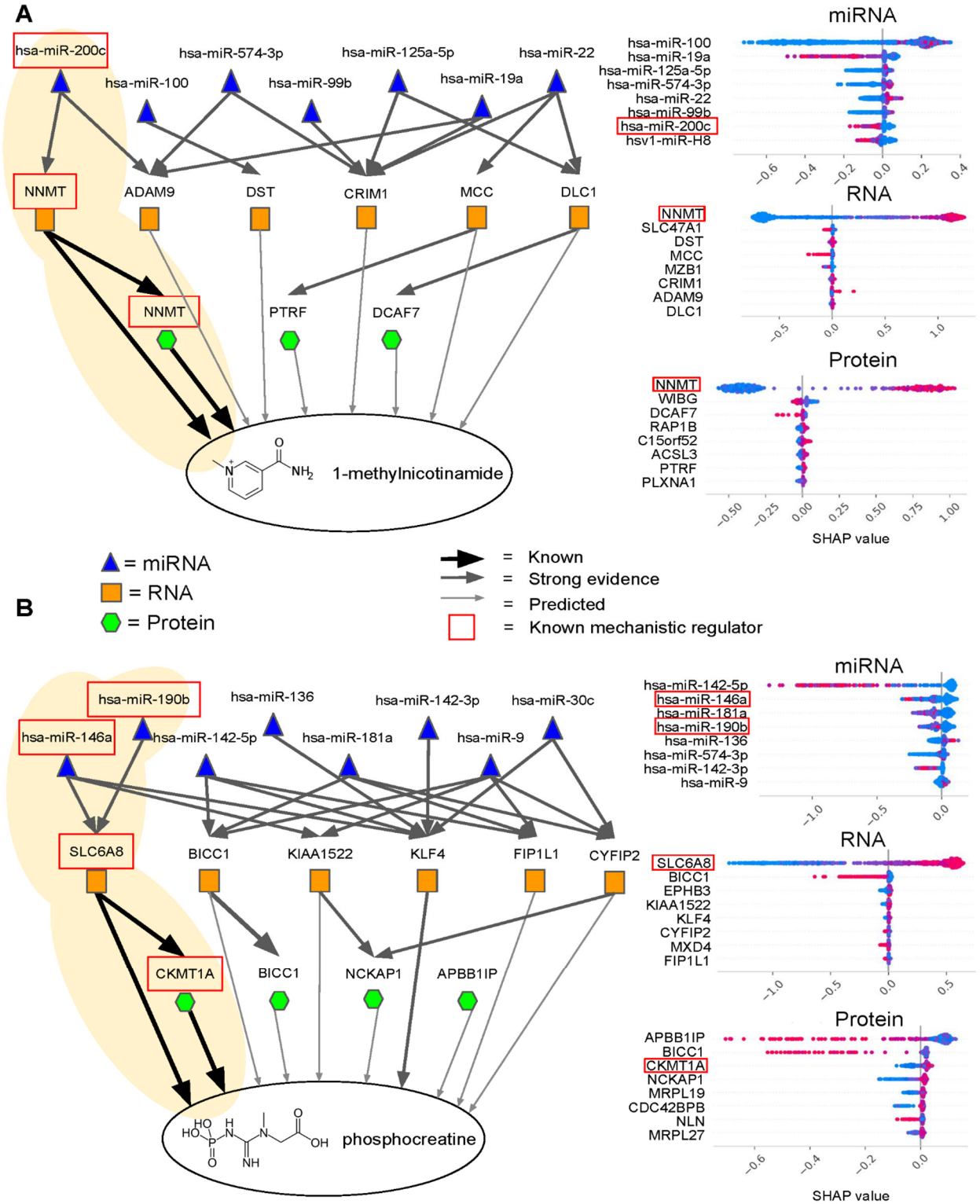


Figure 4. Proposed regulome networks with corresponding Shapley analysis. A. Proposed metabolic regulome network from top 8 features from miRNA, transcriptomic, and proteomic-trained models for predicting 1-methylnicotinamide (highest mean Shap value) as determined by Shapley analysis. **B.** Proposed metabolic influencer network from top 8 features from miRNA, transcriptomic, and proteomic-trained models for predicting phosphocreatine (highest mean Shap values) as determined by Shapley analysis.

YAP1 and GPX8 are globally important features across the metabolome

After identifying the most influential features locally for individual metabolite models, we next pinpointed features that were globally important across the metabolome using the average rank of the features for all 225 metabolite models (Methods, S. Table 8). These results were then scaled by Z-score and top features for each input omics class were identified. Important features of interest were checked for overlap between whole-data models and controls for robustness.

Some top features were several standard deviations above the mean for certain omics classes, such as YAP1, GPX8, and SCD in the transcriptomics models, with Z-scores of 69, 61, and 59, respectively. Well known metabolic regulators such as mTOR also featured in the global top regulators. mTOR phosphorylation was a top 5 feature for several metabolites in central and nucleotide metabolism including oxalate, succinate, xanthine, and phosphogluconate; similarly, phosphorylation of Rictor, which interacts with mTOR was a top feature for amino acids including arginine, consistent with their role in amino acid sensing and regulation[41].

Histone PTMs are increasingly being recognized for their role in sensing and regulating metabolism[42–44]. The CCLE epigenomics data measures total PTM levels and lacks gene-level resolution. Total levels of H3K9 trimethylation and H3K27 di- & trimethylation were the top features in histone PTMs; these are well known repressive marks, and importantly, are also sensitive to the metabolic state of the cell. Histone methylation is sensitive to the level of one-carbon metabolites like SAM and one-carbon sink metabolites such as methyl nicotinamide[43].

Interestingly, top global predictors showed consistencies across omics classes. The most intriguing example of this was YAP1, which was found to be the most important global feature for transcriptomics and phosphoproteomics, as well as the seventh for DNA methylation. YAP1 is a coactivator in Hippo signaling, which is essential for growth signaling, oncogenic adaptations, and the epithelial mesenchymal transition (EMT)[45]. YAP1 was the most important feature for 35 metabolite models using transcriptomics, 2 using proteomics, 5 using DNA methylation, and 132 using phosphoproteomics. Other examples of cross-omic consistencies for top features were GPX8 (transcriptomics and proteomics), miR142 (methylation and miRNA), EPHA2 (transcriptomics and RNA splicing), and ERBB2 (DNA methylation and proteomics). Not only do different omics display conserved global regulators, but globally important features have established relationships. For example, the top global feature from RNA splicing, EPHA2, has been shown to induce accumulation and activation of YAP1, subsequently promoting glutamine metabolism[46].

To validate the association between YAP1 transcripts and the 35 metabolites for which it was the top predictor, we used the DepMap Data Explorer to quantify the relationships between two isoforms of YAP1 protein and each metabolite with linear regression (S. Table 9)[47]. We found that most associations were statistically significant (FDR p-value < 0.05), thus increasing confidence that YAP1 acts as an important predictor for a wide array of metabolites.

Global feature importances from control experiments also identified features integral to specific cellular states. For example, the ML models built using primary or metastatic cell lines revealed

374 distinct miRNA predictors that are known drivers of tumor phenotypes in each condition. miR-
375 142-5p, which has been shown to limit cell invasion and migration[48], was the second most
376 important global feature from the miRNA models and the most important in the primary tumor
377 controls. miR-142 was also a top global feature in the DNA methylation data, 4th in whole-data
378 models and 3rd in primary controls, but is not a top predictor in the metastatic ML models. In the
379 metastatic samples, miR-142 was ranked 233rd out of 734 miRNAs, showing that its relation to
380 metabolite level variation is blunted in the metastatic state. Additionally, miR-181a, which is
381 known to promote proliferation and invasion[49], is the 44th most important feature in the primary
382 tumor control models versus 17th in the metastatic controls, highlighting its influence on
383 metabolism in metastatic cells in particular.
384
385 ***Synthetic lethal interactions are frequent among top metabolite predictors***
386
387 It is difficult to tease out metabolic regulation by individual biomolecules given buffering and
388 combinatorial regulation. By integrating synthetic genetic interactions with multiomic
389 associations, we hypothesized that we can understand how groups of regulators complement
390 each other and maintain metabolic homeostasis.
391
392 Inspired by the synthetic lethal (SL) interaction centered around glutathione highlighted earlier
393 (Figure 5E), we next systematically assessed for SL interactions across all top metabolite
394 predictors in Recon8D. Since metabolite levels are controlled by multiple factors, we
395 hypothesized that the SL interaction analysis can identify compensatory regulators. We compiled
396 curated experimental human SL interactions from SynLethDB and tested if these interactions are
397 enriched among the top 10 predictors of each metabolite (S. table 10)[50]. The transcriptome model
398 had the most striking enrichment for SL interactions among top predictors for each metabolite,
399 with 52 interactions involving 45 metabolites (27 fold higher than random chance, p-value = 1 x
400 10^{-34}) (Figure 5A, 5D).
401
402 Several SL interactions involve a combination of signaling factors, suggesting that combinations
403 of signaling pathways influence variation in metabolite levels, and each gene in the SL pair may
404 compensate for the other's absence. For example, the top predictors of nucleotides and
405 nucleotide-precursors showed several SL interactions involving growth signaling factors like
406 AXL, YAP1, CCN1, PTPN14, CSK, and HRAS. Of note, YAP1 and PTPN14 are both top
407 predictors of hypoxanthine, and they both form an SL pair and also physically interact[51].
408
409 SL interactions among top predictors also highlight potential long range regulatory effects.
410 Several interactions involved both a growth signaling and a redox-related gene and were
411 connected to metabolites in central metabolism (Figure 5D). For example, the top predictive
412 network for lactate contained an SL interaction between the redox enzyme GPX8 and CAV2,
413 which is involved in cell signaling, growth, and lipid metabolism. Similarly, the top predictors
414 for glyceraldehyde-3-phosphate and dihydroxyacetone phosphate, intermediates in glycolysis,
415 also had an SL interaction between GPX8 and the tyrosine kinase AXL.
416
417 ***Synergistic drug combinations converge on nucleotide metabolism pathways***
418

419 Synthetic lethality is directly related to drug synergy, wherein drugs targeting proteins in
420 compensatory pathways show enhanced potency when combined. The abundance of SL
421 interactions among top predictors in Recon8D suggests that drugs that target predictors of
422 specific metabolites are likely to be synergistic. To test this hypothesis, we utilized curated drug-
423 drug interaction data in various cancer cell lines from DrugCombDB which includes 448,555
424 drug combinations covering 2887 unique drugs and 124 human cancer cell lines[52]. The drugs
425 were mapped onto known targets using data from DrugBank[53]. We tested if synergistic
426 interactions were enriched among drugs that target the top 10 predictors of each metabolite. The
427 transcriptome and phosphoproteome models had the strongest enrichment for synergistic
428 interactions among top predictors for each metabolite (Figure 5B). Drug combinations that target
429 top predictors in the transcriptomic model were highly likely to be synergistic with an average
430 bliss interaction score greater than +4, while randomly selected combinations were likely to be
431 antagonistic (bliss score < -5) (Figure 5B).
432
433 Similar to observations with SL interactions, several synergistic interactions involved a
434 combination of signaling pathway inhibitors. We next analyzed metabolites that were associated
435 with synergistic interactions. Notably, nucleotides and nucleotide precursors showed numerous
436 synergistic interactions across a wide range of drugs targeting diverse signaling and metabolic
437 pathways. Guanosine, inosine, pantothenate, thymine, xanthine and hypoxanthine were among
438 the top metabolites associated with most synergistic drug combinations (Figure 5B, S. Table 11).
439 Out of 104 drug combinations associated with these metabolites, 100 were synergistic (S. Table
440 11). Drugs that target diverse signaling pathways may lead to similar downstream effects on
441 nucleotide metabolism. For example, several tyrosine kinase inhibitors including sorafenib and
442 sunitinib are predicted to be associated with these nucleotides (S. table 11). These drugs are
443 known to affect levels of nucleotides including guanine, adenine, adenosine, guanosine,
444 hypoxanthine, and inosine[54]. To further validate these associations, we used the DepMap Data
445 Explorer to quantify the relationships between cellular sensitivity to these signaling drugs and
446 levels of nucleotides, and found several statistically significant associations (S. Table 12)[47].
447 These results confirm the connection predicted by our analysis. In contrast, antagonistic drug
448 interactions were surprisingly associated with metabolites in the TCA cycle. Citrate, isocitrate,
449 and fumarate were the top metabolites associated with most antagonistic drug combinations.
450 Several drug combinations associated with these TCA cycle metabolites involve drugs such as
451 Sirolimus, Everolimus and Tetracycline that are known to reduce the activity of the TCA cycle
452 (S. table 11)[55,56]. Drugs that perturb TCA cycle activity may lead to antagonism with agents that
453 require TCA cycle activity. Thus Recon8D can provide insights into understanding
454 combinatorial metabolic regulation, synthetic lethality, and drug synergy.
455

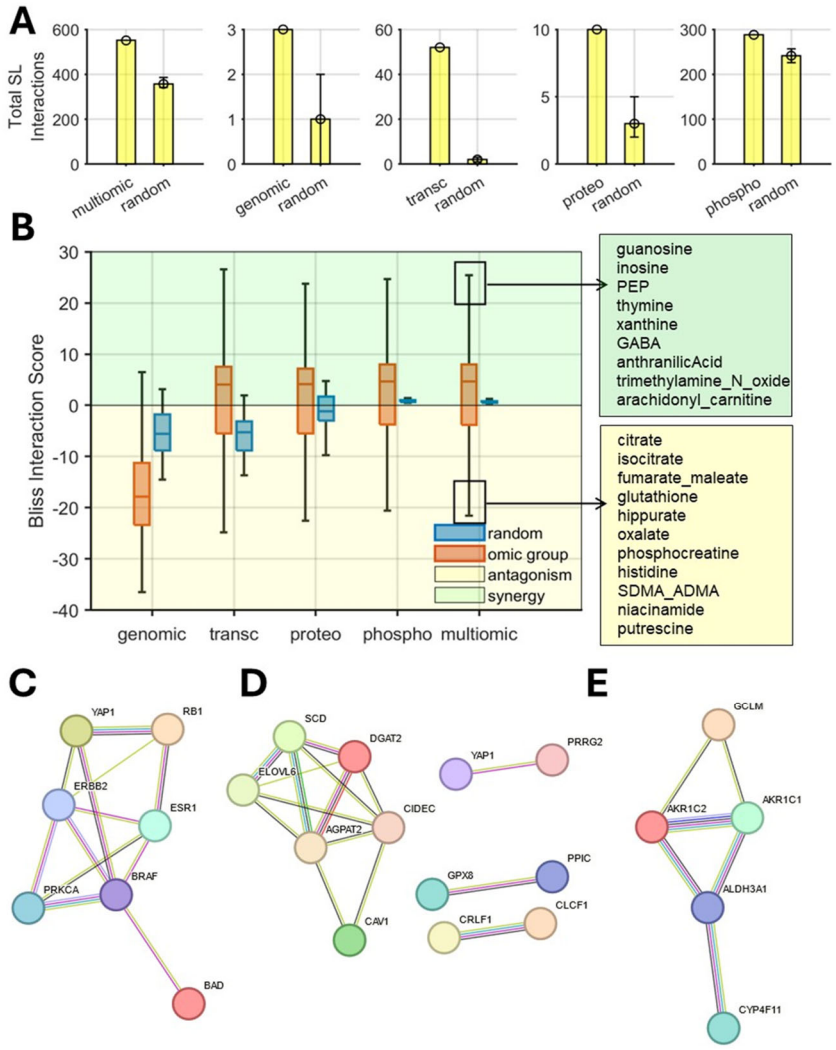


Figure 5. Top metabolite predictors are enriched for synthetic lethality (SL) and drug synergy. A. Frequency of SL gene pairs among top 10 influencers using genomic, transcriptomic, proteomic, phosphoproteomic and multiomic (combination of top 10 influencers from genomic, transcriptomic, proteomic, phosphoproteomic ML models). This is compared with frequency of SL pairs obtained from random sampling using the corresponding omics data. SL pairs were obtained from SynLethDB 2.0 which contains 35,943 human SLs. **B**. Average Bliss drug interaction score (positive values implies synergy) for pairs of drugs that target top predictors of a given metabolite derived from various omics. Top metabolites associated with synergy or antagonism are highlighted. Drug interaction data was derived from DrugCombDB. **C**. High confidence genetic and physical interactions among the top 10 global predictors of phosphoproteomics ML model from the STRING database. **D**. High confidence genetic and physical interactions among the top 10 global predictors of transcriptomics ML model from the STRING database. **E**. High confidence genetic and physical interactions among the top 10 predictors of the metabolite glutathione.

Conclusion

Cellular metabolism is controlled by multiple layers of regulatory networks. Here, we quantified the predictability of metabolomic variation from eight different omics layers. The metabolome was surprisingly highly predictable from transcriptomic data. While the absolute levels of metabolites may depend on thermodynamics, variation may result from differences in

transcriptional regulation. Further, the relationship between transcripts and metabolites is nonlinear, and a combination of transcripts influences metabolite levels. In many cases, there is no direct correspondence between metabolites and enzymes they are involved with. Signaling-related transcripts, proteins, and phosphoproteins were top predictors across omics layers. A second common global predictor was redox metabolism, exemplified by GPX enzymes, that influence glutathione levels and overall redox state of the cell. Levels of histone PTMs, miRNA and phosphoproteins were also highly predictive of the metabolome despite having the fewest features by an order of magnitude. Analysis of feature importance levels of individual metabolites revealed that metabolites in peripheral metabolism, like taurine, were most predictable via levels of enzymes related to those metabolites, while metabolites in central metabolism were more predictable by a combination of signaling and redox regulators. Since metabolites in peripheral metabolism were characterized by a single strong top predictor, we found that these patterns also held up in independent datasets from NCI60 and single-cell omics. Our analysis thus uncovered key systems-level insights on metabolome variation between cells, the generalizability of ML models, and predictability of individual metabolites.

A limitation of our study is that our analysis was done in cancer cell lines in culture and may not reflect metabolism *in vivo*; however, patterns observed here related to transcriptional variation, and topological relatedness of gene-metabolite associations have also been observed *in vivo* in humans and model systems[38,57,58]. Further, the 'guilt by association' strategy used by network inference methods may not identify causative mechanisms. We address this through integration of multiple layers of orthogonal omics information that we found to be consistent with each other[59]. Further, these interactions are robust across hundreds of cell lines in independent datasets spanning multiple tissue lineages and growth conditions. Our model also correctly recalls several known interactions reported in literature. The above suggests that the statistically inferred associations may represent real interactions. Additionally, some high-throughput measurements do not provide a comprehensive panel of molecules that exist within a cell. For example, the metabolite panel used for this study consisted of 225 metabolites, a fraction of those that exist within a cell across distinct cellular compartments. Furthermore, there exist more proteins in human cells than used in this study, and the 90 phosphoproteins present were measured using a targeted approach, and therefore are not representative of the entire cellular array of phosphoproteins. To make a strong inference as to the relative influence of these classes on the metabolome, these analyses would need to be conducted on more comprehensive, untargeted data.

Although our focus was primarily on predictability of the metabolome, we find that top predictors from ML are mechanistically related to the associated metabolites. For example, in the transcriptomics ML model, the metabolites' top predictors were enzymes closely associated, either directly (17%) or first degree neighbor (81%), with those metabolites. The top predictors for each metabolite across different omics layers were highly enriched for genetic and pharmacological interactions. This further supports the ability of our approach to uncover mechanistic interactions and highlights potential multiomic compensatory buffering between top metabolomic predictors. Finally, we find that drugs that target top predictors in nucleotide metabolism are likely to be more synergistic. This builds upon our observation of increased synthetic lethality among the top predictors of each metabolite. Notably, a wide range of signaling drugs intersect with nucleotide metabolism and are predominantly synergistic when

522 used in combination (mean bliss score > 10, S. table 11, Figure 5B). Our analysis suggests that
523 these drugs, which target signaling pathways, may also impact nucleotide homeostasis. This is
524 consistent with a prior study that found combinations of Mek and PI3 Kinase inhibitors show
525 synergy via nucleotide metabolism[60]. In contrast, drug antagonism was associated with
526 metabolites in the TCA cycle. A similar observation was made in bacteria wherein antibiotics
527 that have opposing effects on TCA cycle and bacterial respiration are antagonistic[61]. Several
528 drugs require active growth and metabolism to be potent. Inhibition of TCA cycle may result in
529 growth arrest which may lead to reduced potency of growth-dependent drugs. In contrast, cells
530 with nucleotide imbalances continue to grow[62]. This may result in errors in DNA replication and
531 transcription leading to cell death. Overall, Recon8D may provide a rational framework for
532 metabolic engineering and designing synergistic combination therapies.
533
534
535 **Methods**
536 *Data Acquisition and Preprocessing*
537 Omics from the Cancer Cell Line Encyclopedia (CCLE) were obtained from the Broad Institute
538 (Cambridge, MA) Dependency Map Portal (with the exception of phosphoproteomics, which
539 were obtained from Li et al)[28,63]. These include genomics (copy number variation), epigenomics
540 (bulk histone H3 modifications (MRM) and DNA methylation (RRBS)), transcriptomics
541 (RNAseq), RNA splicing (RNAseq, exon inclusion ratio), miRNA-omics (expression),
542 proteomics (mass spectrometry), phosphoproteomics (RPPA), and metabolomics (LC-MS). The
543 number of cell lines used for training can be found in S. Table 1.
544
545 Data preprocessing involved removing missing values and taking the median value of repeated
546 observations (if required), standardizing cell line or feature names across datasets, and removing
547 features with 0 variance. RNA splicing ensemble IDs were converted to common names using
548 BioTools.fr[64]. Data were scaled by Z-score when used for training/testing to normalize and
549 integrate all datasets. Each data type was merged independently with the metabolomics dataset
550 by common cell line, where each matrix was organized by cell lines (rows) and features
551 (columns). Input datasets for the models containing metabolically relevant genes were filtered
552 for genes present in the Recon3D genome-scale metabolic model[65]. The same data preprocessing
553 steps and scaling methods were used for validation experiments with NCI60 data, which were
554 obtained from the National Cancer Institute CellMiner portal[29,66].
555
556 *Machine Learning*
557 ML models that predict metabolomic variation were trained on single and multiomic data from
558 the CCLE. A separate model was trained and tested for each metabolite using 5-fold cross-
559 validation. Statistically significant predictions were determined using a Bonferroni-corrected P
560 value (0.05/225) to account for false discovery rate. Because the number of features was much
561 greater than the number of observations in these data, it was important to select algorithms that
562 penalize large coefficients and limit overfitting. Random forests were primarily used for this
563 study, although several algorithms were tested for benchmarking, including Ridge Regression,
564 Lasso Regression, XGBoost, and SVM[67–71]. Each of these algorithms is easily interpretable
565 (provides feature importance scores or coefficients) and characterizes different mathematical
566 relationships. The Python script for these algorithms can be found at this GitHub repository.
567

Random forests are comprised of many decision trees, and are classified as an ensemble learning method, where they aggregate many iterations of decisions. This algorithm works by randomly selecting a training set from the input data (termed bootstrapping) and using that training set to dictate a unique decision tree[72]. These predictions are averaged over all the decision trees, which is what allows for more accurate predictions from the overall model than any individual tree could provide[72].

Model Evaluation and Interpretation

Performance of each model was assessed using Pearson's correlation between predicted and true metabolite concentrations. FDR corrected P values were also calculated for assessing significance. Cross-validation was used to assess overall accuracy of each data type, and a final model trained on the entire data set was used to examine feature importances. Feature importance scores were generated from each random forest model and were analyzed locally (for each metabolite) and globally (averaged across all metabolite models for a given omics type). Global importances were sorted by median rank and scaled by Z-score (S. Table 8).

Feature importances for select predicted metabolites were also quantified using Shapley analysis, a game-theoretic approach that calculates feature importances by adding each feature into the model sequentially, and repeating that process for all orderings of features, thereby assessing the influence of each feature on the model as a whole[73,74]. The three metabolites with the lowest P values were selected for Shapley analysis for each data type (S. Figure 2).

The top one, two, five and ten features (quantified by random forest feature importance scores) from genomics, DNA methylation, transcriptomics and proteomics models (both for the whole-data models and the metabolic gene-only models) were mapped back to the Recon3D metabolic model reactions. This was accomplished by matching metabolites from the CCLE panel to metabolites present in the Recon3D metabolic model. Functions from the COBRA toolbox allowed for the identification of all reactions associated with the relevant metabolites (both primary and first degree neighbors), as well as the genes associated with each reaction[75]. Once identified, gene names were converted to BiggIDs and matched to each set of reactions. The number of metabolite models that had at least one top feature present in the relevant set of reactions (out of the 95 metabolites in both the CCLE panel and Recon3D) were compiled (S. tables 6, 7). Code for this process can be found in the GitHub repository.

Model Validation in independent datasets

The top features from CCLE-trained models were isolated and used to analyze relationships between transcripts and metabolites in the NCI60 data and single-cell data. Random forest feature importance scores from NCI60-trained and single-cell-trained models were used in comparison to CCLE outputs, where top features as identified by CCLE models were isolated in NCI60 data to examine if they were larger than by random chance. Z-score-scaled feature importance ranks were used for this analysis. Student's t tests and Kolmogrov-Smirnov tests compared isolated feature importances to the entire feature set. Data were stacked to create one distribution among all metabolites (pooled). Aggregated importances showed significant differences between the top features and overall feature set. These analyses were conducted on sets of 1, 3, 5, 10, 50, and 100 top features from each metabolite model (except for phosphoproteomics, which only had 90 features, and thus the 100 feature subset was not used).

Effect of feature size, missing features and metabolic features

We then explored how well metabolic genes alone could predict the metabolome. Genomic, DNA methylation, transcriptomic, and proteomic datasets were filtered for genes present in the Recon3D metabolic model[65], and new models were trained on the metabolic gene subset. The difference in feature size between omics classes can affect their predictive power, and filtering for genes present in Recon3D standardized these data to provide mechanistic information within the context of known metabolic networks. Each filtered dataset had a similar number of features, which provided an avenue through which to assess the impact of feature size on model accuracy. We first filtered the feature sets for metabolically-relevant genes from Recon3D (S. Figure 5A). With fewer than 2,000 genes, transcriptomics inputs predicted over 99% of metabolites statistically significantly, and other feature sets conferred similar accuracy to whole-data models in predicting each of the 225 metabolites (S. Figure 5B). The median correlation for each omics class was slightly lower in the metabolic gene-only models as compared to the whole-data models, but the distributions were similar overall (S. Figure 5C). Similar accuracies from the metabolic gene-only models and the whole-data models illustrate the lack of impact of feature size on model predictions. We further explored this by training models on a random set of 2000 genes, which also had similar accuracies. While models built on randomly selected features might not provide mechanistic insights, they show that the metabolome is highly predictable within a reasonable range of input feature sizes. Feature importances from metabolic gene-only models and whole-data models were also conserved significantly (S. Figure 5D and E). We hypothesize that the reduced feature size in the proteomic data may account for the reduced accuracy in these models compared to transcriptomics. We interrogated this by comparing the transcriptomic and proteomic datasets to see if there were discrepancies in the feature set that might influence accuracy. Between the 416 genes that were present in the transcriptomics feature set but were absent from proteomics, 11 of those genes were the top 1 predictor in a metabolite model. These missing features explain the discrepancy in accuracy (~15%) between transcriptomic and proteomic inputs, suggesting that access to more comprehensive proteomic data might be more predictive of relative metabolite levels.

Statistical analysis

All P values were corrected using Bonferroni multiple hypothesis correction. Statistical significance of the overlaps with synthetic lethal interactions from SynLethDB 2.0 and drug interactions from DrugCombDB were assessed by creating 50 sets of random top features for each metabolite in all omics models and quantifying the overlap. This was then compared with the actual overlap using a t-test.

Acknowledgments

This work was supported by faculty start-up funds from the University of Michigan (UM), Camille and Henry Dreyfus Foundation, the Rogel Cancer Center at UM, R35 GM13779501 from NIH to SC, and the Advanced Proteogenomics of Cancer (T32 CA140044) Program to RS. We thank Deepak Nagrath, Maciek Antoniewicz, Analisa Difeo and Yatrik Shah for suggestions and critical feedback on the project.

Competing interests

The authors declare no competing interests.

Data availability

All datasets are provided in figures, the supplementary materials, and in the GitHub repository.

Author contributions

S.C, R.S designed and performed research, and wrote the manuscript;.K.S, R.B, S.M, A.C.V, E.K, M.K, performed research.

References

1. Nielsen, J. Systems Biology of Metabolism. *Annu. Rev. Biochem.* **86**, 245–275 (2017).
2. Chubukov, V., Gerosa, L., Kochanowski, K. & Sauer, U. Coordination of microbial metabolism. *Nat. Rev. Microbiol.* **12**, 327–340 (2014).
3. Designing and interpreting 'multi-omic' experiments that may change our understanding of biology - PMC. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7477987/.
4. Cavicchioli, M. V., Santorsola, M., Balboni, N., Mercatelli, D. & Giorgi, F. M. Prediction of Metabolic Profiles from Transcriptomics Data in Human Cancer Cell Lines. *Int. J. Mol. Sci.* **23**, 3867 (2022).
5. Zelezniak, A. *et al.* Machine Learning Predicts the Yeast Metabolome from the Quantitative Proteome of Kinase Knockouts. *Cell Syst.* **7**, 269-283.e6 (2018).
6. Ortmayr, K., Dubuis, S. & Zampieri, M. Metabolic profiling of cancer cells reveals genome-wide crosstalk between transcriptional regulators and metabolism. *Nat. Commun.* **10**, 1841 (2019).
7. Systematic Analysis of Transcriptional and Post-transcriptional Regulation of Metabolism in Yeast - PMC. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5224888/.
8. Brunk, E. *et al.* Characterizing posttranslational modifications in prokaryotic metabolism using a multiscale workflow. *Proc. Natl. Acad. Sci.* **115**, 11096–11101 (2018).
9. Covert, M. W., Knight, E. M., Reed, J. L., Herrgard, M. J. & Palsson, B. O. Integrating high-

692 throughput and computational data elucidates bacterial networks. *Nature* **429**, 92–96
693 (2004).
694 10. Chandrasekaran, S. & Price, N. D. Probabilistic integrative modeling of genome-scale
695 metabolic and regulatory networks in Escherichia coli and Mycobacterium tuberculosis.
696 *Proc. Natl. Acad. Sci.* **107**, 17845–17850 (2010).
697 11. Daran-Lapujade, P. *et al.* The fluxes through glycolytic enzymes in Saccharomyces
698 cerevisiae are predominantly regulated at posttranscriptional levels. *Proc. Natl. Acad. Sci.*
699 **104**, 15753–15758 (2007).
700 12. Oliveira, A. P. *et al.* Regulation of yeast central metabolism by enzyme phosphorylation.
701 *Mol. Syst. Biol.* **8**, 623 (2012).
702 13. Shen, F., Boccuto, L., Pauly, R., Srikanth, S. & Chandrasekaran, S. Genome-scale network
703 model of metabolism and histone acetylation reveals metabolic dependencies of histone
704 deacetylase inhibitors. *Genome Biol.* **20**, 49 (2019).
705 14. Zaslaver, A. *et al.* Just-in-time transcription program in metabolic pathways. *Nat. Genet.* **36**,
706 486–491 (2004).
707 15. Zhao, S. *et al.* Regulation of Cellular Metabolism by Protein Lysine Acetylation. *Science*
708 **327**, 1000–1004 (2010).
709 16. Lempp, M. *et al.* Systematic identification of metabolites controlling gene expression in E.
710 coli. *Nat. Commun.* **10**, 4463 (2019).
711 17. Yu, L. *et al.* Modeling the Genetic Regulation of Cancer Metabolism: Interplay between
712 Glycolysis and Oxidative Phosphorylation. *Cancer Res.* **77**, 1564–1574 (2017).
713 18. Jia, D. *et al.* Elucidating cancer metabolic plasticity by coupling gene regulation with
714 metabolic pathways. *Proc. Natl. Acad. Sci.* **116**, 3909–3918 (2019).
715 19. Gene regulatory network inference in the era of single-cell multi-omics | Nature Reviews
716 Genetics. https://www.nature.com/articles/s41576-023-00618-5.
717 20. Hecker, M., Lambeck, S., Toepfer, S., van Someren, E. & Guthke, R. Gene regulatory

718 network inference: Data integration in dynamic models—A review. *Biosystems* **96**, 86–103
719 (2009).
720 21. Huynh-Thu, V. A., Irrthum, A., Wehenkel, L. & Geurts, P. Inferring Regulatory Networks from
721 Expression Data Using Tree-Based Methods. *PLOS ONE* **5**, e12776 (2010).
722 22. Rychel, K. *et al.* iModulonDB: a knowledgebase of microbial transcriptional regulation
723 derived from machine learning. *Nucleic Acids Res.* **49**, D112–D120 (2021).
724 23. Hu, T. *et al.* Single-cell spatial metabolomics with cell-type specific protein profiling for tissue
725 systems biology. *Nat. Commun.* **14**, 8260 (2023).
726 24. Patti, G. J., Yanes, O. & Siuzdak, G. Metabolomics: the apogee of the omics trilogy. *Nat.*
727 *Rev. Mol. Cell Biol.* **13**, 263–269 (2012).
728 25. Alexandrov, T. Spatial Metabolomics and Imaging Mass Spectrometry in the Age of Artificial
729 Intelligence. *Annu. Rev. Biomed. Data Sci.* **3**, 61–87 (2020).
730 26. Liu, Y. *et al.* High-plex protein and whole transcriptome co-mapping at cellular resolution
731 with spatial CITE-seq. *Nat. Biotechnol.* **41**, 1405–1409 (2023).
732 27. Zhao, M., He, W., Tang, J., Zou, Q. & Guo, F. A comprehensive overview and critical
733 evaluation of gene regulatory network inference technologies. *Brief. Bioinform.* **22**, bbab009
734 (2021).
735 28. Cancer Cell Line Encyclopedia (CCLE). https://sites.broadinstitute.org/ccle/datasets.
736 29. NCI-60 Human Tumor Cell Lines Screen | Discovery & Development Services |
737 Developmental Therapeutics Program (DTP).
738 https://dtp.cancer.gov/discovery_development/nci-60/.
739 30. Temporal fluxomics reveals oscillations in TCA cycle flux throughout the mammalian cell
740 cycle | Molecular Systems Biology.
741 https://www.embopress.org/doi/full/10.15252/msb.20177763.
742 31. The transcriptome dynamics of single cells during the cell cycle | Molecular Systems
743 Biology. https://www.embopress.org/doi/full/10.15252/msb.20209946.

744 32. Icard, P. *et al.* Understanding the Central Role of Citrate in the Metabolism of Cancer Cells
745 and Tumors: An Update. *Int. J. Mol. Sci.* **22**, 6587 (2021).
746 33. Li, L. *et al.* YAP/ACSL4 Pathway-Mediated Ferroptosis Promotes Renal Fibrosis in the
747 Presence of Kidney Stones. *Biomedicines* **11**, 2692 (2023).
748 34. Zhang, D. *et al.* Kynurenine promotes neonatal heart regeneration by stimulating
749 cardiomyocyte proliferation and cardiac angiogenesis. *Nat. Commun.* **13**, 6371 (2022).
750 35. Jyotsana, N., Ta, K. T. & DelGiorno, K. E. The Role of Cystine/Glutamate Antiporter
751 SLC7A11/xCT in the Pathophysiology of Cancer. *Front. Oncol.* **12**, (2022).
752 36. Okazaki, S. *et al.* Synthetic lethality of the ALDH3A1 inhibitor dyclonine and xCT inhibitors in
753 glutathione deficiency-resistant cancer cells. *Oncotarget* **9**, 33832–33843 (2018).
754 37. Brunk, E. *et al.* Recon3D enables a three-dimensional view of gene variation in human
755 metabolism. *Nat. Biotechnol.* **36**, 272–281 (2018).
756 38. Benedetti, E. *et al.* A multimodal atlas of tumour metabolism reveals the architecture of
757 gene–metabolite covariation. *Nat. Metab.* **5**, 1029–1044 (2023).
758 39. Skoufos, G. *et al.* TarBase-v9.0 extends experimentally supported miRNA–gene interactions
759 to cell-types and virally encoded miRNAs. *Nucleic Acids Res.* **52**, D304–D310 (2024).
760 40. Oughtred, R. *et al.* The BioGRID database: A comprehensive biomedical resource of
761 curated protein, genetic, and chemical interactions. *Protein Sci. Publ. Protein Soc.* **30**, 187–
762 200 (2021).
763 41. Zhang, S. *et al.* Regulation of mTORC1 by amino acids in mammalian cells: A general
764 picture of recent advances. *Anim. Nutr.* **7**, 1009–1023 (2021).
765 42. Reid, M. A., Dai, Z. & Locasale, J. W. The impact of cellular metabolism on chromatin
766 dynamics and epigenetics. *Nat. Cell Biol.* **19**, 1298–1306 (2017).
767 43. Su, X., Wellen, K. E. & Rabinowitz, J. D. Metabolic control of methylation and acetylation.
768 *Curr. Opin. Chem. Biol.* **30**, 52–60 (2016).
769 44. King, J., Patel, M. & Chandrasekaran, S. Metabolism, HDACs, and HDAC Inhibitors: A

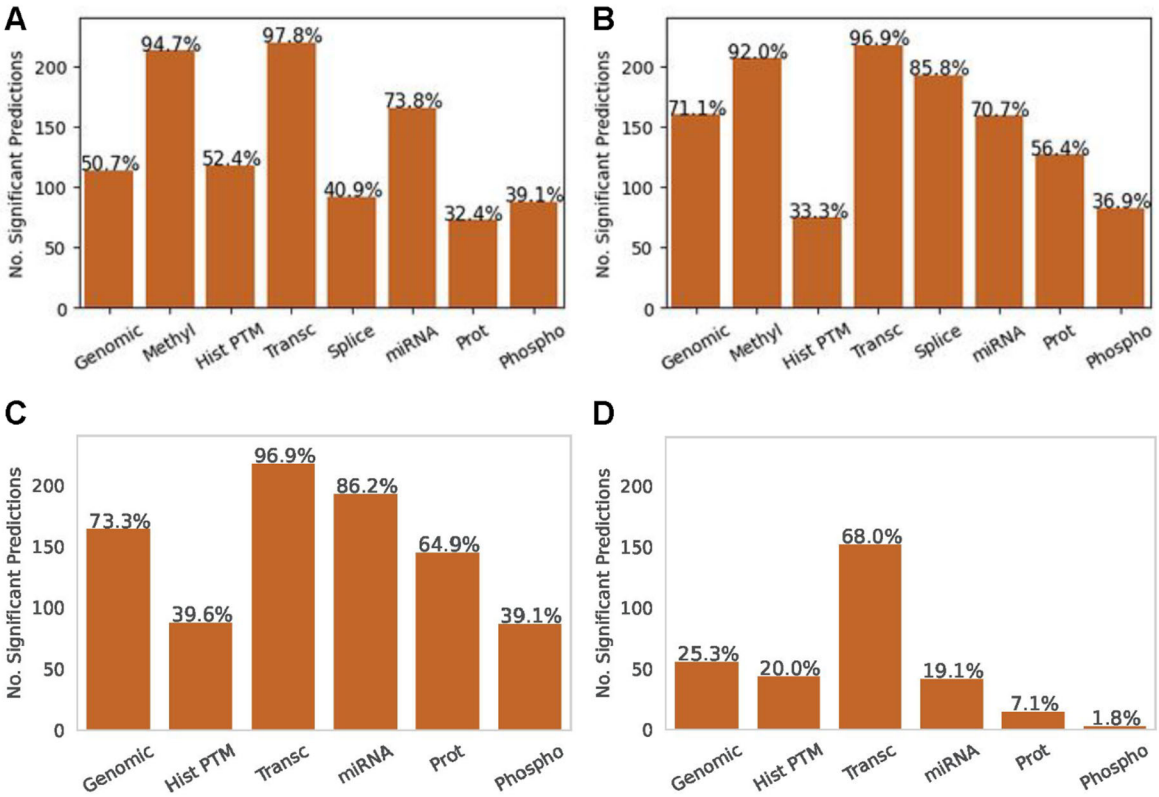
770 Systems Biology Perspective. *Metabolites* **11**, 792 (2021).
771 45. Calvo, F. *et al.* Mechanotransduction and YAP-dependent matrix remodelling is required for
772 the generation and maintenance of cancer-associated fibroblasts. *Nat. Cell Biol.* **15**, 637–
773 646 (2013).
774 46. Edwards, D. N. *et al.* Regulation of cancer glutamine metabolism by EphA2 RTK-dependent
775 activation of transcriptional co-activators YAP and TAZ. *Sci. Signal.* **10**, eaan4667 (2017).
776 47. DepMap Data Explorer 2.0. https://depmap.org/portal/data_explorer_2/.
777 48. Li, H. *et al.* miR-142-5p Inhibits Cell Invasion and Migration by Targeting DNMT1 in Breast
778 Cancer. *Oncol. Res.* **28**, 885–897 (2022).
779 49. Zhai, Z. *et al.* MiR-181a-5p facilitates proliferation, invasion, and glycolysis of breast cancer
780 through NDRG2-mediated activation of PTEN/AKT pathway. *Bioengineered* **13**, 83–95
781 (2022).
782 50. Wang, J. *et al.* SynLethDB 2.0: a web-based knowledge graph database on synthetic
783 lethality for novel anticancer drug discovery. *Database J. Biol. Databases Curation* **2022**,
784 baac030 (2022).
785 51. PTPN14 interacts with and negatively regulates the oncogenic function of YAP | Oncogene.
786 https://www.nature.com/articles/onc2012147.
787 52. Liu, H. *et al.* DrugCombDB: a comprehensive database of drug combinations toward the
788 discovery of combinatorial therapy. *Nucleic Acids Res.* **48**, D871–D881 (2020).
789 53. Wishart, D. S. *et al.* DrugBank 5.0: a major update to the DrugBank database for 2018.
790 *Nucleic Acids Res.* **46**, D1074–D1082 (2018).
791 54. You, X. *et al.* Metabolic reprogramming and redox adaptation in sorafenib-resistant
792 leukemia cells: detected by untargeted metabolomics and stable isotope tracing analysis.
793 *Cancer Commun.* **39**, 17 (2019).
794 55. Chatzispyrou, I. A., Held, N. M., Mouchiroud, L., Auwerx, J. & Houtkooper, R. H.
795 Tetracycline antibiotics impair mitochondrial function and its experimental use confounds

research. *Cancer Res.* **75**, 4446–4449 (2015).

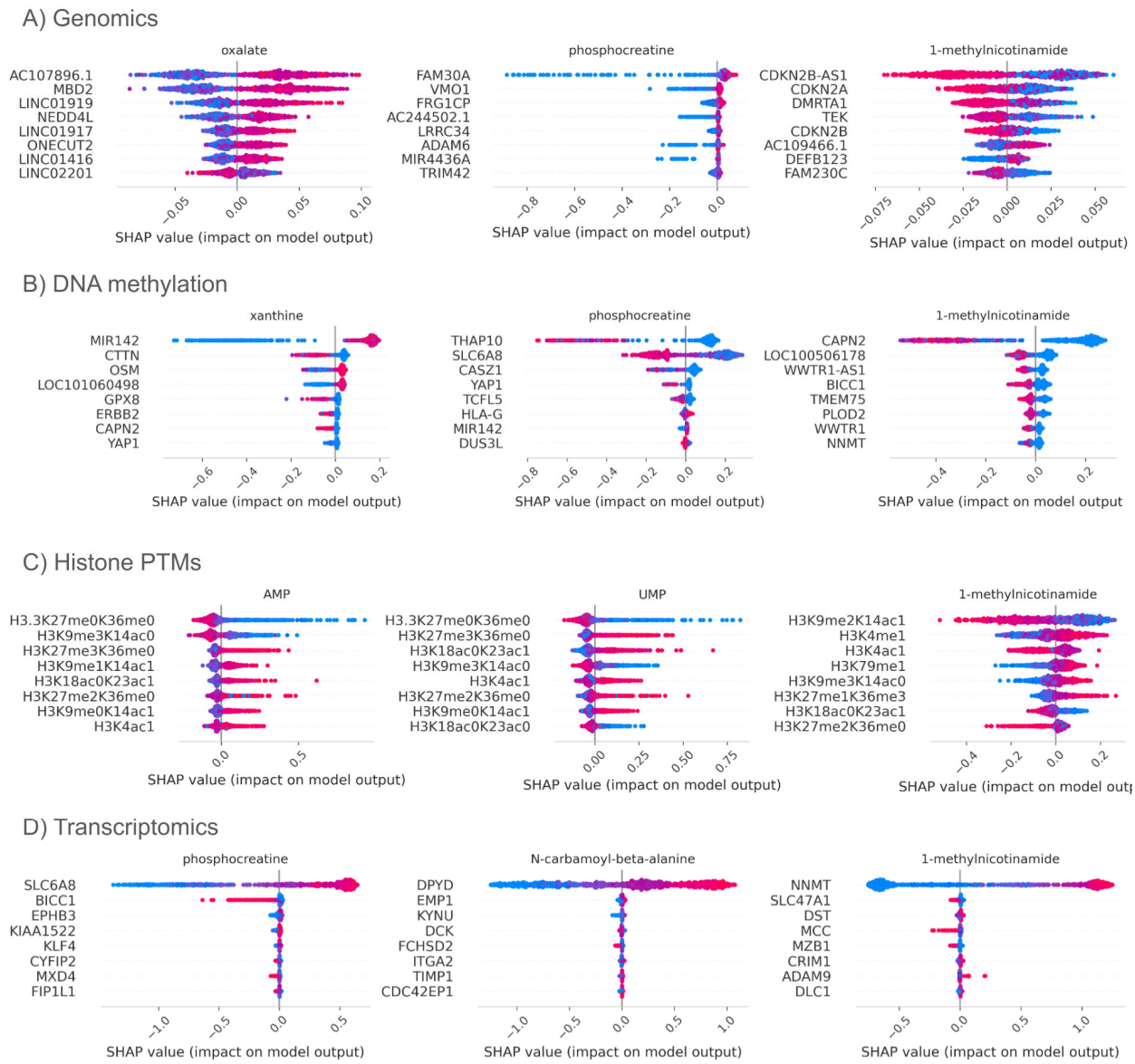
56. Ariaans, G. *et al.* Everolimus decreases [U-13C]glucose utilization by pyruvate carboxylase in breast cancer cells in vitro and in vivo. *Biomed. Pharmacother. Biomedecine Pharmacother.* **173**, (2024).
57. Nanda, S. *et al.* Systems-level transcriptional regulation of Caenorhabditis elegans metabolism. *Mol. Syst. Biol.* **19**, e11443 (2023).
58. Quantifying absolute gene expression profiles reveals distinct regulation of central carbon metabolism genes in yeast | eLife. https://elifesciences.org/articles/65722.
59. Yugi, K., Kubota, H., Hatano, A. & Kuroda, S. Trans-Omics: How To Reconstruct Biochemical Networks Across Multiple 'Omic' Layers. *Trends Biotechnol.* **34**, 276–290 (2016).
60. Shorstova, T. *et al.* Reprogramming of Nucleotide Metabolism Mediates Synergy between Epigenetic Therapy and MAP Kinase Inhibition. *Mol. Cancer Ther.* **20**, 64–75 (2021).
61. Lobritz, M. A. *et al.* Antibiotic efficacy is linked to bacterial cellular respiration. *Proc. Natl. Acad. Sci.* **112**, 8173–8180 (2015).
62. Diehl, F. F. *et al.* Nucleotide imbalance decouples cell growth from cell proliferation. *Nat. Cell Biol.* **24**, 1252–1264 (2022).
63. Li, H. *et al.* The landscape of cancer cell line metabolism. *Nat. Med.* **25**, 850–860 (2019).
64. Dashboard - Genomics Biotools. https://www.biotools.fr/.
65. King, Z. A. *et al.* BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Res.* **44**, D515–D522 (2016).
66. Shankavaram, U. T. *et al.* CellMiner: a relational database and query tool for the NCI-60 cancer cell lines. *BMC Genomics* **10**, 277 (2009).
67. sklearn.ensemble.RandomForestRegressor. *scikit-learn* https://scikit-learn/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html.
68. sklearn.linear_model.Ridge. *scikit-learn* https://scikit-

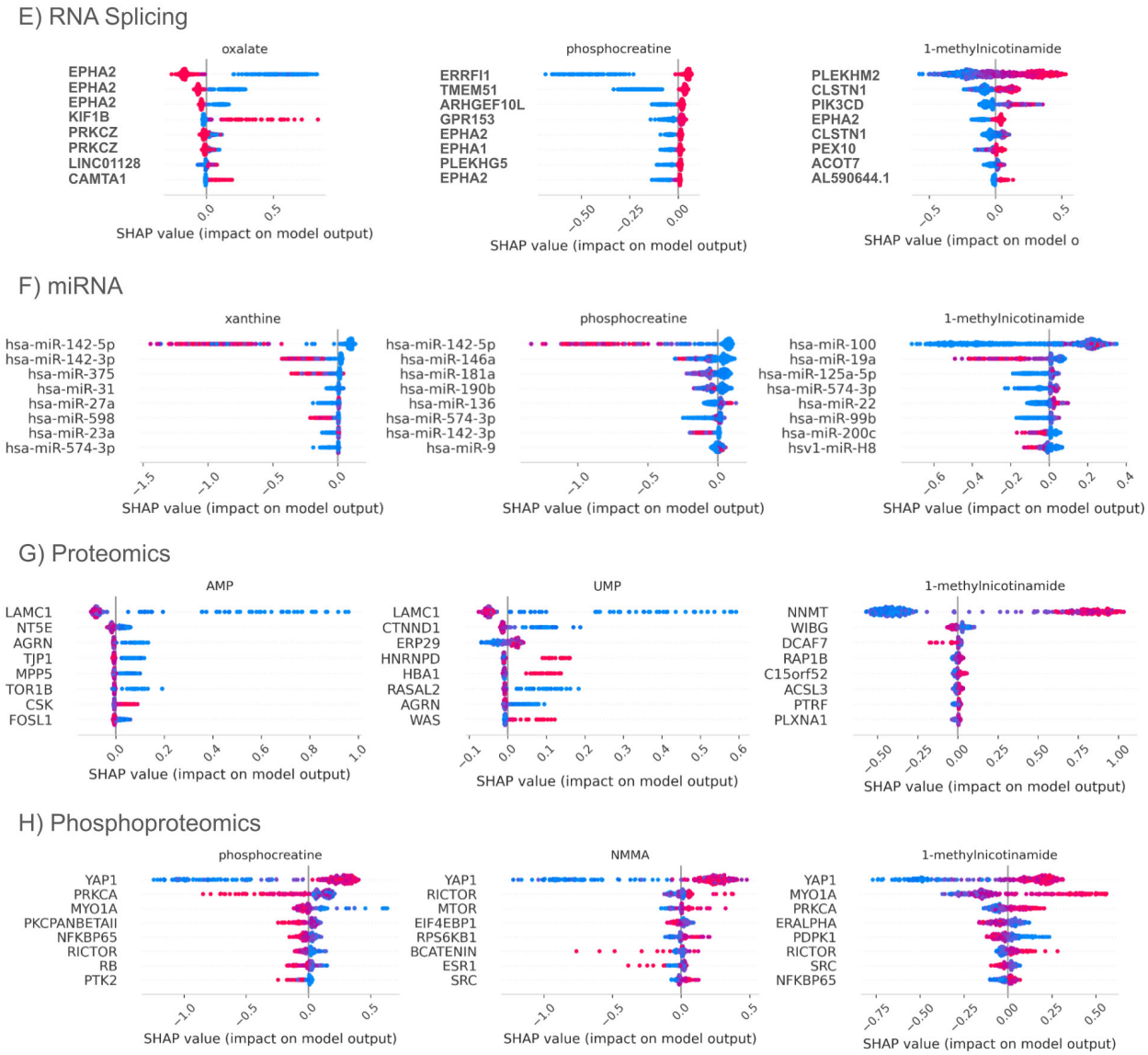
822 learn/stable/modules/generated/sklearn.linear_model.Ridge.html.
823 69. sklearn.linear_model.Lasso. *scikit-learn* https://scikit-
824 learn/stable/modules/generated/sklearn.linear_model.Lasso.html.
825 70. XGBoost Python Package — xgboost 2.1.0-dev documentation.
826 https://xgboost.readthedocs.io/en/latest/python/index.html.
827 71. SVR. *scikit-learn* https://scikit-learn/stable/modules/generated/sklearn.svm.SVR.html.
828 72. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5–32 (2001).
829 73. An introduction to explainable AI with Shapley values — SHAP latest documentation.
830 https://shap.readthedocs.io/en/latest/example_notebooks/overviews/An%20introduction%20
831 to%20explainable%20AI%20with%20Shapley%20values.html.
832 74. Lundberg, S. M. *et al.* Explainable machine-learning predictions for the prevention of
833 hypoxaemia during surgery. *Nat. Biomed. Eng.* **2**, 749–760 (2018).
834 75. Heirendt, L. *et al.* Creation and analysis of biochemical constraint-based models using the
835 COBRA Toolbox v.3.0. *Nat. Protoc.* **14**, 639–702 (2019).
836 76. DepMap Data Downloads. https://depmap.org/portal/download/all/.
837

Supplementary Materials

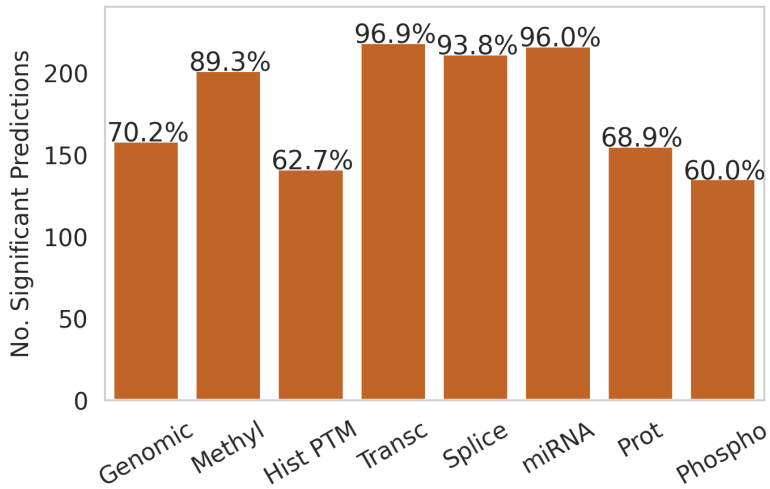


Supplemental Figure 1: Results from training an array of ML models on each input dataset. Bar charts show the number of significantly predicted (FDR p-value < 0.05) metabolites (of 225) per input data type. These algorithms did not outperform Random forests, and therefore were not selected for in-depth analysis. **A.** Ridge Regression **B.** Lasso Regression **C.** XGBoost **D.** Support Vector Machine (SVM) with sigmoid kernel.

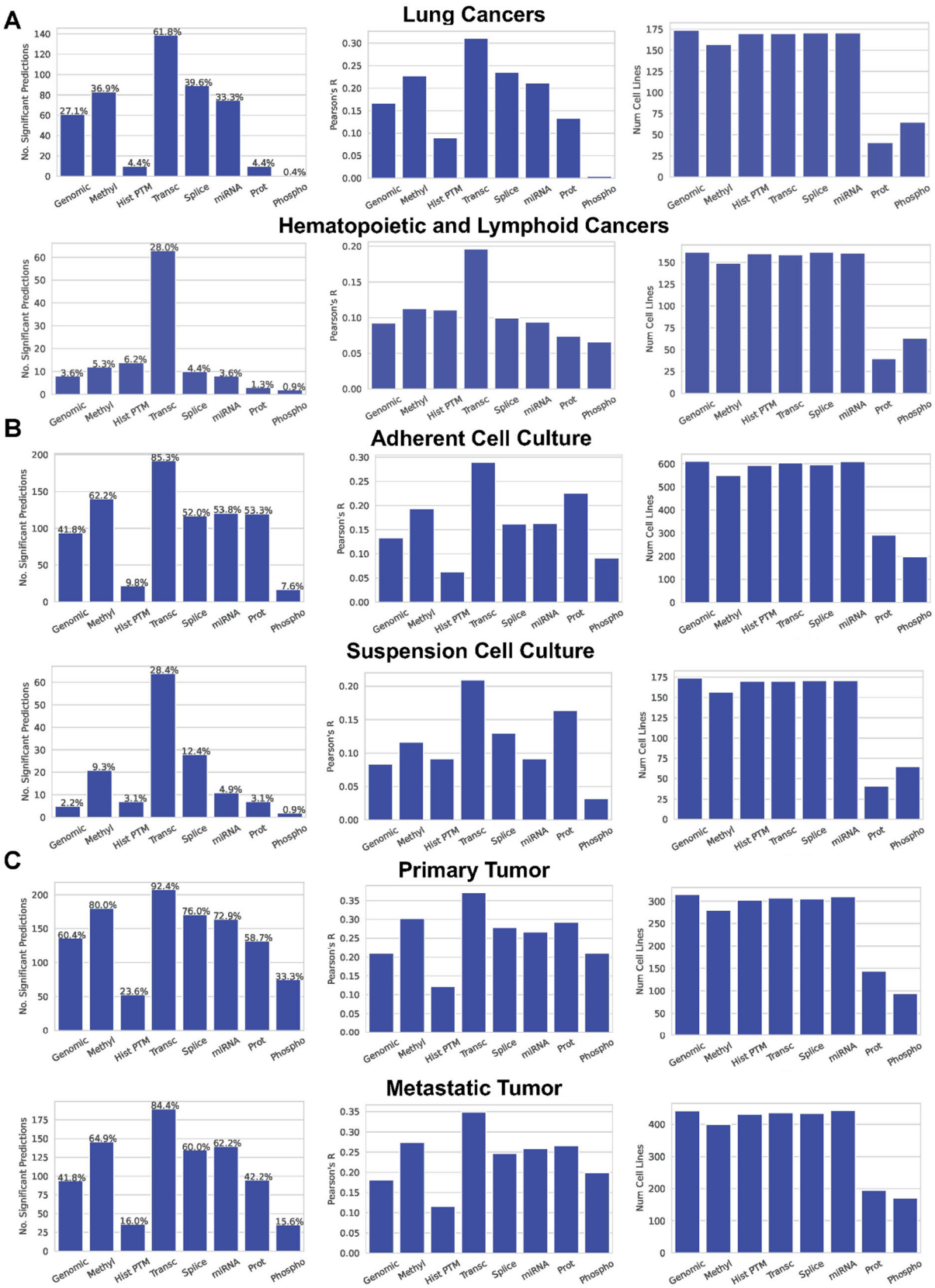

A) Genomics
oxalate
AC107896.1
MBD2
LINC01919
NEDD4L
LINC01917
ONECUT2
LINC01416
LINC02201
−0.05
0.00
0.05
0.10
SHAP value (impact on model output)
phosphocreatine
FAM30A
VMO1
FRG1CP
AC244502.1
LRRC34
ADAM6
MIR4436A
TRIM42
−0.8
−0.6
−0.4
−0.2
0.0
SHAP value (impact on model output)
1-methylnicotinamide
CDKN2B-AS1
CDKN2A
DMRTA1
TEK
CDKN2B
AC109466.1
DEFB123
FAM230C
−0.075
−0.050
−0.025
0.000
0.025
0.050
SHAP value (impact on model output)
B) DNA methylation
xanthine
MIR142
CTTN
OSM
LOC101060498
GPX8
ERBB2
CAPN2
YAP1
−0.6
−0.4
−0.2
0.0
0.2
SHAP value (impact on model output)
phosphocreatine
THAP10
SLC6A8
CASZ1
YAP1
TCFL5
HLA-G
MIR142
DUS3L
−0.8
−0.6
−0.4
−0.2
0.0
0.2
SHAP value (impact on model output)
1-methylnicotinamide
CAPN2
LOC100506178
WWTR1-AS1
BICC1
TMEM75
PLOD2
WWTR1
NNMT
−0.4
−0.2
0.0
0.2
SHAP value (impact on model output
C) Histone PTMs
AMP
H3.3K27me0K36me0
H3K9me3K14ac0
H3K27me3K36me0
H3K9me1K14ac1
H3K18ac0K23ac1
H3K27me2K36me0
H3K9me0K14ac1
H3K4ac1
0.0
0.5
SHAP value (impact on model output)
UMP
H3.3K27me0K36me0
H3K27me3K36me0
H3K18ac0K23ac1
H3K9me3K14ac0
H3K4ac1
H3K27me2K36me0
H3K9me0K14ac1
H3K18ac0K23ac0
0.00
0.25
0.50
0.75
SHAP value (impact on model output)
1-methylnicotinamide
H3K9me2K14ac1
H3K4me1
H3K4ac1
H3K79me1
H3K9me3K14ac0
H3K27me1K36me3
H3K18ac0K23ac1
H3K27me2K36me0
−0.4
−0.2
0.0
0.2
SHAP value (impact on model out
D) Transcriptomics
phosphocreatine
SLC6A8
BICC1
EPHB3
KIAA1522
KLF4
CYFIP2
MXD4
FIP1L1
−1.0
−0.5
0.0
0.5
SHAP value (impact on model output)
N-carbamoyl-beta-alanine
DPYD
EMP1
KYNU
DCK
FCHSD2
ITGA2
TIMP1
CDC42EP1
−1.0
−0.5
0.0
0.5
1.0
SHAP value (impact on model output)
1-methylnicotinamide
NNMT
SLC47A1
DST
MCC
MZB1
CRIM1
ADAM9
DLC1
−0.5
0.0
0.5
1.0
SHAP value (impact on model output)
15



Supplemental Figure 2: Shapley analysis for top 3 metabolites (lowest P value) for each omics class from random forest models. Each dot on the Shapley beeswarm plots corresponds to a cell line observation. Red data points indicate high feature values and blue points indicate low feature values. High SHAP values correspond with higher metabolic concentration outputs from the model, and low SHAP values correspond with lower outputs. Therefore, a gene with a high feature value and SHAP value (red and to the right) illustrates a positive correlation, and a high feature value and a low SHAP value (red and to the left) illustrates a negative correlation. **A.** Genomics **B.** DNA methylation **C.** Histone PTMs **D.** Transcriptomics **E.** RNA Splicing **F.** miRNA **G.** Proteomics **H.** Phosphoproteomics

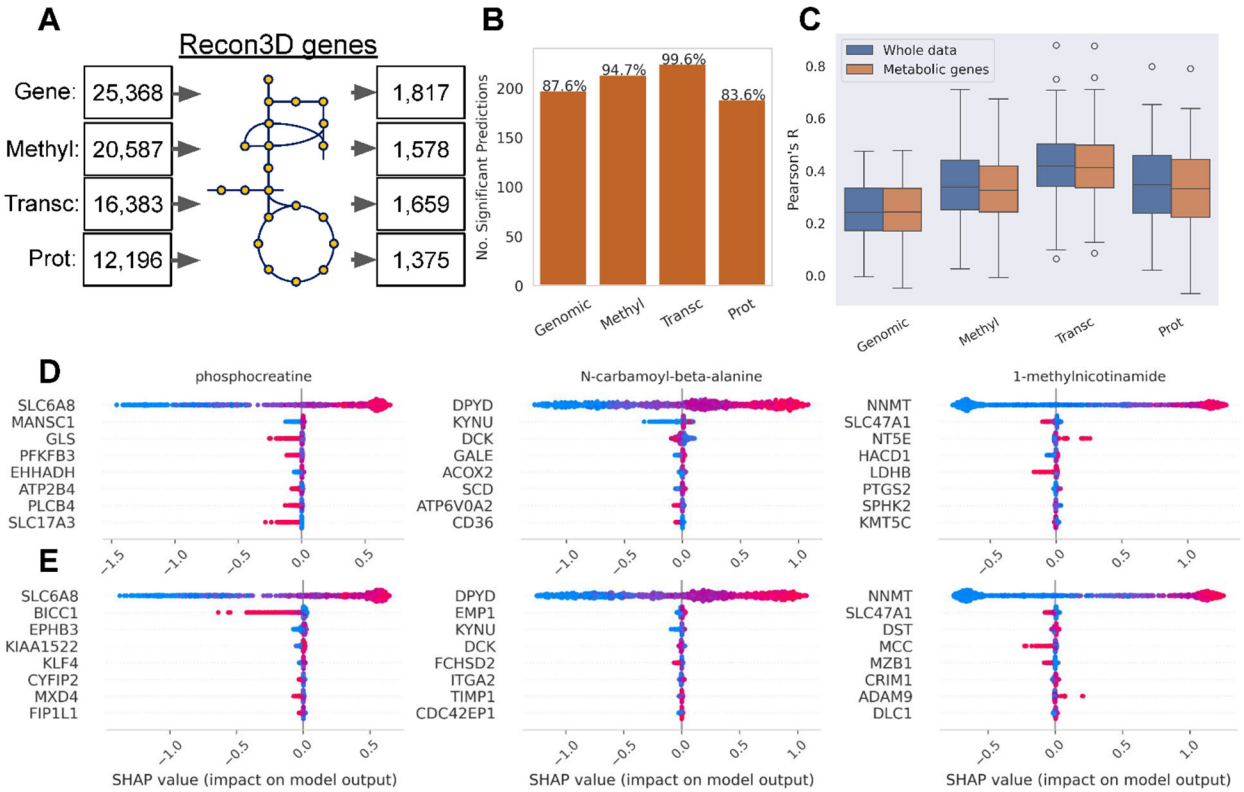


Supplemental Figure 3: Results from Random Forest benchmarking experiments using the top 40 features per input dataset as determined using 75-25% holdout validation. Bar charts show the number of significantly predicted (FDR p-value < 0.05) metabolites per input data type.



Supplemental Figure 4: Control experiment results (random forests), showing the number of significant predictions (FDR p-value < 0.05) out of 225 metabolite models, the average correlations across all models, and the number of

cell lines in each data set (left to right). **A.** Models trained on samples from particular cancer lineages. **B.** Models trained on samples grouped by culture conditions. **C.** Models trained on samples grouped by tumor origin.



Supplemental Figure 5: Results from models trained on genes present in the Recon3D metabolic model. **A.** Number of features in each omics class before and after filtering for genes present in the Recon3D metabolic model. **B.** Number of significant predictions (FDR p-value < 0.05) from metabolic gene-only trained models. **C.** Box plots showing distribution of correlations (true vs predicted) from models built on whole data versus metabolic gene-only models (right). **D.** Shapley analysis for top 3 metabolites (lowest P value) from metabolic gene-only models identifying the top 8 most important features. Each dot on the Shapley beeswarm plots corresponds to a cell line observation. Red data points indicate high feature values and blue points indicate low feature values. High SHAP values correspond with higher metabolic concentration outputs from the model, and low SHAP values correspond with lower outputs. Therefore, a gene with a high feature value and SHAP value (red and to the right) illustrates a positive correlation, and a high feature value and a low SHAP value (red and to the left) illustrates a negative correlation. **E.** Shapley analysis for top 3 metabolites (lowest P value) from models built on whole data identifying the top 8 most important features. These show that feature importances from metabolic gene-only models and whole-data models were conserved. For example, NNMT is the top predictor for 1-methylnicotinamide, SLC6A8 for phosphocreatine, and DPYD for N-carbamoyl-beta-alanine in both the metabolic gene-only and whole-data models. DPYD is involved in N-carbamoyl-beta-alanine biosynthesis.

891 **Supplemental Table 1:** Number of cell lines used to train each model

	Genomics	DNA methylation	Histone PTMs	Transcript-omics	RNA Splicing	miRNA	Proteomics	Phospho-proteomics
Cell Lines	921	828	893	904	900	916	373	296

892

893 **Supplemental Table 2:** Top 40 global features as identified from 75-25% holdout validation random forests.

Genomics	Histone PTMs	DNA methylation	Transcriptomics	RNA Splicing	miRNA	Proteomics	Phosphoproteomics
LSP1P4	H3K9me3K14ac0	CAPN2	FAM162A	EPHA2	hsa-miR-142-5p	PLIN2	YAP1
ADAM6	H3K27me3K36me0	OSM	KLF13	CLSTN1	hsa-miR-142-3p	PTPN13	BRAF
FAM30A	H3.3K27me0K36me0	MIR142	MVD	PLEKHM2	hsa-miR-22	CEP290	PRKCA
TIMP3	H3K4me2	CTTN	SCD	ACOT7	hsa-miR-150	HSPA2	RB
MEF2C-AS1	H3K27me2K36me0	HDAC7	INSIG1	RPL22	hsa-miR-100	FUNDC2	ERBB3
FAM242B	H3K27me1K36me3	YAP1	YAP1	LZIC	hsa-miR-375	TWSG1	SRC
MTRNR2L6	H3K4me1	TCF3	DMAC1	CLSTN1	hsa-miR-24	PLEK	EIF4EBP1
BTNL3	H3K9ac1K14ac0	SLC12A4	RNLS	DVL1	hsa-miR-31	MARCKS	ACC
VMO1	H3K27me1K36me0	EPHA2	MT2A	KIF1B	hsa-miR-23a	DENND4A	BAD
CTSW	H3K18ac0K23ac1	DENND4B	DIAPH2	MINOS1	hsa-miR-21	RCSD1	ERBB2
LILRA6	H3K4ac1	UMODL1-AS1	PLIN2	SZRD1	hsa-miR-130a	RAD17	ERALPHA
LINC00221	H3K27me1K36me2	LANCL2	TRAF3IP2	GNB1	hsa-miR-27a	TREX1	MTOR
SEMA6D	H3K9me2K14ac1	PCF11	SSPN	ICMT	hsa-miR-574-3p	UTP23	ESR1
UGT2B17	H3K27me2K36me2	VAMP1	EBP	PARK7	hsa-miR-148a	COBLL1	NFKBP65
PRSS2	H3K9me0K14ac0	MIR6797	TNFRSF11A	MIIP	hsa-miR-222	FLAD1	MAPK
FAM230C	H3K79me1	KIAA1522	CCDC80	NADK	hsa-miR-574-5p	AKR1B15	AMPK
MIR5195	H3K9me1K14ac0	AMBRA1	CPT1A	SZRD1	hsa-miR-342-3p	TP53	S6
FRG1CP	H3K79me2	PRDM11	IMMP2L	PLOD1	hsa-miR-155	RBKS	MYO1A
AC244502.1	H3K9me3K14ac1	ERBB2	NOS3	CAMTA1	hsa-miR-29a	LHPP	STAT3
H3Y1	H3K79me0	TFDP1	HSPA6	PLEKHM2	hsa-miR-135b	CCBL2	TSC2
PRSS1	H3K4me0	GULP1	GSR	CDK11A	hsa-miR-221	ENDOG	MAP2K1
MIR4436A	H3K27me0K36me0	THOC6	ST6GALNAC4	GNB1	hsa-miR-1975	PPP1R12A	JUN
DUXAP8	H3K9me2K14ac0	TMEM75	RIDA	MFN2	hsa-miR-455-3p	ACSF3	AKT1
RPS9	H3K27me2K36me1	MIR3176	AMOTL2	SLC35E2B	hsa-miR-378	GTPBP6	FOXO3
TGIF2LY	H3K27me0K36me3	SHROOM1	PTGES2	UBE2J2	hsa-miR-23b	ATP5G2	PDPK1
MTAP	H3K27me3K36me1	WDR66	YIPF1	GNB1	hsa-miR-106a+hsa-miR-17	FADS2	P27
TBC1D3J	H3K9me0K14ac1	ARRB2	H1-4	CLSTN1	hsa-miR-210	MICALL2	RAF1
CDKN2A	H3K18ac0K23ac0	MCTS2P	CDC42EP3	RPL22	hsa-miR-149	MAN2B2	MAPK14
GLIDR	H3K56me0	LOC100128531	PHPT1	CAMTA1	hsa-miR-766	ATXN1	BCATENIN
ANXA8	H3K9me1K14ac1	MGA	MTAP	CCNL2	hsa-miR-193b	SNX11	YBX1
ACOT1	H3K18ac1K23ac0	VAV1	ANKRD37	GNB1	hsa-miR-1274a	IRF9	GSK3ALPHABETA
MIR3675	H3K18ac0K23ub1	MIR6772	SQLE	MINOS1	hsa-miR-96	ZFYVE20	CHEK1
TBC1D3F	H3K9ac1K14ac1	SDHAP3	RPL36AL	LINC01128	hsa-miR-181a	SFN	MAPK8
MIR4275	H3K27me0K36me2	INPP5D	COA4	NECAP2	hsa-miR-9	GMPR2	GSK3
OR11H1	H3K27ac1K36me0	SYDE1	PAOX	MRPL20	hsa-miR-183	NT5C3A	RPS6KB1
AL591379.1	H3K27ac1K36me2	STX16-NPEPL1	NDUFA9	MINOS1	hsa-miR-29b	PHLDB2	NDRG1
TRG-AS1	H3K18ac1K23ac1	ZNF512	CDK2AP2	MFN2	hsa-miR-152	CRABP1	MET
POTEH	H3K27ac1K36me3	TMEM14C	SH3BP1	PEX10	hsa-miR-605	M6PR	SHC1
SIGLEC14	H3K27me1K36me1	FXYD1	AK4	ICMT	hsa-miR-129-3p	FGD1	PKCPANBETAII
CHEK2	H3K27me0K36me1	DDI2	CAMKMT	ATP13A2	hsa-miR-365	OTUD4	EPHA2

894
895
896

Supplemental Table 3: Robustness of top features from CCLE transcriptomics in the NCI60 data for full data models and Recon3D genes only. T-test and Kolomogorov-Smirnov test results from the top 1, 3, 5, 10, 50, and 100 CCLE features isolated in NCI60 data as compared to the whole data. Top features for predicting each metabolite were tested against the whole data for the entire metabolite panel (pooled). Data were stacked to create a single vector of feature importance scores across all metabolites for the top feature subsets as well as whole data. Feature importances were used to assess the robustness of the random forest ML framework used to train the CCLE models. Feature importances were significantly higher for all top feature subsets in the pooled analyses using transcriptomics. Top features for individual metabolites also showed significantly higher correlations and feature importance scores for a subset of metabolites (FDR p-value < 0.05). Only using the top 1 feature, CMP, sucrose, GABA, anthranilic acid, N-carbamoyl-beta-alanine, and adenosine displayed conserved top features, which are all in peripheral metabolism. * = significant (p-value < 0.05)

	Whole-data		Recon3D-only	
	T-test	KS-test	T-test	KS-test
Top Features	P value	P value	P value	P value
1	2.54E-05*	6.50E-02	3.02E-02*	5.2E-02
3	2.51E-05*	7.30E-02	2.48E-02*	6.7E-02
5	2.22E-06*	4.30E-02*	1.73E-02*	2.9E-02*
10	4.23E-10*	2.90E-03*	4.12E-03*	2.4E-02*
50	1.44E-10*	4.60E-03*	8.05E-07*	4.28E-06*
100	2.92E-17*	3.68E-05*	3.69E-13*	1.52E-10*

Supplemental Table 4: Robustness of top features from CCLE genomics, miRNA, proteomics, and phosphoproteomics in NCI60 models illustrated by T-test and Kolmogorov-Smirnov tests using pooled feature importances. * = significant (p-value < 0.05)

	Genomics		**miRNA**		**Proteomics**		**Phosphoproteomics**	
	T-test	KS-test	T-test	KS-test	T-test	KS-test	T-test	KS-test
Top Features	P value	P value	P value	P value	P value	P value	P value	P value
1	1.16E-01	5.64E-01	8.80E-01	9.3E-01	2.35E-01	5.7E-01	8.5E-01	9.7E-01
3	2.40E-02*	3.39E-01	4.00E-01	6.6E-01	1.54E-01	4.2E-01	7.7E-2	3.5E-01
5	2.40E-02*	3.46E-01	7.10E-02	1.8E-01	6.80E-02	1.5E-01	9.70E-04*	1.99E-02*
10	1.80E-03*	2.06E-01	7.62E-07*	1.91E-05*	8.80E-02	2.96E-01	1.00E-04*	8.00E-03*
50	4.75E-07*	7.66E-03*	4.79E-07*	3.46E-23*	2.23E-02*	1.88E-01	2.05E-11*	5.75E-08*
100	5.06E-12*	6.10E-04*	2.06E-33*	6.75E-26*	2.60E-03*	1.05E-01	N/A	N/A

Supplemental Table 5: Robustness of top features from CCLE transcriptomics in single-cell models illustrated by T-test and Kolmogorov-Smirnov tests using pooled feature importances. * = significant (p-value < 0.05)

	T-test	KS-test
Top Features	P value (pooled)	P value (pooled)
1	2.00E-01	4.9E-01
3	2.10E-02*	1.6E-01
5	6.00E-04*	2.6E-02*
10	6.06E-08*	3.80E-04*
50	8.74E-16*	3.80E-07*
100	2.12E-25*	4.36E-11*

Supplemental Table 6: Number of metabolite models trained on the whole data that had at least one top feature map to Recon3D for the reactions directly associated with or within one neighbor of the predicted metabolite (out of 95 common metabolites).

		Top Feature		Top 2 Features		Top 5 Features	
Whole data	Total Model Training Features	Primary Reaction	1st degree neighbor	Primary Reaction	1st degree neighbor	Primary Reaction	1st degree neighbor
Genomics	25,368	0	1	0	3	0	10
Methylation	20,587	1	6	7	10	8	20
Transcriptomics	16,383	4	21	14	27	20	50
Proteomics	12,196	3	26	17	32	21	43

Supplemental Table 7: Number of metabolite models trained on the Recon3D metabolic gene subset that had at least one top feature map to Recon3D for the reactions directly associated with or within one neighbor of the predicted metabolite (out of 95 common metabolites).

		Top Feature		Top 2 Features		Top 5 Features	
Recon3D Genes	Total Model Training Features	Primary Reaction	1st degree neighbor	Primary Reaction	1st degree neighbor	Primary Reaction	1st degree neighbor
Genomics	1,817	1	56	2	78	7	88
Methylation	1,578	8	60	10	75	16	90
Transcriptomics	1,659	16	77	23	90	32	92
Proteomics	1,375	20	74	22	83	27	89

929 **Supplemental Table 8:** Global feature importance ranks for the top 20 features from each omics class. Median rank
930 was used to determine the most important features over all 225 metabolite models for each feature.

Global Average Rank	Gene	Hist	Methyl	Transc	Splice	miRNA	Prot	Phos
1	LSP1P4	H3K9me3K14ac0	CAPN2	YAP1	EPHA2	hsa-miR-142-5p	LAMC1	YAP1
2	FAM30A	H3K27me2K36me0	OSM	GPX8	CLSTN1	hsa-miR-142-3p	PURA	PRKCA
3	ADAM6	H3.3K27me0K36me0	CTTN	SCD	PLEKHM2	hsa-miR-22	PLIN2	BRAF
4	FRG1CP	H3K27me1K36me3	MIR142	DGAT2	CLSTN1	hsa-miR-150	FTL	ACC
5	MTRNR2L6	H3K4me1	HDAC7	PPIC	DVL1	hsa-miR-100	TLK1	BAD
6	LILRA6	H3K27me3K36me0	LOC101060498	IKZF1	ACOT7	hsa-miR-31	HBA1	RB
7	H3Y1	H3K27me1K36me2	YAP1	KLF13	KIF1B	hsa-miR-27a	COL6A3	ESR1
8	VMO1	H3K4me2	C5orf66-AS2	FAM162A	MINOS1	hsa-miR-375	TSPAN6	MYO1A
9	TIMP3	H3K9ac1K14ac0	INPP5D	CLCF1	EPHA2	hsa-miR-23a	SLC39A14	ERALPHA
10	FAM242B	H3K27me1K36me0	AMOTL2	CAV1	SZRD1	hsa-miR-24	GLUL	ERBB2
11	CTSW	H3K27me2K36me2	SHROOM1	FAM53B	LZIC	hsa-miR-96	ZFYVE16	SRC
12	PRSS1	H3K9me1K14ac0	THOC6	PPARD	RPL22	hsa-miR-222	PSTPIP2	STAT3
13	FAM230C	H3K18ac0K23ac1	NRROS	CCN1	GNB1	hsa-miR-183	ECHDC3	ERBB3
14	MIR4436A	H3K9me2K14ac1	GTPBP2	EPHA2	RERE	hsa-miR-193b	TWSG1	MTOR
15	AC244502.1	H3K4ac1	HIST1H4G	TNFRSF12A	GNB1	hsa-miR-342-3p	LRPAP1	MAPK
16	PRSS2	H3K27me0K36me2	ERBB2	IRF2BP2	GNB1	hsa-miR-148a	NMES1	BCATENIN
17	LINC00221	H3K79me1	LPCAT1	F12	MFN2	hsa-miR-574-3p	PTPN6	NFKBP65
18	CDKN2A	H3K9me0K14ac0	ARRB2	WWTR1	RPL22	hsa-miR-29a	RAB13	MAP2K1
19	DUXAP8	H3K27me3K36me1	WIZ	POLR1H	ICMT	hsa-miR-221	STX7	EIF4EBP1
20	SEMA6D	H3K79me0	ZFAT	BATF2	MINOS1	hsa-miR-455-3p	LAMB2	FOXO3

931

Supplemental Table 9: Correlation of two protein isoforms of YAP1 with metabolites for which ML found YAP1 to be the top feature from transcriptomics inputs. Correlations were obtained from proteomic and metabolomic CCLE data using the DepMap portal[76]. Associations between YAP1 and the metabolites it predicted well from ML were assessed for statistical significance using a Bonferroni corrected P value (FDR p-value < 0.05/35).

	YAP1 (P46937-7)		YAP1 (P46937-8)	
Metabolite	Pearson's R	P value	Pearson's R	P value
aconitate	0.117	3.08E-02	0.159	2.11E-03
dCMP	-0.212	7.60E-05*	-0.302	2.97E-09*
hypoxanthine	-0.134	1.33E-02	-0.122	1.90E-02
oxalate	-0.325	7.02E-10*	-0.377	5.92E-14*
uracil	0.249	2.94E-06*	0.3	3.93E-09*
xanthine	0.34	1.04E-10*	0.437	1.13E-18*
3-methyladipate/pimelate	-0.112	3.80E-02	-0.136	8.75E-03
α-hydroxybutyrate	-0.204	1.39E-04*	-0.258	5.12E-07*
Met	0.241	6.61E-06*	0.356	1.59E-12*
Leu	0.245	4.27E-06*	0.363	5.34E-13*
Ile	0.187	4.94E-04*	0.297	5.97E-09*
Phe	0.231	1.59E-05*	0.338	2.41E-11*
Tyr	0.204	1.45E-04*	0.308	1.35E-09*
Trp	0.16	2.97E-03*	0.261	3.72E-07*
dimethylglycine	0.311	3.87E-09*	0.368	2.69E-13*
SDMA/ADMA	0.326	6.55E-10*	0.409	2.22E-16*
NMMA	0.333	2.68E-10*	0.403	7.12E-16*
allantoin	0.28	1.42E-07*	0.372	1.36E-13*
kynurenic acid	0.38	3.24E-13*	0.427	8.61E-18*
carnosine	0.352	1.85E-11*	0.39	6.57E-15*
creatinine	0.324	7.93E-10*	0.386	1.46E-14*
methionine sulfoxide	0.291	4.27E-08*	0.346	7.71E-12*
propionylcarnitine	0.188	4.67E-04*	0.179	5.52E-04*
C32:0 PC	-0.248	3.23E-06*	-0.222	1.68E-05*
C34:1 PC	-0.246	4.19E-06*	-0.28	4.46E-08*
C38:4 PC	-0.38	3.04E-13*	-0.352	2.99E-12*
C40:6 PC	-0.306	7.46E-09*	-0.298	5.25E-09*
C18:2 SM	-0.179	8.61E-04*	-0.238	3.74E-06*
C16:1 SM	0.213	7.07E-05*	0.252	9.27E-07*
C16:0 CE	0.399	1.62E-14*	0.453	4.09E-20*
C18:2 CE	0.317	1.82E-09*	0.401	9.55E-16*
C18:1 CE	0.349	3.06E-11*	0.416	7.07E-17*
C18:0 CE	0.224	2.75E-5*	0.255	6.51E-7*
C20:4 CE	0.334	2.25E-10*	0.424	1.43E-17*
C56:8 TAG	-0.313	3.19E-09*	-0.37	1.82E-13*