Accurate prediction of gene deletion phenotypes with Flux Cone Learning

Charlotte Merzbacher,[1] Oisin Mac Aodha,[1] and Diego A. Oyarzún[1, 2, 3]

[1)] *School of Informatics, University of Edinburgh, UK*

[2)] *School of Biological Sciences, University of Edinburgh, UK*

[3)] *Corresponding author: d.oyarzun@ed.ac.uk*

(Dated: 15 June 2025)

Summary: Predicting the impact of gene deletions is crucial for biological discovery, biomedicine, and biotechnology. For example, identifying lethal deletions is key for new cancer therapies or antimicrobial treatments that bypass drug resistance. In biotechnology, non-lethal deletions are a powerful strategy to redirect chemical flux toward production of high-value compounds for the food, energy, and pharmaceutical sectors, using genetically engineered cells as an alternative to petrochemicals. Owing to the cost and complexity of large-scale deletion screens, there is a growing interest in computational models that can leverage such data for predictive modelling. Here, we present Flux Cone Learning, a best-in-class framework for predicting the impact of metabolic gene deletions on many phenotypes of interest. Flux Cone Learning is based on high-dimensional Monte Carlo sampling of the metabolic space in tandem with supervised learning of fitness scores from deletion screens. We demonstrate unparalleled predictive accuracy for metabolic gene essentiality in organisms of varied complexity (*Escherichia coli*, *Saccharomyces cerevisiae*, Chinese Hamster Ovary cells), outperforming the gold standard predictions of Flux Balance Analysis. Our method does not rely on an optimality principle and thus can be applied to a range of organisms where cellular objectives cannot be encoded as an optimization task. We showcase the versatility of Flux Cone Learning in other phenotypes by training a predictor of small molecule production from deletion screening data. Flux Cone Learning provides a widely applicable framework for phenotypic prediction and lays the groundwork for the development of metabolic foundation models across the kingdom of life.

Understanding the phenotypic impact of gene deletions often relies on genome-wide screens of mutational effects. Deletion screens based on high-throughput technologies such as RNAi or CRISPR-Cas9[1–3] have revealed foundational insights across numerous applications, including the genetic basis of disease[4,5], drug target discovery[6,7], and genetic engineering[8,9]. Computational methods hold substantial promise as a complement to experimental deletion screens, for example for filling gaps in coverage, extrapolating predictions to new variants or conditions, and aiding experimental design.

In the case of metabolic genes, the gold standard is Flux Balance Analysis (FBA), a computational method that predicts metabolic phenotypes by combining genome-scale metabolic models[10] (GEM) with an optimality principle[11]. This technique can model many metabolic tasks such as growth capabilities in various substrates[12], cell-specific auxotrophies[13], or responses to drug interventions[14]. FBA is particularly effective at predicting gene essentiality in microbes, i.e. whether a gene deletion leads to cell death. For various model microbes[15], FBA predicts metabolic gene essentiality with high accuracy, but its predictive power drops when applied to higher-order organisms where the optimality objective is unknown or nonexistent[16–18]. Other methods build on FBA to extend essentiality predictions to other relevant tasks; for example, gene Minimal Cut Sets[19] were developed to identify combinations of deletions that block specific cellular functions, with particular success for predicting synthetic lethal genes in cancer[20]. Alternative strategies for essentiality prediction include network-based methods that leverage protein-protein interactions[21] or sequence-based approaches that employ machine learning to extract predictive features from DNA or protein sequences[22,23]. Progress in deep learning has led to a renewed interest in sequence-based approaches with increased predictive power[24–26].

Here, we describe Flux Cone Learning (FCL), a versatile machine learning strategy for predicting deletion phenotypes from the shape of the metabolic space. Flux Cone Learning utilizes mechanistic information encoded in a genome-scale metabolic model to produce a large corpus of training data for each deletion. These data can be paired with experimental fitness readouts for a phenotype of interest and then employed for training predictive models with supervised learning. Flux Cone Learning can be adapted to multiple prediction tasks, provided that the fitness scores correlate with metabolic activity. This includes prediction of metabolic signals already encoded in a genome-scale metabolic model, e.g. growth rate or the activity of specific pathways, as well as non-metabolic readouts absent from the model

but associated with metabolic activity. We show that FCL produces the most accurate predictions of metabolic gene essentiality, surpassing FBA predictions in all tested organisms. Crucially, FCL predictions do not require an optimality assumption and thus can be applied to a broader range of organisms than FBA. We demonstrate the flexibility of FCL for predicting other deletion phenotypes by building a predictor of small-molecule synthesis from deletion screen data.

LEARNING THE SHAPE OF THE METABOLIC SPACE

Our approach is based on learning the shape of the metabolic space of an organism through random sampling. Flux Cone Learning (FCL) has four components (Figure 1): a genome-scale metabolic model (GEM), a Monte Carlo sampler to produce features for model training, a supervised learning algorithm trained on fitness data, and a score aggregation step. A GEM is defined by:

$$\mathbf{S}\mathbf{v} = 0, \tag{1}$$

$$V_i^{\min} \leq v_i \leq V_i^{\max}, \tag{2}$$

where $\mathbf{S}$ is an $m \times n$ integer matrix describing the metabolic stoichiometry, $\mathbf{v}$ is an $n-$dimensional vector of metabolic fluxes, and $(V_i^{\min}, V_i^{\max})$ are flux bounds that can be used to model gene deletions through a Gene-Protein-Reaction (GPR) map. Upon deletion of gene g_j, the GPR determines which flux bounds need to be zeroed out in the GEM, i.e. by setting $V_i^{\min} = V_i^{max} = 0$ in Eq. (1); a single gene deletion can affect more than one reaction flux in the GEM. From a geometric standpoint, a GEM defines a convex polytope in a high-dimensional space, which is known as the *flux cone* of an organism[27]. The dimensionality of the cone equals that of the null space of $\mathbf{S}$, which for current GEMs can be up to several thousand dimensions depending on model complexity (Supplementary Figure S1).

Flux Cone Learning relies on the observation that gene deletions perturb the shape of the flux cone, because zeroing out the flux bounds in Eq. (2) alters the boundaries of the polytope. The correlations between such geometric changes and a phenotype of interest can then be learned with supervised learning algorithms trained on experimental fitness scores. To test if the shape differences between flux cones can be captured from random

samples, we first sampled five metabolically diverse pathogens (*Bordetella pertussis*, *Pseudomonas aeruginosa*, *Helicobacter pylori*, *Mycobacterium tuberculosis*, *Streptococcus pneumoniae*) from programatically generated GEMs to avoid confounders introduced by variations in model quality[28]. We trained a variational autoencoder (VAE)[29] based on neural networks to compute low-dimensional representations of each species cone[30], using a large set of Monte Carlo samples of metabolic reactions shared across the five species and removing species-specific reactions. The learned representations are well separated across species, despite being trained on reactions shared by diverse species (Supplementary Figure S2). This suggests that the cone geometry can be learned from Monte Carlo samples, and offers a path toward the construction of metabolic foundation models across many species and genomic perturbations.

To train predictive models of deletion phenotypes, FCL utilizes a Monte Carlo sampler to capture the shape of each deletion cone (Figure 1). A supervised machine learning model is then trained on the flux samples alongside measured phenotypic fitness labels for each deletion; all samples in a deletion cone get assigned the same label. FCL does not prescribe the choice of machine learning model and can be applied to both regression and classification tasks. The feature matrix for model training has $k \times q$ rows and n columns, where k is the number of gene deletions, q is the number of flux samples per deletion cone, and n is the number of reactions in the GEM. This approach leads to large datasets; for example, in the case of the iML1515 model of *Escherichia coli*[12], acquiring 100 Monte Carlo samples for the 2,712 reactions and 1,502 gene deletions leads to a dataset over 3Gb in single-precision floating-point format. In the final step, FCL aggregates sample-wise predictions with a majority voting scheme to produce deletion-wise predictions.

BEST PREDICTIVE ACCURACY OF METABOLIC GENE ESSENTIALITY

We first tested FCL as a predictor of gene essentiality in *Escherichia coli*, which has the best curated GEM in the literature and thus mitigates the impact of poor model quality on predictive performance. When tested across different carbon sources, FBA delivers a maximal accuracy of 93.5% correctly predicted genes for *E. coli* growing aerobically in glucose with biomass synthesis as optimization objective[12]. We employed FCL using N =1,202 gene deletions (80%) with $q = 100$ samples/cone for training a binary classifier of gene essential-

ity; the biomass reaction was removed from training to prevent the model from learning the correlation between biomass and essentiality that support FBA predictions (Supplementary Figure S3, Supplementary Table S3). This led to a training dataset with N =120,285 samples and n =2,712 features. We opted for a random forest classifier as a suitable compromise between model complexity and interpretability. Test results in a random set of N =300 held-out genes (20%) outperformed the state-of-the-art FBA predictions in accuracy, precision and recall, achieving an average 95% accuracy for all test genes across training repeats (Figure 2A); moreover, FCL achieved a 1% and 6% improvement in classification of non-essential and essential genes, respectively, as compared to FBA (Figure 2B).

Inspection of sample-wise prediction scores show that a small number of deletions get incorrectly classified, likely due to GEM misspecifications (Figure 2C). Interpretability analysis revealed that a few as ∼100 reactions can explain model predictions, with top predictors being enriched for transport and exchange reactions (Figure 2D). Thanks to its excellent predictive power, FCL can be employed to define a distance metric between deletions and the wild type strain, with statistically significant differences between non-essential and essential deletions (Supplementary Figure S4).

To investigate which factors determine FCL performance, we first retrained the model with sparser sampling data and fewer gene deletions (Figure 2E); predictive accuracy dropped in both cases, but models trained on as few as 10 samples/cone already matched the current state-of-the-art FBA accuracy. We additionally retrained FCL with earlier and less complete GEMs for *E. coli* and found that only the smallest GEM (iJR904) displayed a statistically significant drop in performance (Figure 2F). Given the high dimensionality of the feature space, we retrained the random forest model on a reduced feature set computed with Principal Component Analysis, but this resulted in lower accuracy in all tested cases, possibly because correlations between essentiality and small changes in cone shape can only be captured in a high-dimensional feature space. We also explored the use of deep learning models, including feedforward and convolutional neural networks, but these did not improve performance even when trained on larger data with more than q =5,000 samples/cone (not shown). This is likely because such models are deliberately overparameterized to accommodate highly nonlinear correlations among features, but in our case flux samples are linearly correlated through the stoichiometric constraint in Eq. (1).

We tested FCL for essentiality prediction in *Saccharomyces cerevisiae* and Chinese Ham-

ster Ovary (CHO) cells, two more complex organisms with well-curated GEMs[31,32] widely employed for synthesis of heterologous proteins and metabolites. These models have 52% and 130% more reactions than *E. coli*, respectively, leading to a higher dimensionality of the flux cone and more features for training. Following the same data generation protocol as in Figure 2A, we trained FCL on 80% of deleted genes with similar sample density. FCL achieved better classification results than FBA in both organisms across multiple performance scores (Figure 3, Supplementary Table S5). We found that FCL showed similar prediction errors as FBA, with a tendency to misclassify some essential genes as non-essential, likely due to the class imbalance in the training data (most genes are non-essential).

The performance improvements against FBA predictions on three organisms of varied complexity (*E. coli*, *S. cerevisiae*, and CHO cells) suggests that FCL provides the most accurate predictions for metabolic gene essentiality in the literature. This result further demonstrates that optimality assumptions are not required for prediction of metabolic gene essentiality, which supports the early evidence provided by recent studies[18,33].

PREDICTION OF SMALL MOLECULE SYNTHESIS

To explore the power of FCL for predicting other phenotypes, we focused on small molecule biosynthesis in microbial strains engineered with heterologous pathways[34]. Recent studies have showcased the utility of genome-wide deletion screens for improving production titers[35,36]. Non-essential deletions can both suppress or boost metabolite production; for example, deletions that disrupt enzymatic co-factor homeostasis are deleterious for product synthesis, while other non-essential deletions can re-direct metabolic flux away from non-essential pathways toward increased production[8].

We focused on a large deletion screen of *S. cerevisiae* mutants engineered to synthesize betaxanthin[35], a tyrosine-derived pigment widely employed in the food sector. The screen includes a total of 4,223 gene deletions, out of which $N = 811$ genes code for metabolic enzymes present in the latest yeast GEM[32]. Fitness scores for each deletion strain were quantified via betaxanthin autofluorescence averaged across four non-clonal cultures (Figure 4A). We first binned betaxanthin autofluorescence into three classes for low, medium and high producing cultures. We employed FCL to build a 3-class classifier that predicts betaxanthin synthesis using Monte Carlo sampling of the deletion GEMs. Due to the im-

balanced data size across classes (17.1%, 67.2% and 15.7%, resp.), we trialed various model architectures in combination with re-balancing strategies (Figure 4B); the best performing model delivered promising accuracy (69.8%). We observed a tendency to underpredict the high-producing deletions due to these being underrepresented in the training data, though high producer accuracy improvements between 5.5% and 28.3% could be obtained via various class balancing techniques (Figure 4C, Supplementary Table S6).

To the best of our knowledge, this is the first demonstration that small molecule synthesis can be predicted from deletion screening data, and adds to the growing number of tools to predict metabolite production using various data modalities and computational approaches[37–39]. Since FCL relies purely on the wild-type GEM and experimental fitness readouts, it does not require extending the GEM with a heterologous pathway of interest, which can be beneficial in cases where pathway stoichiometry is not well characterized.

DISCUSSION

With the rapid progress in high-throughput genetic engineering and automated screening technologies, there is a growing opportunity to utilize such data for building predictors of the phenotypic response to gene deletions. Flux Cone Learning offers a general strategy to detect correlations between metabolic genotypes and phenotypic readouts. It combines experimental fitness data with mechanistic knowledge into a machine learning system able draw phenotypic predictions for a specific gene deletion.

Our model evaluations demonstrate that FCL outperforms the state-of-the-art FBA predictions of metabolic gene essentiality. Flux Balance Analysis has the advantage of being a zero-shot predictor, in the sense that it does not need to be trained on fitness data. Instead, FBA draws predictions based on a biological optimality assumption; for microbial systems, maximal growth rate or biomass synthesis rate are well validated metabolic objectives. But for most organisms beyond the microbial world, such optimality assumptions are not warranted and there is no consensus on how to define suitable metabolic objectives for higher-order organisms[15,17]. Various studies have built strategies to accommodate for the multiobjective nature of metabolic optimality[40,41] or to reverse engineer metabolic objectives[42,43] and tradeoffs[17,44]. Yet even in cases where an optimal objective of the wild-type can be validated, there is little evidence that such an objective would be preserved

upon a gene deletion. Mutants are likely to be subject to different evolutionary pressures that shift their genetic programs away from the physiological objectives of the wild-type. Flux Cone Learning thus allows essentiality predictions in a much wider range of cell types than current methods, including those with unknown optimality principles such as human cell lines[45] or the gut microbiome[46], as well as prediction of other deletion phenotypes beyond essentiality, such as single-cell metabolic capabilities[47], synthetic lethality[48], or gain-of-function deletions[49]. Although FCL is agnostic to the fitness score employed for training, its effectiveness is limited by the strength of correlations between metabolic activity and the phenotype of interest. In the case of gene essentiality, for example, FCL works well because deletions in pathways that supply key metabolites for growth can strongly impact cell viability. Other phenotypes with weaker or no associations to metabolic activity may require additional modalities of data for accurate prediction.

The integration of learning algorithms with genome-scale metabolic models has shown substantial promise for improved predictivity across various tasks[17,18,39,50–52]. The novel paradigm behind FCL is to learn the shape of the metabolic space through random sampling of genome-scale metabolic models. High-dimensional sampling remains a key challenge in statistical learning[53], because in high dimensions samples tend to be equidistant and concentrate on the boundaries of the space[54]. While expectation would suggest that dense sampling is needed to accurately capture the cone geometry, in our tests we consistently found that accurate FCL models could be trained from shallow sampling with as few 100 samples per deletion. We hypothesize this is a case of the curse-of-dimensionality working to our benefit: to capture changes to the cone, FCL only requires samples at the boundary, and therefore a relatively small number of samples are sufficient for accurate prediction.

The performance of FCL suggests that that predictive representations of metabolic capabilities can be learned from Monte Carlo sampling of genome-scale metabolic models. This advancement lays the groundwork for the development of metabolic foundation models via large sampling across species, growth conditions and deletion genotypes, thus extending the breadth of biological foundation models across additional data modalities[55,56]. We expect that Flux Cone Learning will open new routes for computational prediction of many cellular phenotypes, with applications in basic discovery, biotechnology and future therapies.

ACKNOWLEDGMENTS

CM and DAO were supported by the United Kingdom Research and Innovation (grant EP/S02431X/1, UKRI Centre for Doctoral Training in Biomedical AI).

REFERENCES

[1]Wang, T. *et al.* Identification and characterization of essential genes in the human genome. *Science* **350**, 1096–1101 (2015).

[2]Cacheiro, P. *et al.* Human and mouse essentiality screens as a resource for disease gene discovery. *Nature Communications 2020 11:1* **11**, 1–16 (2020).

[3]Bock, C. *et al.* High-content CRISPR screening. *Nature Reviews Methods Primers* **2**, 1–23 (2022).

[4]Shohat, S. & Shifman, S. Genes essential for embryonic stem cells are associated with neurodevelopmental disorders. *Genome Research* **29**, 1910–1918 (2019).

[5]Tsherniak, A. *et al.* Defining a Cancer Dependency Map. *Cell* **170**, 564–576.e16 (2017).

[6]Chang, L., Ruiz, P., Ito, T. & Sellers, W. R. Targeting pan-essential genes in cancer: challenges and opportunities. *Cancer Cell* **39**, 466–479 (2021).

[7]Minikel, E. V. *et al.* Evaluating drug targets through human loss-of-function genetic variation. *Nature* **581**, 459–464 (2020).

[8]Rancati, G., Moffat, J., Typas, A. & Pavelka, N. Emerging and evolving concepts in gene essentiality. *Nature Reviews Genetics* **19**, 34–49 (2018).

[9]Meyer, A. J., Segall-Shapiro, T. H., Glassey, E., Zhang, J. & Voigt, C. A. Escherichia coli "Marionette" strains with 12 highly optimized small-molecule sensors. *Nature Chemical Biology* **15**, 196–204 (2019).

[10]King, Z. A. *et al.* Bigg models: A platform for integrating, standardizing and sharing genome-scale models. *Nucleic acids research* **44**, D515–D522 (2016).

[11]Orth, J. D., Thiele, I. & Palsson, B. Ø. What is flux balance analysis? *Nature Biotechnology* **28**, 245–248 (2010).

[12]Monk, J. M. *et al.* i ml1515, a knowledgebase that computes escherichia coli traits. *Nature Biotechnology* **35**, 904–908 (2017).

[13]Wang, H. *et al.* RAVEN 2.0: A versatile toolbox for metabolic network reconstruction and a case study on Streptomyces coelicolor. *PLOS Computational Biology* **14**, e1006541 (2018).

[14]Kim, H. U. *et al.* Integrative genome-scale metabolic analysis of Vibrio vulnificus for drug targeting and discovery. *Molecular Systems Biology* **7**, 460 (2011).

[15]Bordbar, A., Monk, J. M., King, Z. A. & Palsson, B. O. Constraint-based models predict metabolic and associated cellular functions. *Nature Reviews Genetics* **15**, 107–120 (2014).

[16]Segrè, D., Vitkup, D. & Church, G. M. Analysis of optimality in natural and perturbed metabolic networks. *Proceedings of the National Academy of Sciences* **99**, 15112–15117 (2002).

[17]Lin, D.-W., Zhang, L., Zhang, J. & Chandrasekaran, S. Inferring metabolic objectives and trade-offs in single cells during embryogenesis. *Cell Systems* **16** (2025).

[18]Hasibi, R., Michoel, T. & Oyarzún, D. A. Integration of graph neural networks and genome-scale metabolic models for predicting gene essentiality. *NPJ Syst Biol Appl* **10** (2024).

[19]Apaolaza, I. *et al.* An in-silico approach to predict and exploit synthetic lethality in cancer metabolism. *Nature communications* **8**, 459 (2017).

[20]Olaverri-Mendizabal, D., Valcárcel, L. V., Barrena, N., Rodríguez, C. J. & Planes, F. J. Review and meta-analysis of the genetic minimal cut set approach for gene essentiality prediction in cancer metabolism. *Briefings in Bioinformatics* **25**, bbae115 (2024).

[21]Hahn, M. W. & Kern, A. D. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Molecular biology and evolution* **22**, 803–806 (2005).

[22]Campos, T. L., Korhonen, P. K., Gasser, R. B. & Young, N. D. An Evaluation of Machine Learning Approaches for the Prediction of Essential Genes in Eukaryotes Using Protein Sequence-Derived Features. *Computational and Structural Biotechnology Journal* **17**, 785–796 (2019).

[23]Aromolaran, O., Aromolaran, D., Isewon, I. & Oyelade, J. Machine learning approach to gene essentiality prediction: a review. *Briefings in Bioinformatics* **22** (2021).

[24]Hasan, M. A. & Lonardi, S. Deeplyessential: a deep neural network for predicting essential genes in microbes. *BMC bioinformatics* **21**, 1–19 (2020).

[25]Zhang, X., Xiao, W. & Xiao, W. Deephe: Accurately predicting human essential genes based on deep learning. *PLOS Computational Biology* **16**, e1008229 (2020).

[26]Kang, B., Fan, R., Cui, C. & Cui, Q. Comprehensive prediction and analysis of human protein essentiality based on a pretrained large language model. *Nature Computational Science* **5**, 196–206 (2025).

[27]Palsson, B. Ø. *Systems biology: simulation of dynamic network states* (Cambridge University Press, 2011).

[28]Devoid, S. *et al.* Automated genome annotation and metabolic model reconstruction in the SEED and Model SEED. *Methods in Molecular Biology* **985**, 17–45 (2013).

[29]Kingma, D. P., Welling, M. *et al.* An introduction to variational autoencoders. *Foundations and Trends in Machine Learning* **12**, 307–392 (2019).

[30]Cain, S., Merzbacher, C. & Oyarzún, D. A. Low-dimensional representations of genome-scale metabolism. *Foundations of Systems Biology in Engineering Conference* 2024–05 (2024).

[31]Yeo, H. C., Hong, J., Lakshmanan, M. & Lee, D.-Y. Enzyme capacity-based genome scale modelling of cho cells. *Metabolic Engineering* **60**, 138–147 (2020).

[32]Zhang, C. *et al.* Yeast9: a consensus genome-scale metabolic model for S. cerevisiae curated by the community. *Molecular Systems Biology* **20**, 1134–1150 (2024).

[33]Sharma, K., Marucci, L. & Abdallah, Z. S. Fluxgat: Integrating flux sampling with graph neural networks for unbiased gene essentiality classification. *arXiv preprint arXiv:2403.18666* (2024).

[34]Han, T., Nazarbekov, A., Zou, X. & Lee, S. Y. Recent advances in systems metabolic engineering. *Current Opinion in Biotechnology* **84**, 103004 (2023).

[35]Cachera, P. *et al.* CRI-SPA: a high-throughput method for systematic genetic editing of yeast libraries. *Nucleic Acids Research* **51**, e91 (2023).

[36]Fang, L. *et al.* Genome-scale CRISPRi screen identifies pcnB repression conferring improved physiology for overproduction of free fatty acids in Escherichia coli. *Nature Communications* **16**, 3060 (2025). URL `https://www.nature.com/articles/s41467-025-58368-3`. Publisher: Nature Publishing Group.

[37]Djoumbou-Feunang, Y. *et al.* Biotransformer: a comprehensive computational tool for small molecule metabolism prediction and metabolite identification. *Journal of cheminformatics* **11**, 1–25 (2019).

[38]Schneider, P., von Kamp, A. & Klamt, S. An extended and generalized framework for the calculation of metabolic intervention strategies based on minimal cut sets. *PLoS computational biology* **16**, e1008110 (2020).

[39]Merzbacher, C., Mac Aodha, O. & Oyarzún, D. A. Modelling dynamic host-pathway interactions at the genome scale with machine learning. *Metabolic Engineering (in press)* (2025).

[40]Schuetz, R., Zamboni, N., Zampieri, M., Heinemann, M. & Sauer, U. Multidimensional optimality of microbial metabolism. *Science* **336**, 601–4 (2012).

[41]Shoval, O. *et al.* Evolutionary trade-offs, Pareto optimality, and the geometry of phenotype space. *Science* **336**, 1157–60 (2012).

[42]Zhao, Q., Stettner, A. I., Reznik, E., Paschalidis, I. C. & Segrè, D. Mapping the landscape of metabolic goals of a cell. *Genome Biology* **17**, 109 (2016).

[43]Richelle, A. *et al.* Model-based assessment of mammalian cell metabolic functionalities using omics data. *Cell Reports Methods* **1**, 100040 (2021).

[44]Hausser, J. & Alon, U. Tumour heterogeneity and the evolutionary trade-offs of cancer. *Nature Reviews Cancer* **20**, 247–257 (2020).

[45]Brunk, E. *et al.* Recon3D enables a three-dimensional view of gene variation in human metabolism. *Nature Biotechnology* **36**, 272–281 (2018).

[46]Heinken, A. *et al.* Genome-scale metabolic reconstruction of 7,302 human microorganisms for personalized medicine. *Nature Biotechnology* **41**, 1320–1331 (2023).

[47]Gustafsson, J. *et al.* Generation and analysis of context-specific genome-scale metabolic models derived from single-cell RNA-Seq data. *Proceedings of the National Academy of Sciences* **120**, e2217868120 (2023).

[48]Srivatsa, S. *et al.* Discovery of synthetic lethal interactions from large-scale pan-cancer perturbation screens. *Nature Communications* **13**, 7748 (2022).

[49]Ye, L. *et al.* A genome-scale gain-of-function CRISPR screen in CD8 T cells identifies proline metabolism as a means to enhance CAR-T therapy. *Cell Metabolism* **34**, 595–614.e14 (2022).

[50]Yang, J. H. *et al.* A White-Box Machine Learning Approach for Revealing Antibiotic Mechanisms of Action. *Cell* **177**, 1649–1661.e9 (2019).

[51]Faure, L., Mollet, B., Liebermeister, W. & Faulon, J.-L. A neural-mechanistic hybrid approach improving the predictive power of genome-scale metabolic models. *Nature Com-*

munications **14**, 4669 (2023).

[52]Gopalakrishnan, S. *et al.* COSMIC-dFBA: A novel multi-scale hybrid framework for bioprocess modeling. *Metabolic Engineering* **82**, 183–192 (2024).

[53]Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning* (Springer, 2009).

[54]Wainwright, M. J. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint.* Cambridge Series in Statistical and Probabilistic Mathematics (Cambridge University Press, Cambridge, 2019).

[55]Cui, H. *et al.* scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods* **21**, 1470–1480 (2024).

[56]Nguyen, E. *et al.* Sequence modeling and design from molecular to genome scale with Evo. *Science* **386**, eado9336 (2024).

[57]Megchelenbrink, W., Huynen, M. & Marchiori, E. optGpSampler: An Improved Tool for Uniformly Sampling the Solution-Space of Genome-Scale Metabolic Networks. *PLOS ONE* **9**, e86587 (2014).

[58]Xiong, K. *et al.* An optimized genome-wide, virus-free CRISPR screen for mammalian cells. *Cell Reports Methods* **1**, 100062 (2021).

[59]Ebrahim, A., Lerman, J. A., Palsson, B. O. & Hyduke, D. R. Cobrapy: constraints-based reconstruction and analysis for python. *BMC systems biology* **7**, 1–6 (2013).

METHODS

Flux Cone Learning utilizes sampling data generated with a random walk on a deletion-specific genome-scale metabolic model (GEM). First, a wild type GEM is modified with a gene deletion by setting the corresponding reaction bounds to zero. The high-dimensional flux cone of the deletion GEM is sampled using a random walk sampler; in our implementations we opted for OptGPSampler[57], a fast Monte Carlo method that aims to uniformly sample the flux cone. The sampler first transforms the problem into a convex optimization problem in logarithmic space using geometric programming, then employs a hit-and-run algorithm to sample the interior of the cone. The resulting flux data has a number of rows equal to the number of samples and a number of columns equal to the number of reactions in the GEM. Each deletion produces a collection of flux sampling vectors, all which are labeled with a fitness score obtained from experimental data. The fitness score can be either discrete or continuous depending on the fitness readout under study.

The resulting labeled dataset is then employed for training supervised machine learning models. The model predictions are made at the level of flux samples, i.e. one row of the flux sampling data frame from each deletion is passed through the trained machine learning model to produce a predicted fitness score. Therefore, every sample from each deletion GEM is assigned an individual predicted score, and the distribution of these scores is finally averaged to obtain a gene-level prediction. Flux Cone Learning can deliver high predictive accuracy because it is trained to learn correlations between the geometry of the flux cone and the resulting phenotype.

Generation of flux sampling data

Flux sampling is a collection of methods for randomly generating flux distributions from the solution space of a genome-scale metabolic model. Flux sampling algorithms are random walk algorithms optimized for the high-dimensional and non-isotropic geometry of the convex polytope defined by the GEM. OptGPSampler uses artificial centering hit-and-run to bias the random walk towards the elongated sections of the flux cone. After an initial random location in the flux space is selected and a warm-up phase, every kth point following is generated by the sampler until N points are generated. These two parameters (k, N) control

the number of flux samples generated by the algorithm. Flux sampling is computationally costly because it requires running a random walk on an high-dimensional flux space that needs to reach mixing time to achieve uniform coverage.

We ran OptGPSampler on all single-gene deletions in four *Escherichia coli* models, the Yeast9 model for *Saccharomyces cerevisiae*, and the iCHO2291 model for Chinese Hamster Ovary cells. For training supervised machine learning models, sampling data were normalized to zero mean and unit variance. There were a small number of deletions in each GEM where the sampling failed to converge; these were not included in training or testing. A summary of GEM sizes and sampling data can be found in Supplementary Tables S1–S2. For example, in the Yeast9 model we sampled 1,159 single-gene deletions with a step size of k =5,000 for a sampling density of N =124 samples/cone, leading to a total of 143,716 samples with D =4,130 fluxes each (total data size 4.43Gb).

For all models except *Escherichia coli*, we sampled with a high step size of k =5,000. To ensure robust performance evaluations in the *Escherichia coli* iML1515 model (Figure 2)A–D, we retrained models many times using different training sets. For computational efficiency and due to large data sizes, after computing an initial large set of samples, using a fine step size of $k = 100$, we subsampled the data 10 times to have the same number of samples per deletion (N =100 samples/cone). Three smaller *E. coli* models (iAF1260, iJO1366, iJR904) were employed for the comparison in Figure 2F. In these models, deletion GEMs were sampled with N =100 samples/cone and k =5,000. To equalize the amount of training features between models, only the deletions present in all models (D =864 reactions) were included in the training and test sets for the models in Figure 2F. The biomass reactions were removed to ensure the models were learning from the true reaction fluxes, not the biomass reaction used to compute FBA predictions (see Supplementary Table S3).

Experimental fitness labels

The gene essentiality labels for *E. coli*, *S. cerevisiae* and CHO cells were obtained from the literature[12,32,58]. The yeast essentiality labels included nonmetabolic genes and were labeled with both gene and ORF labels. Gene names were standardized to their systematic names from the Saccharomyces Genome Database, resulting in $N = 1121$ metabolic gene deletions labeled with essentiality data, sampled, and included in the final dataset for model

training. ORFs and gene names were linked using a tool from Yeastract+ `http://www.yeastract.com/formorftogene.php`.

For the results in Figure 4, betaxanthin autofluorescence readouts for $N = 811$ yeast deletions were taken from Cachera *et al*[35] and averaged across four cultures. While one genes (YBR011C) was also identified as essential in other studies, we included it in our analysis as we hypothesized this could be a conditionally essential deletion which can grow in alternative strain and media conditions. The average autofluorescence was normalized to the (0,1) range. We first framed the problem as a regression task, but this proved challenging with the limited number of knockouts at the high and low ends of the autofluorescence distribution. Recognizing that predicting high or low producers is a core task in several applications, we chose to train a three-class classifier by binning the data into three classes of high, medium, and low producers (Figure 4A). We set the thresholds qualitatively to label 67% of samples as medium producers (within ∼1 standard deviation from the mean). In all our case studies, labels were highly class imbalanced, as shown in Supplementary Table S4.

Training of supervised learning models

All models were trained using the scikit-learn package in Python.

Escherichia coli A random forest model classifier was trained on a 80% training set stratified to maintain the class imbalance. Model hyperparameters were fixed to `max_depth` None, `min_samples_split` 2. The random forest was retrained N =5 times with different test sets to confirm that performance was not significantly affected by the composition of the training set. The FBA baseline was obtained using the `single_gene_deletion` function in the CobraPy package[59] applied to all genes in the iML1515 model with default biomass objective function, aerobic conditions, and glucose as the carbon source. Genes were classified as essential using a cutoff of 0.4 1/h on the predicted growth rate. The naive baseline was compared by predicting all genes as non-essential (majority class). Once trained on the sample level, the prediction score of all samples from a single deletion was averaged; if this score was less than 0.5, the deletion was classified as essential. A representative model was used to create the prediction score distributions in Figure 2C. In Figure 2D we trained $N = 50$ random forest classifiers on one subsample with random held-out test sets. Feature importance scores of all reactions were extracted from these random forest objects.

Saccharomyces cerevisiae For essentiality prediction, a class-stratified 20% of deletions (192 non-essential; 31 essential) was held out as a consistent test set. The remaining 80% of deletions (772 non-essential; 126 essential) were split into 5-fold cross validation sets and a random forest model was trained on each fold. The `max_depth`, `n_estimators`, and `min_samples_split` hyperparameters were tuned using a grid search and the model with the highest average cross validation accuracy was selected and the confusion matrix and ROC curve computed for Figure 3A. The best `max_depth` value was 30, the best `n_estimators` value (the number of trees) was 300, and the best `min_samples_split` (the minimum number of data points to split a leaf on the random forest) value was 2. The minimum deletion-level accuracy was 87.5%, the maximum was 90.3%. The test set results were computed by running the held-out test set through all 5 fold models and averaging the deletion-level scores across all models. The FBA baseline was computed using the `single_gene_deletion` function in Cobrapy for all genes with glucose as the carbon source and the standard biomass reaction.

For prediction of betaxanthin synthesis (Figure 4), multiple models were trained on a class-stratified 80% training set split with a consistent held-out test set. The following model types were trained: HistGradientBoostingClassifier, Linear Support Vector Classifier, Logistic Regression Classifier, Random Forest Classifier. We implemented two class balancing techniques to improve the minority class performance: balancing, which weights the class labels to account for the class imbalance, and resampling, which subsamples the majority class to be the same size as the minority classes.

Chinese Hamster Ovary cells A HistGradientBoosting Classifier was trained on a 5-fold cross validation of the training set. 20% of the original training dataset was held out as a test set and not included in the cross validation. The large training set data size required training models across 4 CPU nodes of to load all training data into memory. The hyperparameters `learning_rate`, `max_iter`, and `max_depth` parameters were tuned via grid search and the model with the highest average cross validation accuracy was selected and the confusion matrix and ROC curve computed for Figure 3B. The `learning_rate` was varied between 0.01 and 0.2, the `max_iter` between 100 and 500, and the `max_depth` set to 5, 10, or None. The best model had a `learning_rate` of 0.05, a `max_iter` of 100, and a `max_depth` of None. The test set results were computed by running the held-out test set through all 5 fold models and averaging the deletion-level scores across all models. The confusion matrix was

computed for a class threshold value of 0.5 for each fold and counts were averaged across all 5 folds. The FBA results were computed using the `single_gene_deletion` function in Cobrapy for all deletions and the default carbon source and objective function in the iCHO2291 model.

481 **FIGURES**

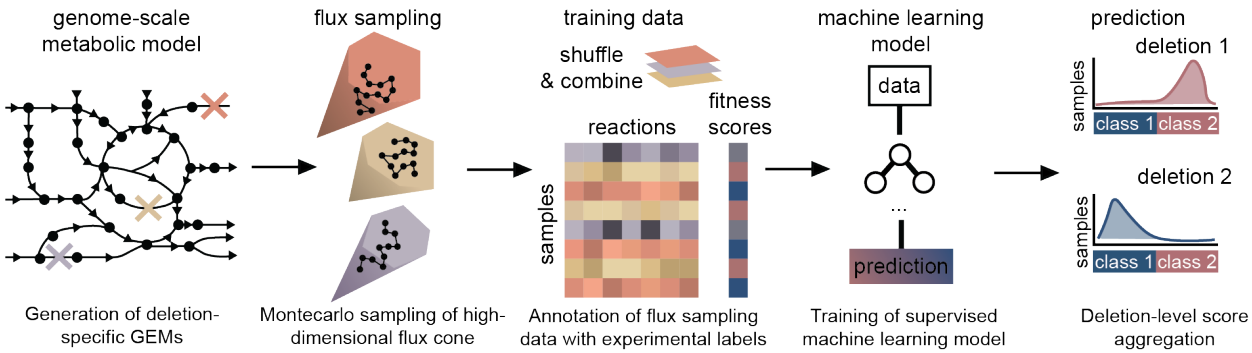

FIG. 1. **Flux Cone Learning of metabolic deletion phenotypes.** The Flux Cone Learning pipeline predicts experimental fitness scores from changes to the shape of the metabolic space. Starting from a genome-scale metabolic model (GEM) of the wild-type, we first create deletion-specific GEMs, which are then sampled via a random walk defined in a high-dimensional cone. Flux samples from each cone are then stacked, randomly shuffled, and then paired with experimental fitness scores for supervised training. Predicted sample-level fitness scores are finally averaged across each deletion to produce gene-level predictions. The choices of supervised machine learning model and task (regression/classification) are flexible and can accommodate various types of deletion screens depending on the nature of the fitness readouts.

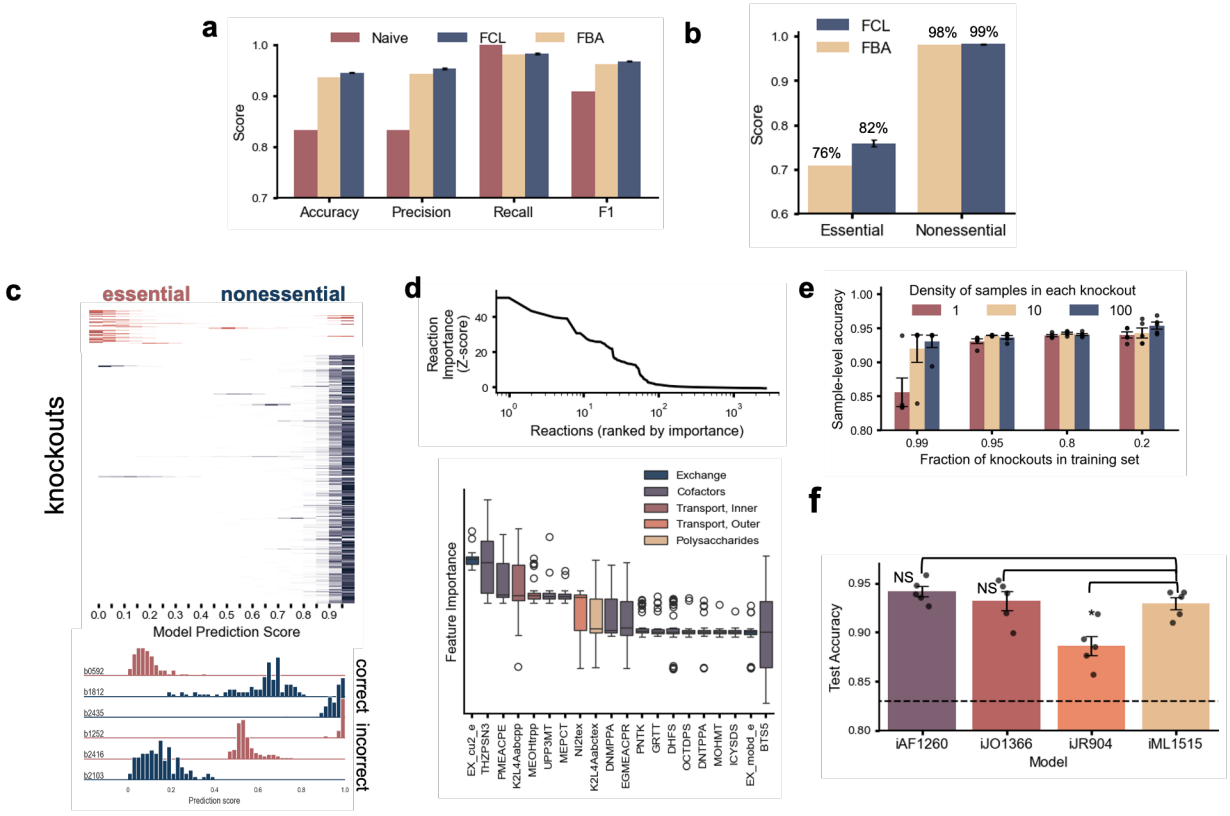

FIG. 2. **Prediction of metabolic gene essentiality in *Escherichia coli*.** **(A)** Flux Cone Learning (FCL) delivers best results for metabolic gene essentiality prediction in *E. coli*, outperforming the current gold standard predictions from Flux Balance Analysis (FBA). FCL predictions were computed on a class-stratified test set with 300 genes held-out from training; FBA predictions were computed for all genes in the iML1515 genome-scale model with default media with glucose as carbon source and aerobic growth[12]. The naive baseline was computed by predicting all genes as non-essential (majority class); error bars denote standard error computed across $N = 5$ training repeats using $N = 10$ training subsamples with $q = 100$ samples/cone; training sets were subsampled from large flux sampling data with more than 5,000 samples/cone. **(B)** Prediction accuracy for essential and non-essential genes. **(C)** Top: Distribution of sample-level FCL prediction scores across 300 test genes for one representative random forest model. Bottom: Representative prediction score distributions for correctly and incorrectly predicted genes. **(D)** Top: reaction feature importance across all genes employed for training, using the random forest feature importance scores. Bottom: Importance of top 20 features across 50 repeats; box plots show mean, IQR, and whiskers are all samples not determined to be outliers. **E)** Performance of FCL with smaller and less dense training data; error bars are standard error across 5 training repeats with different initializations. **(F)** Performance of FCL with earlier versions of the *E. coli* genome-scale metabolic model. Test results were computed across N =848 genes shared by the four models[10]. Significance was determined at $p < 0.05$ with a one-sided t-test; error bars are standard error across 5 training repeats.

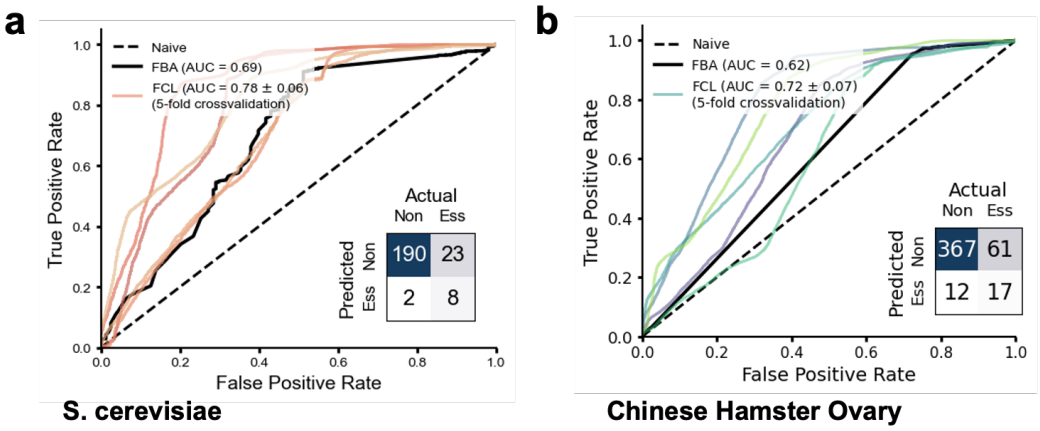

FIG. 3. **Prediction of metabolic gene essentiality for higher-order organisms.** **(A)** Receiver operating characteristic (ROC) curves of FCL model for *Saccharomyces cerevisiae* in 5-fold cross validation; models were trained on N =897 class-stratified gene deletions with q =124 samples/cone computed from the Yeast9 genome-scale model[32]. Solid black line are FBA baseline predictions computed for all genes in the GEM. **(B)** ROC curves of FCL model for Chinese Hamster Ovary cells in 5-fold cross validation; models were trained on N =1,832 class-stratified gene deletions with q =127 samples/cone computed from a well-adopted genome-scale model[31]. In both panels, insets show confusion matrices for FCL predictions computed on class-stratified test genes held out from training (20%) and averaged across all folds. AUROC metrics for FCL models were computed as an average across folds ± one standard deviation. Solid black line are FBA baseline predictions computed for all genes in the GEM. Details on model training can be found in the Methods; additional classification metrics can be found in Supplementary Table S5.

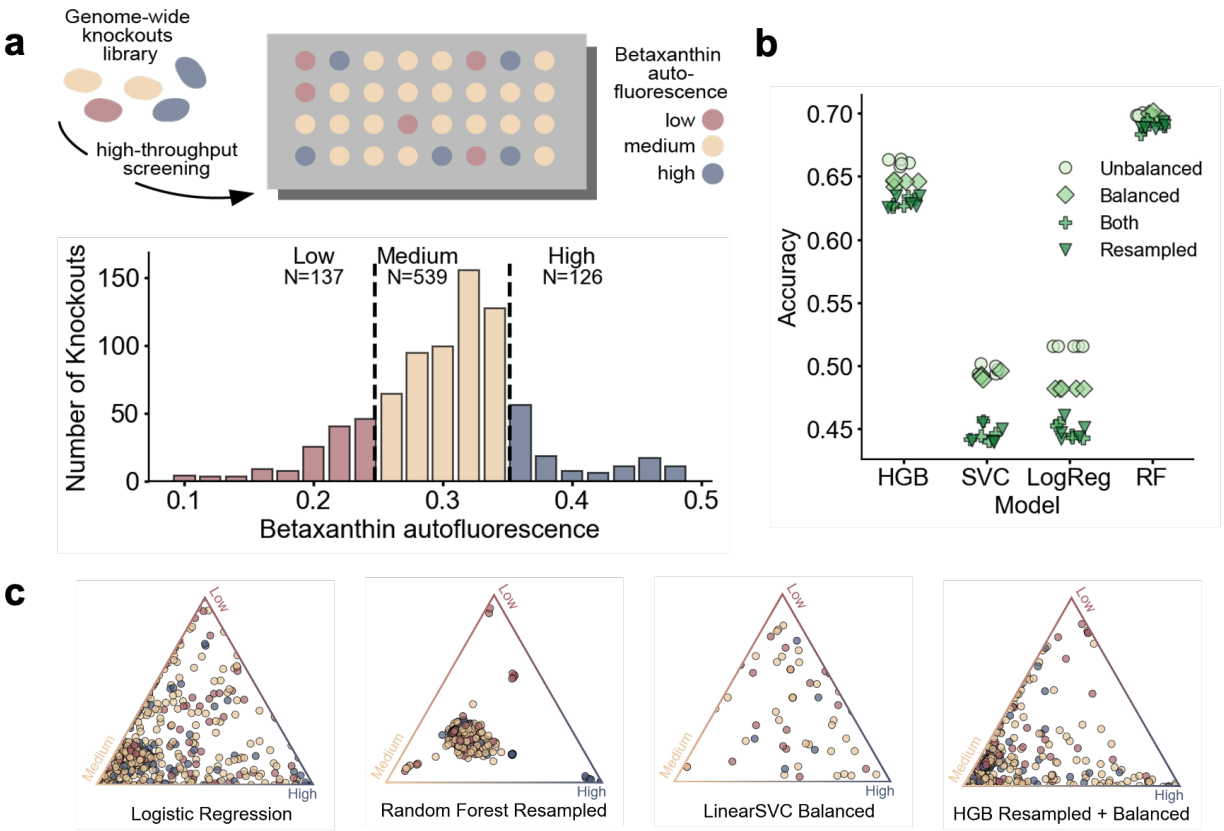

FIG. 4. **Prediction of small molecule synthesis in *Saccharomyces cerevisiae*. (A)** High-throughput deletion screening data of *S. cerevisiae* strains engineered to produce betaxanthin[35]; fitness scores were determined from betaxanthin autofluorescence, normalized and binned into three classes for model training. The majority of deletions are medium producers, with ∼15% of deletions being high producers. **(B)** Accuracy results for several FCL models with different algorithms for multiclass classification of deletion strains (Histogram-based Gradient Boosting, HGB; Support Vector Classifier, SVC; Logistic Regression, LogReg; random forest, RF) and various strategies to re-balance the three classes. Accuracy was computed across all three classes and shown for N =5 training repeats. Balanced: class labels were weighted to account for class imbalance; Resampled: majority class was subsampled to be the same size as the minority classes. Class re-balancing can increase accuracy for high producers, often at the expense of overall accuracy. For full class balancing results on a held-out 20% test set (N =659 deletions), see Supplementary Table S6. **(C)** Ternary plots of model predictions on the test set for representative models with varying predictive accuracy across the three classes. Vertices represents class prediction with probability one (full confidence), whereas central points are deletions predicted to be equally likely to be any of the three classes. Each sample has been colour coded according to their ground truth class labels.

SUPPLEMENTARY FIGURES AND TABLES

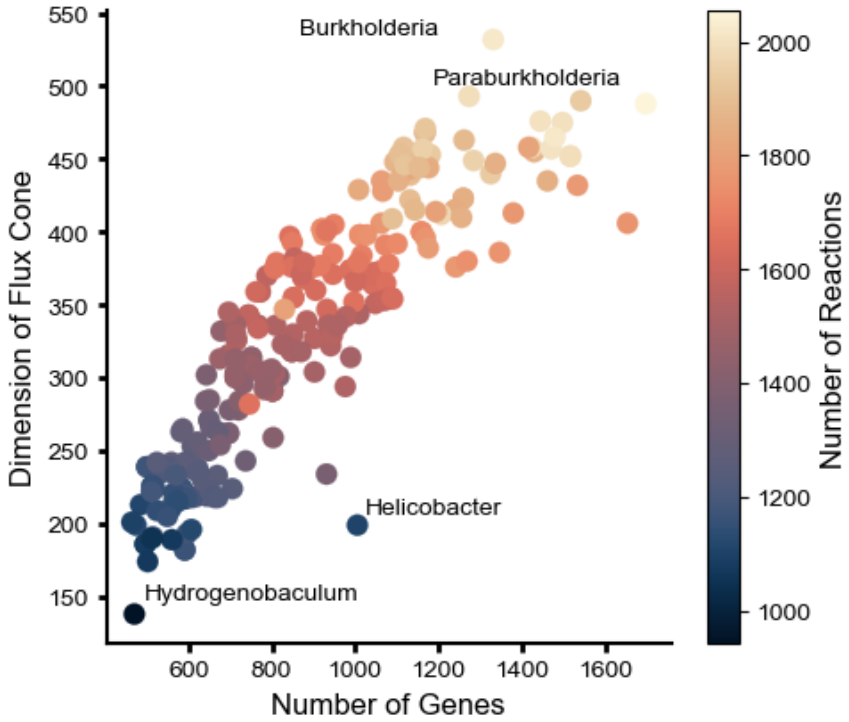

FIG. S1. **Genome-scale metabolic models from across the bacterial kingdom.** Models for $N = 244$ bacterial species were obtained from Plata et al, Nature, 2015 and species names were obtained from Ramon and Stelling, Nature Communications, 2023. The stoichometric matrices of each model were extracted, along with the number of genes (x-axis) and reactions (hue) in each model. The nullity of the stoichiometric matrix (the number of columns in the null space) was computed for each species (y-axis).

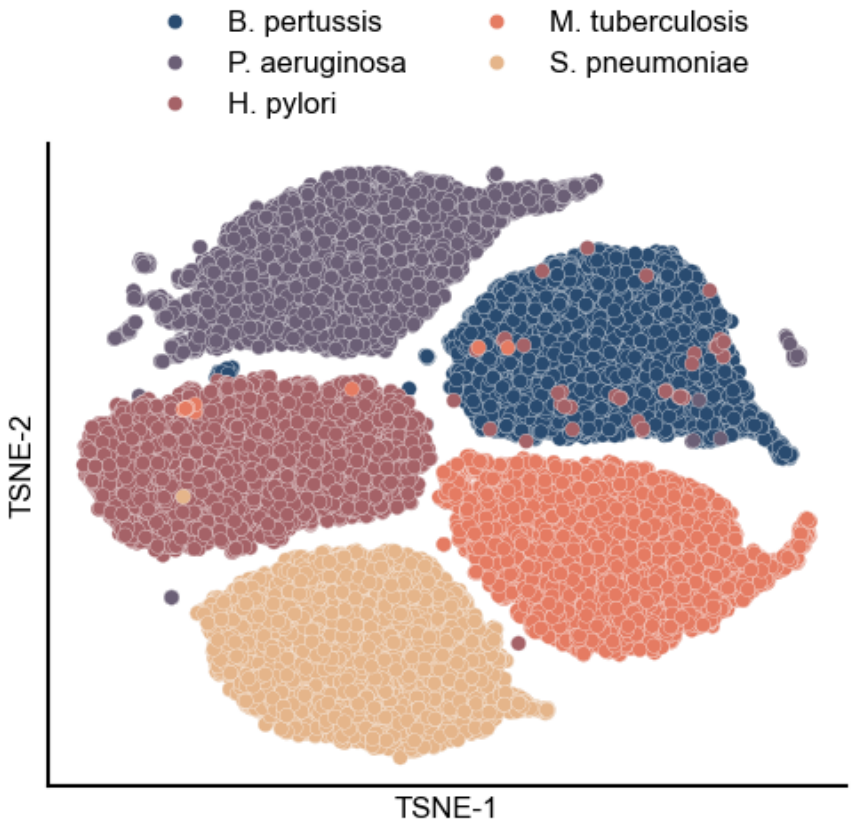

FIG. S2. **Variational autoencoder compression of five bacterial pathogens.** OptGPSampler was used to obtain $N = 5000$ samples ($k = 100$) from five bacterial GEMs selected from Plata et al, Nature, 2015. A variational autoencoder (VAE) was trained on the common $N = 494$ reactions shared across all species. The final embedding dimension D =8 was reached via 5 linearly decreasing ($D = 494$, $D = 397$, $D = 300$, $D = 203$, $D = 107$, $D = 8$) fully connected layers and the VAE was trained with PyTorch for 100 epochs with default hyperparameters using the Adam optimizer; t-SNE embeddings were employed for 2-dimensional visualization. The embeddings captures the separation of metabolic spaces across species. While the learned representation displays a small number of outliers, overall we found that all 5 pathogens retain distinct clustering despite the VAE only being trained on non species-specific reactions.

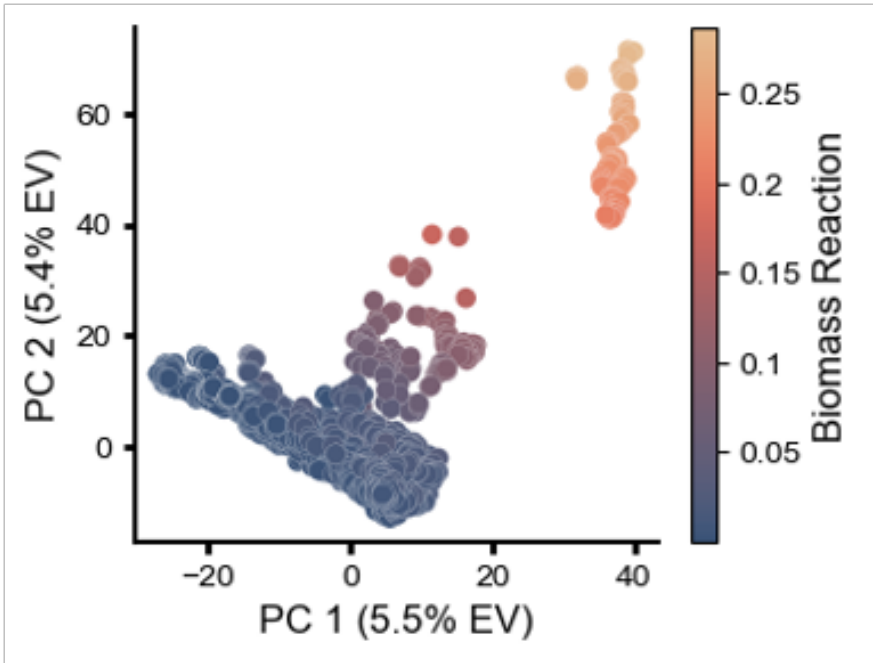

FIG. S3. **PCA representation of flux cone of wild type *E. coli* metabolism.** The unmodified (WT) iML1515 model was sampled 5000 times with a step size k=100.The resulting data was normalized using zero mean and unit variance and Principal Component Analysis (PCA) was performed. 621 PCs were required to explain 95% of variance in the scaled data. The first two PCs accounted for 5.5% and 5.4% of variance, respectively. The biomass reactions were removed from the PC calculations (see Supplementary Table S3) but samples on the plot are colored based on the value of the forward biomass reaction (BIOMASS_Ec_iML1515_core_75p37M).

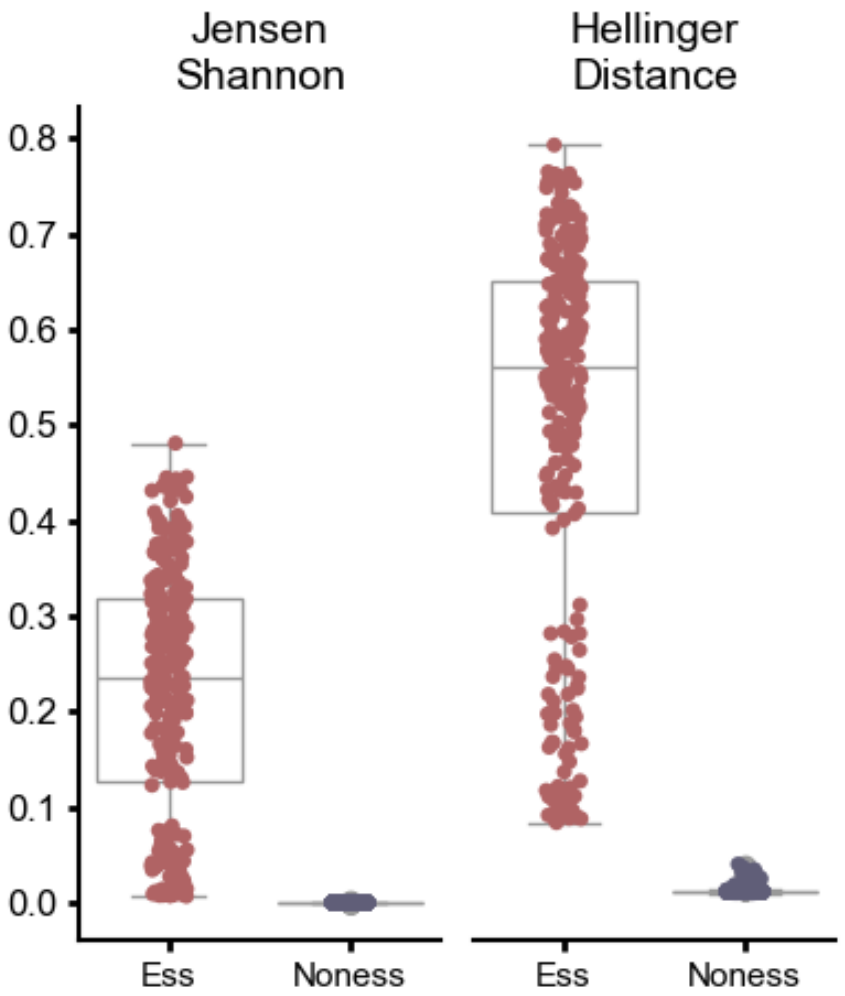

FIG. S4. **Complete classifier sample scoring.** We sought to exploit the excellent predictive performance of FCL to define a distance metric between deletions and the wild type strain. To this end, we retrained FCL on all *E. coli* gene deletions (N =1,502 genes), and computed the distribution of prediction scores for all flux samples in the wild type and each deletion strain. We then queried the model with flux samples of the wild-type GEM and each deletion cone (100 samples/cone), to produce distributions of prediction scores for the WT and each deletion strain. We scored each strain with the Jensen-Shannon divergence and Hellinger distance between the score distributions of each deletion and the wild type. We found statistically significant differences in score distribution of non-essential and essential genes, which reinforces the conclusion that perturbations to the flux cone are indicative of gene essentiality (one-sided t-test with p 0.05).

GEM	Number of genes	Number of reactions	Stoichiometric nullity
iML1515	1516	2712	867
iJO1366	1367	2583	817
iAF1260	1261	2382	752
ijr904	904	1075	332
Yeast9	1162	4130	1539
iCHO2291	2291	6236	2390

TABLE S1. Summary of genome-scale metabolic models employed in this study.

GEM	Sampled KOs	Total S	Density per cone	Step size	Subsampling	Size of NPZ file
iML1515	1502	150313	101	100	50	3.06Gb
iJO1366	1318	163680	124	5000	1	3.17Gb
iAF1260	1214	150784	124	5000	1	2.69Gb
ijr904	866	107508	124	5000	1	0.89Gb
Yeast9	1121	143716	124	1	1	4.43Gb
iCHO2291	2290	291028	127	1	1	13.53Gb

TABLE S2. Summary of model sampling details. All models were sampled with the same step size, except iML1515 which employed a smaller step size to generate more samples and the subsampled for the analysis in Figure 2.

GEM	Biomass Reactions Removed [IDs]
iML1515	BIOMASS_Ec_iML1515_core_75p37M [2669], BIOMASS_Ec_iML1515_core_75p37M_reverse_35685 [2670]
iJO1366	BIOMASS_Ec_iJO1366_core_53p95M [19], BIOMASS_Ec_iJO1366_core_53p95M_reverse_5c8b1 [14]
iAF1260	BIOMASS_Ec_iAF1260_core_59p81M [926]
ijr904	BIOMASS_Ecoli [269], BIOMASS_Ecoli_reverse_bf7a1 [270]
Total Common Reactions	864

TABLE S3. Biomass reactions removed for *E. coli* models and their common reactions.

Task	Class imbalance	Class breakdown
E. coli essentiality	83/17	N=1252 non-essential, N=251 essential
Yeast essentiality	86/14	N=964 non-essential, N=157 essential
CHO essentiality	83/17	N=1898 non-essential, N=392 essential
Yeast production	17/67/16	N=138 low, N=545 medium, N=128 high

TABLE S4. Class imbalances for each task.

Organism	Model	Accuracy	Precision	Recall	F1 Score	PR AUC	ROC AUC
Yeast	FBA	0.81	0.91	0.97	0.89	0.92	0.69
Yeast	FCL (5-fold CV)	0.90±0.008	0.91±0.008	0.98±0.010	0.94±0.007	0.94±0.021	0.78±0.065
Yeast	FCL (test)	0.89	0.89	0.99	0.94	0.93	0.73
CHO	FBA	0.85	0.86	0.97	0.91	0.86	0.62
CHO	FCL (5-fold CV)	0.85±0.014	0.86±0.010	0.97±0.008	0.91±0.008	0.91±0.029	0.72±0.068
CHO	FCL (test)	0.84	0.86	0.97	0.91	0.86	0.65

TABLE S5. Classification performance metrics for Flux Cone Learning (FCL) and Flux Balance Analysis (FBA) in yeast (*S. cerevisiae*) and Chinese Hamster Ovary (CHO) cells (Figure 3).

Model Name	Baseline	Resampled	Balanced	Both	Max % improvement
HistGradientBoostingClassifier	11.4	13.6	14.1	14.6	28.3
LinearSVC	23.8	24.0	27.2	26.8	27.2
LogisticRegression	23.3	26.2	29.5	28.9	14.2
RandomForestClassifier	18.1	11.7	19.1	18.4	5.5

TABLE S6. Accuracy for high producer deletion yeast deletion strains for each class balancing methods in Figure 4. Models were assessed across all genes in a held-out, class-stratified 20% test set (N =649 deletions).