Master's Thesis

Report of Reading the Paper [Yan+19]

Chunhao Gu

March 10, 2021

1 Introduction

Antibiotics, a cornerstone of modern medicine, are threatened by the increasing burden of drug resistance, which is compounded by a diminished antimicrobial discovery pipeline [BW16]. In recent years, there is growing appreciation for secondary processes, such as altered metabolism, which actively participate in antibiotic efficacy. A deeper understanding of how bacterial metabolism interfaces with antibiotic lethality has the potential to open new drug discovery paradigms [MMP14]. In the project of the paper, the researchers invent a new white-box machine learning approach to discover a novel antibiotic mechanism of action.

2 Methods and Results

2.1 Metabolite Screening Experiments

There were two parts of biological experiments in the project. One was metabolite screening experiments conducted before the data analysis to obtain the data for analysis. The other was validation experiments conducted after the data analysis to verify the discovery from the data analysis. This subsection will briefly introduce the first part and the other will be introduced in Subsection 2.3.

The researchers did metabolite screening experiments with *Escherichia coli* strain K-12 MG1655 (ATCC 700926). The *E. coli* cells were cultured in MOPS minimal medium with ^{13}C glucose. After overnight culture, the cells were grown to early exponential phase, and back-diluted to $OD_{600} = 0.1$. Then they were dispensed into Biolog phenotype microarray (PM) plates 1–4 [Boc08] with different concentrations of ampicillin (AMP), ciprofloxacin (CIP), or gentamicin (GENT) added and systematically screened with an unbiased and semi-comprehensive library of metabolites, which contains 206 unique amino acids, carbohydrates, nucleotides, and organic acids that are included in the iJO1366 genome-scale model of *E. coli* metabolism. After 4 hours' incubation, OD_{600} was measured on a SpectraMax M5 Microplate Reader. Then IC_{50} were estimated from each set of $n > 3$ independent biological replicates by fitting logistic functions to each set of OD_{600} measurements for each metabolite well by MATLAB (Figure 1). After the experiments, the researchers obtained the data of Table S2.

2.2 Data Analysis

2.2.1 Hierarchical Clustering (Unsupervised Learning)

Hierarchical clustering of the measured IC_{50} values revealed that the metabolite response profiles differed between AMP, CIP, and GENT, highlighting their different biochemical targets (Figure 2). However, several metabolites appeared to commonly potentiate or inhibit efficacy across multiple antibiotics, indicating shared metabolic mechanisms of action. Many carbon

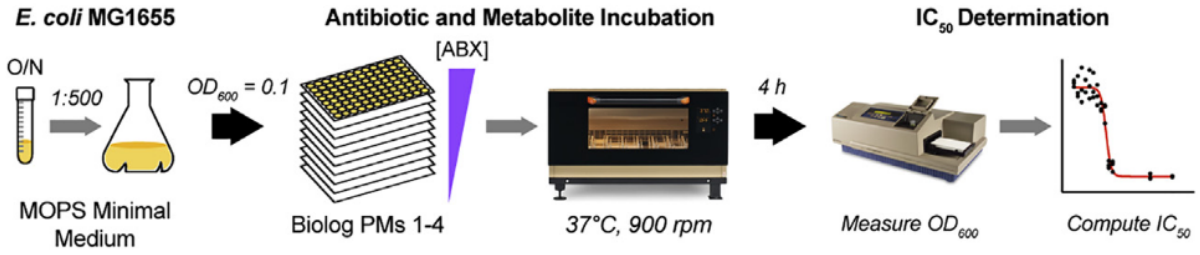

Figure 1: Overall Flow of Metabolite Screening Experiments

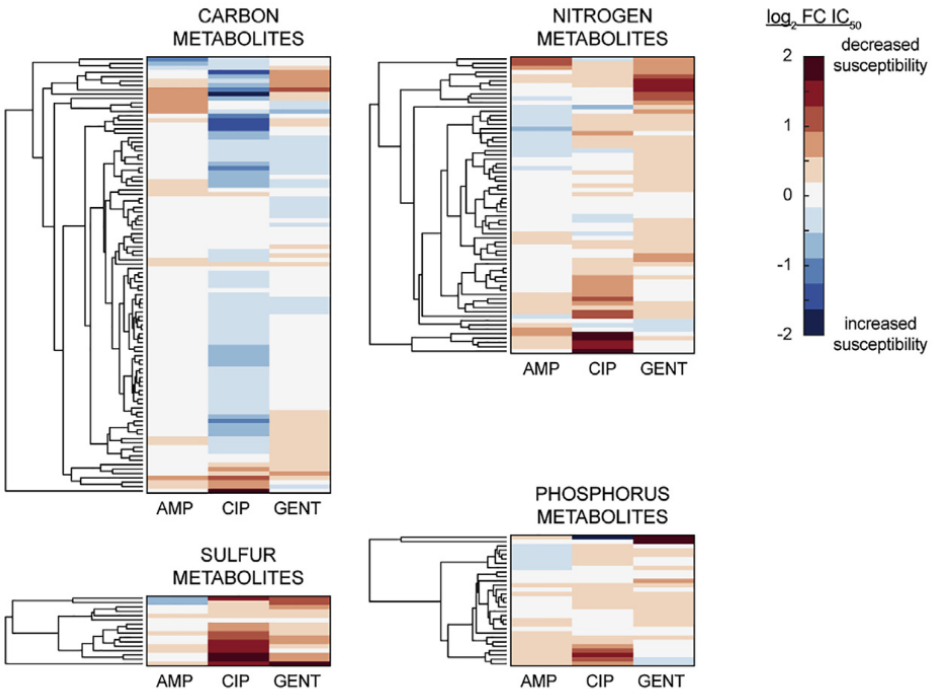

Figure 2: Hierarchical Clustering of IC_{50} Values

metabolic pathways were susceptible to antibiotic, whereas nitrogen, phosphorus, and sulfur were less susceptible. These raw data indicate that the measured antibiotic lethality responses to metabolite perturbations occurred through specific metabolic pathways rather than generically as a response to medium enrichment.

2.2.2 Metabolite Set Enrichment Analysis

For each antibiotic, a metabolite set enrichment analysis was performed in Ecocyc. A SmartTables was created comprised of metabolites eliciting a $\geq$ 2-fold change in IC_{50} for at least one antibiotic (Table S3). Pathways were identified using the 'Enrichment' analysis type. The Fisher Exact test was performed for each enrichment analysis with false discovery rate (FDR) correction by the Benjamini-Hochberg method (Table S4). These findings are consistent with previous observations that protein translation inhibitors generally exert antagonistic effects on antibiotic lethality and thus do not provide novel mechanistic insights.

2.2.3 White-Box Machine Learning

Conventional hierarchical clustering and enrichment analysis revealed few novel mechanistic insights. The researchers combined machine learning and metabolic networks to invent a new

white-box machine learning approach to analyze the data.

Genome-Scale Metabolic Network Simulation The first step of the white-box machine learning was to obtain training data. The researchers had obtained the antibiotic lethality (IC_{50}) for each metabolite perturbation through the metabolite screening experiments. Next, they decided to obtain the metabolic flux for each metabolite perturbation through the simulations in iJO1366 genome-scale metabolic network of *E. coli*.

In order to simulate the cell growth in MOPS minimal medium, reaction bounds from the exchange reactions corresponding to each metabolite present in MOPS minimal medium were set to a value of '1,000', to permit uptake. Reaction bounds for oxygen exchange, glucose exchange and cobalamin exchange were as set to values of '18.50', '10' and '0.1', respectively, as previously described in [Ort+11]. For each metabolite screening condition, additional exchange reactions were added to represent supplementation with each metabolite on the Biolog phenotype microarray plates (Table S1), with reaction bounds set to '1,000' to permit uptake of each metabolite from the extracellular environment. The simulations were performed using the COBRA Toolbox in MATLAB and Gurobi Optimizer. COBRA also has the Python version COBRApy.

Next, parsimonious flux balance analysis (pFBA) [Lew+10] optimized for the biomass objective function was performed on each metabolite condition-specific model 10,000 times with sampling by optGpSampler. For each reaction in the condition-specific models, the mean flux across all 10,000 samples was computed and used to represent flux in each condition (Table S5).

Multi-Task Least-Squares Regression with Elastic Net Regularization (Supervised Learning) The elastic net is a regularized regression method that linearly combines the L1 and L2 penalties of the lasso and ridge methods. The method is particularly useful when the number of predictor variables (independent variables) is much bigger than the number of observations [ZH05]. In this project, there were 1098 iJO1366 reactions (predictor variables) and 208 metabolite condition-specific models (observations). Thus, the elastic net was a good choice.

Given m learning tasks $\{T_i\}_{i=1}^{m}$ where all the tasks or a subset of them are related but not identical, multi-task learning aims to help improve the learning of a model for T_i by using the knowledge contained in the m tasks [ZY17]. In this project, a task of least-squares regression with elastic net regulation were conducted for each of the three antibiotics. The fluxes of iJO1366 reactions were used as predictor variables (independent variables). The normalized and $\log_2$-transformed IC_{50} values of each antibiotic was used as the response variables (dependent variables) in the regression.

The scikit-learn MultitaskElasticNetCV has integrated multi-task regression and elastic net regularization. The researchers used it with 50-fold cross-validation, 1e4 max iterations and tolerance of 1e-6. After the regression coefficients were computed, the standard deviation of the coefficients was also computed for each antibiotic. Reactions (predictor variables) whose coefficients possessed magnitude less than half the standard deviation were filtered and removed. For AMP, CIP, and GENT, this yielded 189, 208, and 204 reactions, respectively (Table S6).

So far, the machine learning had built the predictive model to correlate the iJO1366 metabolic reactions (fluxes) and antibiotic lethality (IC_{50}). Because metabolic networks are mechanistically constructed, the predictor variables (reactions) comprising the predictive model are, in principle, mechanistically causal and represent tangible biochemical species that can be directly tested experimentally. This is why the approach is called white-box machine learning. General black-box machine learning can only be used to predict what will happen but this white-box machine learning can also reveal why it will happen.

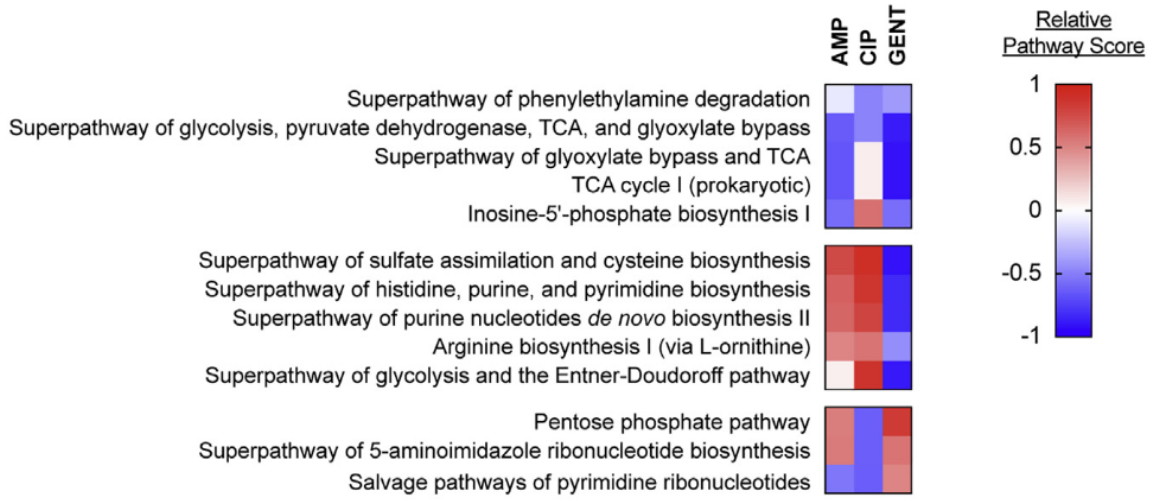

Figure 3: Pathway Scores after Hypergeometric Pathway Identification

Hypergeometric Pathway Identification This step was similar to metabolite set enrichment analysis. A hypergeometric statistical testing on metabolic pathways was performed in Ecocyc. For each antibiotic, the reactions selected in Table S6 were converted to their Ecocyc counterparts and hypergeometric p values were computed for each pathway-reaction set. For each antibiotic pathway combination, FDR statistics were estimated using the Benjamini-Hochberg method. Pathways that exhibited $p \leq 0.05$ and $q \leq 0.05$ for at least one antibiotic were selected. Of the 431 metabolic pathways curated by Ecocyc, only 13 were found to be statistically significant, with less than a 5% FDR for at least one antibiotic (Table S7).

Pathway and Reaction Score Computation The regression coefficients for each antibiotic had been computed and put in Table S5 through the previous steps. *Comment:* I find there is only one column of regression coefficients in Table S5. I think there should be three columns for the three antibiotics. So I'm not sure what this unique column of regression coefficients represents.

Pathway scores were computed by first computing the average of the non-zero regression coefficients for all reactions in each pathway. The magnitudes for these pathway scores were then $\log_{10}$-transformed and normalized by the largest magnitude of all pathway scores.

The pathways primarily clustered into three groups based on their pathway scores (Figure 3). The first cluster possessed central carbon metabolism pathways (superpathway of glycolysis, pyruvate dehydrogenase, TCA, and glyoxylate bypass; superpathway of glyoxylate bypass and TCA; and TCA cycle I (prokaryotic)) with similar pathway directionality for AMP, CIP, and GENT. These findings are consistent with several studies demonstrating the TCA cycle to be a shared mechanism in antibiotic lethality and validate the fidelity of the white-box machine learning approach.

Interestingly, the second cluster possessed purine biosynthesis pathways (superpathway of histidine, purine, and pyrimidine biosynthesis; and superpathway of purine nucleotides *de novo* biosynthesis II) with shared directionality between AMP and CIP and opposite directionality for GENT. To biological knowledge, purine biosynthesis has not been implicated previously as a mechanism of antibiotic lethality from any biochemical or chemogenomic screening experiments.

To better understand these differences in pathway directionality, the researchers further computed reaction scores for each reaction in the superpathway of purine nucleotides *de novo* biosynthesis II. Reaction scores were computed by taking the $\log_{10}$-transformation of each re-

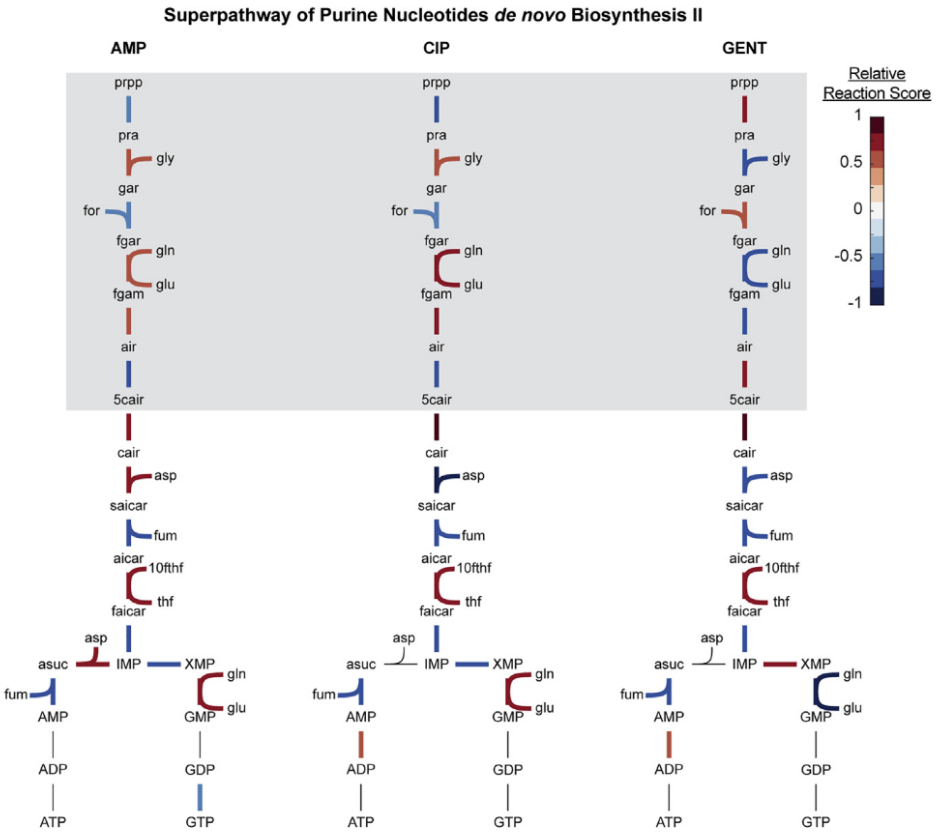

Figure 4: Reaction Scores for Purine Biosynthesis Pathways

gression coefficient for each antibiotic. These analyses identified the early steps in the purine biosynthesis pathway as being primarily responsible for the predicted differences for AMP and CIP from GENT (Figure 4). These findings illustrate how white-box machine learning can reveal new mechanisms of action with high biochemical specificity.

2.3 Validation Experiments

The above data analysis shows purine biosynthesis pathway may participate in antibiotic lethality. For AMP and CIP, stimulating the pathway may enhance antibiotic efficacy and inhibiting it weaken antibiotic efficacy. But for GENT, the effect may be different or even opposite. *Comment:* The next experiments focused on the feature of AMP and CIP and didn't delve why GENT showed the opposite feature.

2.3.1 Purine Biosynthesis Indeed Participate in Antibiotic Lethality

The researches did experiments to validate whether perturbations to purine biosynthesis would alter antibiotic lethality. First, they hypothesized that genetic deletion of enzymes involved in purine metabolism would exert differential effects on AMP and CIP lethality compared with GENT lethality. They constructed *E. coli* $\Delta glyA$, $\Delta purD$, $\Delta purE$, $\Delta purK$, $\Delta purM$, $\Delta pyrC$ and $\Delta pyrE$ gene knockout mutants by P1 phage transduction using the Keio collection. These mutants were supposed to inhibit their purine biosynthesis pathway. The experiments showed *E. coli* $\Delta purD$, $\Delta purE$, $\Delta purK$, and $\Delta purM$ mutants, in the early steps in purine biosynthesis, indeed exhibited significant decreases in AMP and CIP lethality but increased GENT lethality compared with the wild type. And biochemical inhibition of purine biosynthesis showed the similar results. They further hypothesized that stimulation of purine biosynthesis would elicit

opposite effects on antibiotic lethality than inhibition by these genetic and biochemical perturbations. Indeed, biochemical supplementation with the purine biosynthesis substrates phosphoribosyl pyrophosphate (prpp) and glutamine (gln) led to increased AMP and CIP lethality and decreased GENT lethality. Of note, these effects appear to be specific to purine metabolism because genetic deletion of enzymes involved in pyrimidine biosynthesis did not elicit significant differences in AMP, CIP, or GENT lethality.

2.3.2 Adenine Limitation and Pyrimidine Supplementation Increases Antibiotic Lethality

The researches further hypothesized that purine supplementation would rescue antibiotic-induced purine depletion and, consequently, decrease the demand for purine biosynthesis, reducing antibiotic lethality. The experimental results showed supplementation with adenine, but not guanine, decreased antibiotic lethality in wild-type cells. These results suggest that adenine limitation, but not guanine limitation, stimulate purine biosynthesis pathway and enhance antibiotic lethality. They also hypothesized that pyrimidine supplementation would inhibit pyrimidine biosynthesis and promote purine biosynthesis activity via prpp accumulation and, consequently, increase antibiotic lethality. Indeed, supplementation with uracil or cytosine potentiated antibiotic lethality.

2.3.3 Adenine Limitation Increases Antibiotic Lethality by Increasing Central Carbon Metabolism Activity and Cellular Respiration

The researchers reused the metabolic network model to simulate the states corresponding to adenine or uracil supplementation. The model predicted that adenine supplementation would decrease purine biosynthesis, central carbon metabolism activity and cellular oxygen consumption. They tested these model predictions by quantifying the intracellular concentrations of central carbon metabolism and energy currency metabolites from *E. coli* cells grown in MOPS minimal medium and supplemented with either adenine or uracil. The results support the model predictions that adenine supplementation decreases central carbon metabolism activity. They also tested these model predictions using a Seahorse XF analyzer and measured changes in the oxygen consumption rate (OCR) following antibiotic treatment with or without adenine or uracil supplementation. Antibiotic treatment with AMP, CIP, or GENT increased the cellular oxygen consumption but adenine supplementation significantly repressed changes in cellular oxygen consumption under antibiotic treatment, whereas uracil enhanced cellular oxygen consumption. These results support the previous observations that cellular respiration is important for antibiotic lethality [Lob+15] and further indicate that adenine limitation can fuel the process and increase antibiotic lethality (Figure 5).

3 Summary

Recent years, machine learning has shown its power in clinical medicine, such as image-based cancer diagnosis. However, its applications in fundamental biomedical studies are still limited. A reason is that a machine learning model is a predictive model that automatically builds the input-output association based on a large number of data and provides predictions for future data. It is like a black box and doesn't interpret the the input-output causality. Particularly, when the input is biomedical data, a machine learning model learns input-output associations according to its algorithms rather than simulates input-output causality according to the biological mechanisms. On the other hand, biological networks, such as metabolic networks, are mechanistic models that are built based on our biological knowledge and hypothesis. They can simulate the input–output causality according to the known biological mechanisms. Thus, the integration of machine learning approaches and biological networks may enhance the strength of each other [Bak+18].

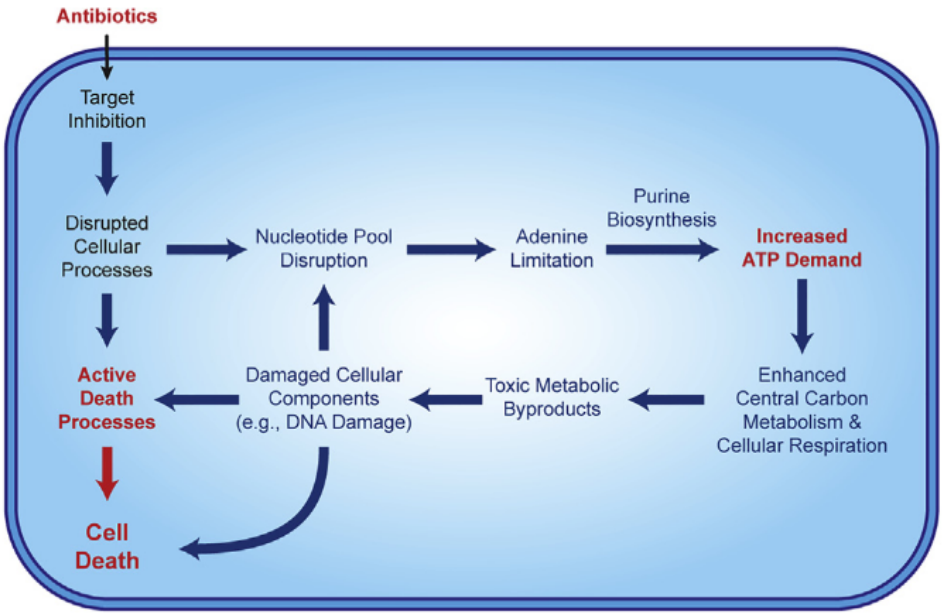

Figure 5: Antibiotic-Induced Adenine Limitation Induces Purine Biosynthesis, Increasing ATP Demand and Driving Cellular Respiration

Before this paper, there had been some good review articles on how to integrate the two approaches, such as [Cam+18]. One idea is to use machine learning and multi-omics data to build more complex and biologically realistic mechanistic models [Ma+18]. The idea in this paper is so innovative that the researchers used machine learning to correlate experimental data with mechanistic model simulation data and then fortunately discovered a novel antibiotic mechanism of action.

Though the new approach is attractive, but the following shortcomings may undermine the effectiveness of the approach and the reliability of the biological discovery:

1. The paper assumed that the metabolic network simulation states were only impacted by metabolite perturbations and the antibiotic effects were not counted during the simulations. Thus, the simulations were biased to the real metabolic reactions occurred in the screening experiments.

2. Hierarchical clustering of the measured IC_{50} values revealed that the metabolite response profiles differed between AMP, CIP, and GENT (Figure 2). And when analyzing pathway scores, the pathways were manually separated into three clusters. We see only the second cluster of AMP and CIP were similar and other pathway scores were pretty different between AMP, CIP, and GENT (Figure 3). The researchers dismissed these differences but favored those similarities from which they draw the paper's conclusion. It seems that they had assumed the conclusion according to some preceding observations and tried to find evidence via this computational approach to support it. If true, it is dubious if this new approach is effective in discovering new biological mechanisms without plausible assumptions. And I have checked some papers that cite this paper recently. I haven't found a paper which reuses the same white-box learning approach to reveal a novel biological mechanism.

In addition, the paper tried to weigh the lengths of biological experiments and computational approaches sections. But this made both parts were not detailed enough. So it may be inevitable to contact the authors if we want to identically reproduce the experiments in the paper.

4 Further Work

4.1 Biological Directions

1) The paper doesn't answer why there is an opposite in antibiotic lethality between AMP/CIP and GENT. We could reproduce the validation experiments of the paper but replace the antibiotics with other types. This could help us check the generality of the paper's conclusion.

2) The paper has validated *in vitro* the novel antibiotic mechanism of action. But a *in vitro* conclusion is not always applied to *in vivo* experiments. We could conduct *in vivo* studies to further validate the conclusion.

3) The paper used *E. coli* as the study subject. But a conclusion drawn from one species is not always applied to other species. We could conduct studies on other bacterial species to further validate the generality of the paper's conclusion.

Comment: If we choose these directions as the topic of the master's theis, it is inevitable to collaborate with the biological department. It may be difficult to coordinate the schedule in a short term.

4.2 Computational Directions

1) The paper covers a variety of computational methods. We could find other similar methods to replace the paper's ones and test them on the paper's data to see if we can have more discoveries. For example, the paper chose elastic net regularization because the method is particularly useful when the number of predictor variables is much bigger than the number of observations. We could delve if there is other machine learning models that can perform better for this case.

   Comment: This direction will almost reproduce the white-box machine learning in the paper yet with other computational methods. The work may help to improve the performance of the original white-box machine learning.

2) The paper used a lot of different tools which makes the flow of analysis look chaotic. We could study if it is possible to simplify the steps or conduct most steps in an integrative platform, such as MetaboAnalyst.

   Comment: This direction is a little boring but the work can help other researchers more conveniently use the white-box machine learning approach.

3) The paper has showed a paradigm to combine machine learning and metabolic networks. We could complete a survey on other paradigms. For example, as the paper has suggested, we could perform simulations on signaling networks from LINCS, BioGRID, Biobank, etc. to learn signaling mechanisms of epigenetic regulation, etc. Similarly, we could perform simulations on gene regulatory networks of [Wan+14] to learn transcriptional programs underlying screened phenotypes. Also, we could combine machine learning and protein–protein interaction networks, such as [Kon+20]. Furthermore, we could try to design a new paradigm.

   Comment: This direction is a greater topic. But if the master's thesis is just a survey without innovative points, it seems available to complete in a short term as long as we could find many paradigms.

4) The paper has applied the white-box machine learning to study cell metabolism in antibiotic efficacy. We could delve its applications to other biomedical areas related to

metabolism. For example, the paper has suggested cell metabolism in cancer pathogenesis, histidine metabolism in efficacy of cancer therapeutics.

Comment: The work will help spread the white-box machine learning to wider biomedical areas.

5) The paper has showed a computational approach in antibiotic studies. We could look for other computational approaches for this biological topic, such as [Sto+20], and complete a survey to compare them. Furthermore, we could try to design a new approach.

   Comment: The work will help summarize and extend computational approaches for antibiotic studies.

In addition to the above, we could discuss more master's thesis directions based on this paper.

References

[Bak+18] Ruth E. Baker et al. "Mechanistic models versus machine learning, a fight worth fighting for the biological community?" In: *Biology Letters* 14.5 (2018), p. 20170660. ISSN: 1744-9561. DOI: 10.1098/rsbl.2017.0660.

[Boc08] Barry R. Bochner. "Global phenotypic characterization of bacteria." In: *FEMS Microbiology Reviews* 33.1 (Dec. 2008), pp. 191–205. ISSN: 0168-6445. DOI: 10.1111/j.1574-6976.2008.00149.x.

[BW16] Eric D. Brown and Gerard D. Wright. "Antibacterial drug discovery in the resistance era." In: *Nature* 529.7586 (2016), pp. 336–343. ISSN: 0028-0836. DOI: 10.1038/nature17042.

[Cam+18] Diogo M. Camacho et al. "Next-Generation Machine Learning for Biological Networks." In: *Cell* 173.7 (2018), pp. 1581–1592. ISSN: 0092-8674. DOI: 10.1016/j.cell.2018.05.015.

[Kon+20] JungHo Kong et al. "Network-based machine learning in colorectal and bladder organoid models predicts anti-cancer drug efficacy in patients." In: *Nature Communications* 11.1 (2020), p. 5485. DOI: 10.1038/s41467-020-19313-8.

[Lew+10] Nathan E Lewis et al. "Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models." In: *Molecular Systems Biology* 6.1 (2010), p. 390. ISSN: 1744-4292. DOI: 10.1038/msb.2010.47.

[Lob+15] Michael A. Lobritz et al. "Antibiotic efficacy is linked to bacterial cellular respiration." In: *Proceedings of the National Academy of Sciences* 112.27 (2015), pp. 8173–8180. DOI: 10.1073/pnas.1509743112.

[Ma+18] Jianzhu Ma et al. "Using deep learning to model the hierarchical structure and function of a cell." In: *Nature Methods* 15.4 (2018). ISSN: 1548-7091. DOI: 10.1038/nmeth.4627.

[MMP14] Paul Murima, John D. McKinney, and Kevin Pethe. "Targeting Bacterial Central Metabolism for Drug Development." In: *Chemistry & Biology* 21.11 (2014), pp. 1423–1432. ISSN: 1074-5521. DOI: https://doi.org/10.1016/j.chembiol.2014.08.020.

[Ort+11] Jeff Orth et al. "A comprehensive genome-scale reconstruction of E." In: *Molecular systems biology* 7 (Oct. 2011), p. 535. DOI: 10.1038/msb.2011.65.

[Sto+20] Jonathan M. Stokes et al. "A Deep Learning Approach to Antibiotic Discovery." In: *Cell* 180.4 (2020), 688–702.e13. ISSN: 0092-8674. DOI: 10.1016/j.cell.2020.01.021.

[Wan+14] Tim Wang et al. “Genetic Screens in Human Cells Using the CRISPR-Cas9 System.” In: *Science* 343.6166 (2014), pp. 80–84. ISSN: 0036-8075. DOI: 10.1126/science.1246981.

[Yan+19] Jason H. Yang et al. “A White-Box Machine Learning Approach for Revealing Antibiotic Mechanisms of Action.” In: *Cell* 177.6 (2019). ISSN: 0092-8674. DOI: 10.1016/j.cell.2019.04.016.

[ZH05] Hui Zou and Trevor Hastie. “Regularization and variable selection via the elastic net.” In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.2 (2005), pp. 301–320. ISSN: 1467-9868. DOI: 10.1111/j.1467-9868.2005.00503.x.

[ZY17] Yu Zhang and Qiang Yang. “An overview of multi-task learning.” In: *National Science Review* 5.1 (2017). ISSN: 2095-5138. DOI: 10.1093/nsr/nwx105.