On Modelability and Generalizability: Are Machine Learning Models for Drug Synergy Exploiting Artefacts and Biases in Available Data?

Arushi G K Majha
University of Cambridge
ag920@cam.ac.uk

Ian Stott
Unilever
ian.stott@unilever.com

Andreas Bender
University of Cambridge
ab454@cam.ac.uk

Abstract

Synergy models are useful tools for exploring drug combinatorial search space and identifying promising sub-spaces for in vitro/vivo experiments. Here, we report that distributional biases in the training-validation-test sets used for predictive modeling of drug synergy can explain much of the variability observed in model performances (up to 0.22 Δ*AUPRC*). We built 145 classification models spanning 4,577 unique drugs and 75,276 pair-wise drug combinations extracted from DrugComb, and examined spurious correlations in both the input feature and output label spaces. We posit that some synergy datasets are easier to model than others due to factors such as synergy spread, class separation, chemical structural diversity, physicochemical diversity, combinatorial tests per drug, and combinatorial label entropy. We simulate distribution shifts for these dataset attributes and report that the drug-wise homogeneity of combinatorial labels most influences modelability (0.16 ± 0.06 Δ*AUPRC*). Our findings imply that seemingly high-performing drug synergy models may not generalize well to broader medicinal space. We caution that the synergy modeling community's efforts may be better expended in examining data-specific artefacts and biases rigorously prior to model building.

1 Introduction

For complex, multifactorial diseases such as cancer, combination therapies offer the possibility of enhanced efficacies [19], with reduced effective doses and associated host toxicities [9], as well as a strategy for slowing the evolved drug resistance commonly observed in monotherapies [32]. It is, however, more challenging to perform clinical trials for combination therapies [22] and the large number of possible drug combinations renders exhaustive testing by brute-force heuristics infeasible. Machine learning is a useful tool for exploring the vast drug combinatorial search space and identifying promising sub-spaces for in vitro/vivo experiments.

Currently, research in the field of predictive modeling for drug synergy is largely focused on model generation and the optimization of performance metrics, such as Area under the Receiver Operating Characteristic curve (AUROC or simply AUC), which overestimates model performance on imblanaced datasets [30, 15], rather than the context in which models are built and deployed. There is, at present, no consensus definition for drug synergy [17, 29] and drug combination screens are generally discordant across independent studies [18], yet model improvements are rarely reported in tandem with descriptive statistics characterizing the quality and modelability of datasets. The experimental endpoints are often proxies of drug response that can be easily measured in a high-throughput fashion, but lack clinical relevance or even reproducibility [20].

Biases have been reported in datasets used for model generation in adjacent research fields, such as PDBBind and CASF for the prediction of ligand-protein binding affinities [27]. In a systematic review of 41 genomic machine learning studies, Barnett et al. [2] investigated which components of a study contributed to improvements in model performance and whether reported improvements represent a true improvement or an unaddressed bias inflating performance. They found that data leakage due to feature selection and the number of hyperparameter optimizations were significantly associated with an increase in reported model performance. In a review of 62 machine learning studies on the detection and prognostication of COVID-19 using chest radiographs and chest computed tomography images, Roberts et al. [26] found that none of the models identified were of potential clinical use due to biases in either the methodology or underlying data.

Previous studies on drug synergy prediction have not examined artefacts and biases in dataset composition. To the best of our knowledge, no attempt has been made to quantify the sensitivity of synergy models to underlying distributions in either input feature or output label spaces. Alsherbiny et al. [1] note that the source of drug combination screening data, i.e. NCI-ALMANAC [8] versus ONEIL [21], has a more significant impact on model performance than feature engineering. Similarly, Rani et al. [25] note that synergy models built using NCI-ALMANAC tend to outperform those built using ONEIL. Here, we report that distributional biases in the datasets used for predictive modeling of drug synergy explain much of the variability observed in model performances (up to 0.22 Δ*AUPRC*). We built 145 binary classification models using drug combination screens extracted from DrugComb [35] spanning 4,577 unique drugs and 75,276 pair-wise drug combinations. We characterize the central tendencies and dispersions of various dataset attributes, and subsequently simulate distribution shifts to demonstrate that model performance can improve or deteriorate depending on the direction of attribute shift.

2 Methodology

2.1 Synergy Definition

We use the Bliss Independence model [3], one of several synergy reference models [17, 29], to qualify and quantify the expected additive or null response of administering a drug combination. Operating under assumptions of statistical independence between drugs (i.e., the modes of action of constituent drugs in a combination differ), symmetry in drug interactions, no variability in responses, and continuous dose-response relationships, Bliss excess is defined mathematically as:

$$E_{Bliss} = E_{AB} - (E_A + E_B - E_A \times E_B)$$

where E_{AB} is the observed effect of the drug combination, and E_A and E_B are the observed individual effects of drugs A and B, respectively. E_{Bliss} = 0 is the threshold for additivity, while E_{Bliss} > 0 indicates synergy and E_{Bliss} < 0 indicates antagonism.

2.2 Data Collection and Pre-Processing

Drug pair synergy data targeting 142 cancer cell lines and 3 malarial parasites was extracted from DrugComb v1.5 [35]. Thirty-three percent of drug-drug-cell line tuples were replicate experiments, which we deduplicated by computing the geometric mean synergy score across replicate samples. Thirty-nine percent (N = 306,282) of the combination-cell line tuples were sourced from NCI-ALMANAC [8] and twenty-five percent (N = 198,722) were sourced from FRIEDMAN [12], with the remainder sourced from twenty-two other combination screens including ONEIL [21] (twelve percent; N = 92,208) and CLOUD [14] (five percent; N = 40,160). In total, 75,276 pair-wise drug combinations comprising 4,577 unique drugs were obtained for 145 cell-line synergy endpoints defined by the Bliss Independence model. We selected the top and bottom fifteen percent of each cell-line dataset's distribution of Bliss synergy scores to obtain balanced classes after filtering out additive samples.

2.3 Dataset Attributes and Metrics

Synergicity Synergicity measures the degree to which a given drug is associated with synergistic combinatorial labels: it is defined in this work, as in previous work [34], as the fraction of combinations for which individual drugs have been labelled synergistic as opposed to antagonistic. At the cell-line dataset level, the interquartile range or H-spread was used to capture the bimodality of synergicity distributions and test the hypothesis that cell-line datasets with drugs found primarily in antagonistic-only combinations ($\texttt{synergicity} = 0$) and synergistic-only combinations ($\texttt{synergicity} = 1$) are easier to model with higher AUPRC scores.

Combinatorial Label Entropy Combinatorial label entropy measures the level of disorder or heterogeneity of combinatorial labels. It is defined mathematically as Shannon entropy:

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log_2(P(x_i))$$

where $H(X)$ is the Shannon entropy of a discrete random variable X and $P(x_i)$ is the probability of outcome x_i occurring in the system. The sum is taken over all n possible outcomes x_i. In our case, $H(X)$ has range [0, 1] and measures how homogeneous the combinatorial labels associated with a given drug are: if a drug occurs predominantly in drug combinations labelled synergistic-only or antagonistic-only, then its combinatorial label entropy is low (close to 0); if a drug occurs in drug combinations labelled synergistic approximately half of the time and antagonistic approximately half of the time, then its combinatorial label entropy is high (close to 1).

Feature Similarity Feature similarity in chemical structural and physicochemical spaces was defined in two steps: cosine similarity computed pair-wise amongst all drugs tested per cell line, followed by the cell-line fraction of pair-wise similarities above 0.15. Mathematically, the cosine similarity between two feature vectors A and B is defined as:

$$\text{cosine\_similarity}(A, B) = \frac{A \cdot B}{\|A\| \cdot \|B\|}$$

Non-Additivity A drug's tendency for non-additivity when combined was scored as the median absolute distance from Bliss additivity across combinations. This measure was used to test the hypothesis that a drug's combinatorial label entropy decreases with its tendency for non-additivity in combinations. In other words, non-additivity thus defined was used to test whether the degree of synergism or antagonism achieved by a drug was associated with the consistency or homogeneity of its combinatorial labels.

2.4 Model Generation and Evaluation

We formulate drug synergy prediction as a supervised classification task: we construct one binary model per cell-line dataset, resulting in a total of 145 binary models, to predict synergistic versus antagonistic class labels for drug-drug pairs using the CRAN "randomForest" [13, 24] implementation of the traditional random forest learner by Breiman [4] under default hyperparameter optimizations. Given that the focus of this work is the influence of dataset composition on model performance, and not the influence of model architecture on model performance, we required a single learner to serve as our baseline before and after shifting attribute distributions. We deliberately chose a decision tree ensemble learner as our baseline due to its computational efficiency on high-dimensional data, adequate interpretability and explainability, as well as state-of-the-art model performance on balanced and minority classes [6]. We constructed two sets of drug features: structural 2048-bit Morgan fingerprints (with radius 3) and 43-element long physicochemical profiles of all available molecular descriptors on RDKit [11]. Feature vectors were concatenated for each drug-drug pair in both permutations. Our 80%-20% train-test split strategy was drug-pair–stratified with five-fold cross-validation. To evaluate model performance, we computed Area under the Precision-Recall curve (AUPRC), which is less sensitive to class imbalance and thus more practically relevant and

actionable than Area under the Receiver Operating Characteristic curve (AUROC) [30, 15]. The mean AUPRC across all models ($n = 145$) was 0.76 ± 0.09. For our categorical analyses, we categorized cell-line models with AUPRC greater than or equal to 0.8 as high-performing ($n = 50$), and cell-line models with AUPRC less than 0.8 as low-performing ($n = 95$).

2.5 Simulating Distribution Shifts in Dataset Attributes

We simulated distribution shifts in dataset attributes by sub-sampling each cell-line dataset. For originally high-performing models, we selected subsets of drugs with high combinatorial label entropy (upper 15%), few combinatorial tests per drug (lower 15%), low physicochemical similarity to other drugs (lower 15%), and low structural similarity to other drugs (lower 15%). Conversely, for originally low-performing models, we selected subsets of drugs with low combinatorial label entropy (lower 15%), many combinatorial tests per drug (upper 15%), high physicochemical similarity to other drugs (upper 15%), and high structural similarity to other drugs (upper 15%). This simulated shifts in attribute distributions such that high-performing models now resembled low-performing models, and vice versa. Cell-line models with insufficient drugs remaining were discarded, yielding 103 models for structural similarity, 109 models for physicochemical similarity, 117 models for combinatorial tests per drug, and 91 models for combinatorial label entropy per drug. The simulations were run for each of the dataset attributes identified individually, as well as pair-wise, but the latter yielded datasets too small for model generation. To distinguish change in model performance due to shifting bias versus reduction in dataset size, models were trained, validated, and tested on shifted and non-shifted subsets of comparable size for each cell line.

3 Results

3.1 Synergy Spread and Class Separation

We first analyzed the effect of dataset span, measured as standard deviation of Bliss synergy scores, and class separation, measured as difference in mean Bliss synergy scores of antagonistic vs synergistic classes, on cell-line model performance, measured as AUPRC. The results are shown in Figure 1. It can be seen that high-performing cell-line models tended to exhibit broader synergy spread with difference in means between high– and low–performing models of 15.4–24.1 (95% CI) Bliss synergy units (Welch's two-sample t = 9.13, df = 71.3, p = 1.26e-13). This is consistent with the relationship between potency span and achievable model performance reported by Brown et al. [5] in the context of predicting binding affinity of small-molecule ligands for protein targets. High-performing cell-line models also tended to exhibit greater class separation in synergy space with difference in means between high– and low–performing models of 12.9–17.6 (95% CI) Bliss synergy units (Welch's two-sample t = 13.1, df = 94.4, p < 2.20e-16). Easier class splits may inflate model performance, particularly on AUROC [30, 15] but also AUPRC: DeepSynergy, for instance, defined the top 10% of combinations as the synergistic or positive class and modeled the remainder as the negative class [23]. Our findings show that both synergy spread and class separation influence modelability.

3.2 Synergicity and Entropy of Combinatorial Labels

We then analyzed the effect of combinatorial label homogeneity on model performance (Sub-Figures 2A-B). It can be seen that the cell-line H-spread of synergicity, defined as the fraction of combinations for which individual drugs have been labelled synergistic as opposed to antagonistic, is positively correlated with cell-line model performance, measured as AUPRC (Spearman's $\rho = 0.539$, p = 1.77e-10). Conversely, the cell-line arithmetic mean heterogeneity of combinatorial labels, measured as Shannon entropy for individual drugs, is negatively correlated with cell-line model performance, measured as AUPRC (Pearson's $r = -0.691$, p < 2.20e-16). The more bimodal a cell line's drug synergicity distribution, the more homogeneous its drug-wise combinatorial labels and the easier to predict combinations unseen during training with at least one seen-before drug. Our findings imply that cell lines comprising drugs with homogeneous combinatorial labels, i.e., drugs occurring

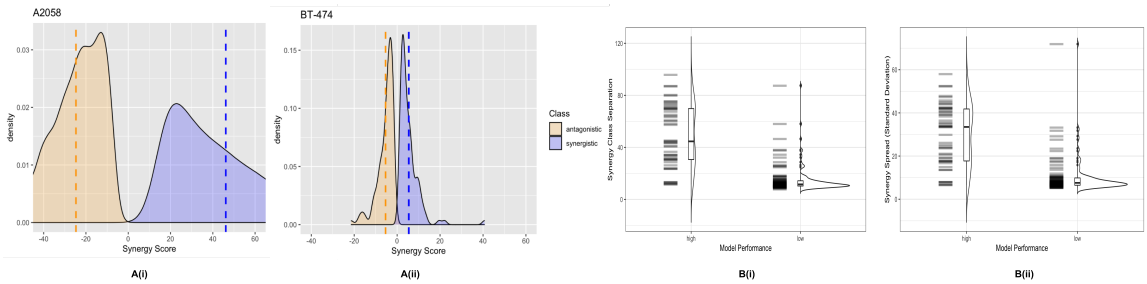

Figure 1: **Panel A.** Distribution of Bliss synergy scores for the best-performing cell-line model **(i)** and the worst-performing cell-line model **(ii)**. **Panel B.** Each barcode line in the violin plots represents one cell-line model. Differences in synergy class means **(i)** and standard deviations of overall synergy distributions **(ii)** plotted for all 145 cell-line models investigated.

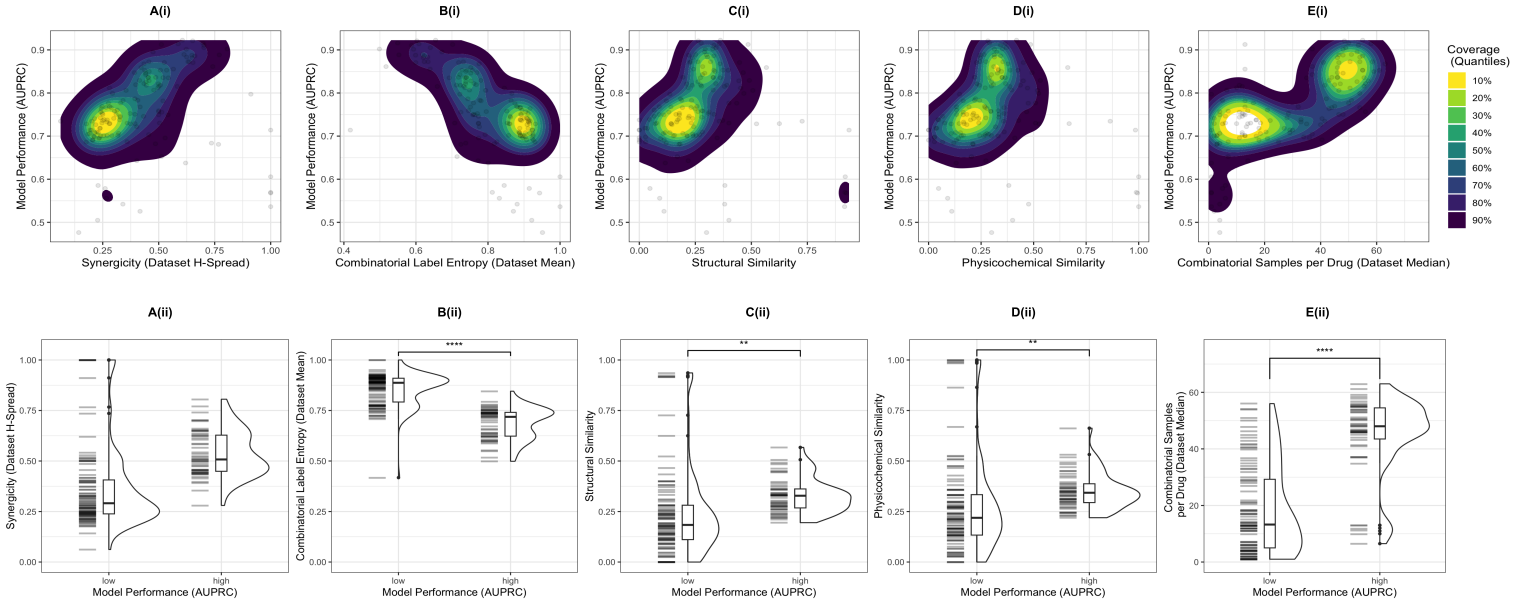

Figure 2: Cell-line model performance plotted against cell-line synergicity (**Panel A**), combinatorial label entropy (**Panel B**), cosine similarity in structural space (**Panel C**), cosine similarity in physicochemical space (**Panel D**), and number of combinatorial samples per drug (**Panel E**). Each dot in the density plots (upper panels) and each barcode line in the violin plots (lower panels) represents one cell-line model.

primarily in antagonistic-only combinatorial labels and synergistic-only combinatorial labels, tend to be easier to model with higher AUPRC scores.

3.3 Structural Diversity, Physicochemical Diversity, Combinatorial Tests Per Drug

We then analyzed the effects of drug diversity in structural Morgan fingerprint and physicochemical spaces, both measured as fraction of drugs in a cell-line dataset with pair-wise cosine similarity above a defined threshold, on cell-line model performance, measured as AUPRC. Panel C of Figure 2 shows that the dataset attribute, compound structural similarity, is positively correlated with model performance (Spearman's ρ = 0.359, p = 1.012e-05): high-performing cell-line models exhibited 3.91%–13.8% (95% CI) higher pair-wise cosine similarity between drugs in Morgan fingerprint space than low-performing cell-line models (Welch's two-sample t = 3.54, df = 132.64, p = 0.0005). Similarly, Panel D of Figure 2 shows that the dataset attribute, compound physicochemical similarity, is positively correlated with model performance (Spearman's ρ = 0.327, p = 6.282e-05): high-

performing cell-line models exhibited 2.28%–12.9% (95% CI) higher pair-wise cosine similarity between drugs in physicochemical space than low-performing cell-line models (Welch's two-sample t = 2.83, df = 131.33, p = 0.005). Summarily, the breadth of compound structural and physicochemical spaces both appear to influence modelability, which one might expect as it is easier to model a smaller space with greater overlap between train and validation/test sets. We subsequently investigated the relationship between cell-line model performance, measured as AUPRC, and number of combinatorial tests per drug. It can be seen in Panel E of Figure 2 that this dataset attribute is positively correlated with model performance (Pearson's r = 0.504, p = 1.24e-10). High-performing cell-line models comprised 17.1-31.0 (95% CI) more combinations tested per drug than low-performing cell-line models (Welch's two-sample t = 6.86, df = 141.19, p = 1.99e-10), which one might expect as it is easier to model a smaller space with fewer distinct drugs tested in more combinations. These findings imply that seemingly high-performing drug synergy models do not generalize well to broader medicinal space.

3.4 Simulating Distribution Shifts in Dataset Attributes

To test whether the differences in model performance observed across cell lines was due to underlying data modelability versus biological variability, we simulated shifts in dataset attribute distributions and compared resulting changes in model performance (Δ*AUPRC*). We selected subsets of drug-drug samples to shift distributions for low-performing cell-line models to resemble high-performing cell-line models, and vice versa. The simulations were run for each of the dataset attributes identified individually, as well as pair-wise, but the latter yielded datasets too small for model generation. The results are summarized in Figure 3.

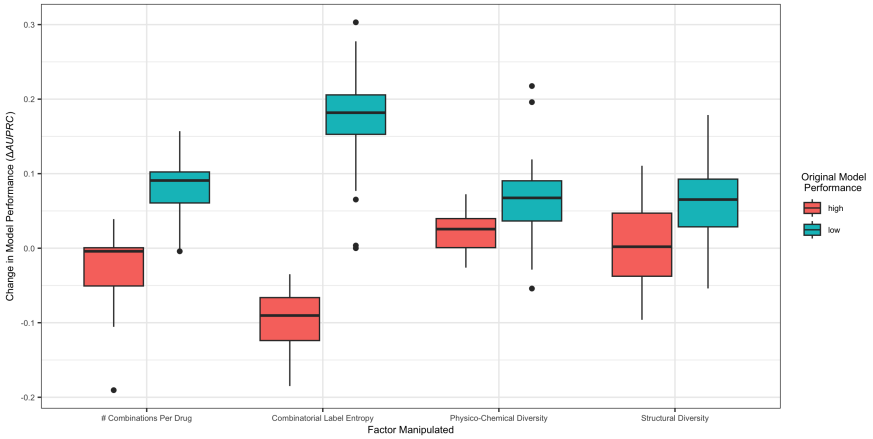

Figure 3: Change in model performance, Δ*AUPRC*, after simulating distribution shifts for each dataset attribute individually. Performance improved for previously low-performing models (blue) under all simulations, albeit to varying degrees ($+0.06 \pm 0.04$ Δ*AUPRC* for physicochemical diversity versus $+0.18 \pm 0.05$ Δ*AUPRC* for combinatorial label entropy). Performance deteriorated most noticeably for previously high-performing models (red) following shifts in distributions for combinatorial label entropy (-0.10 ± 0.04 Δ*AUPRC*).

Sub-sampling that resulted in greater class separation, broader synergy spread, lower structural diversity, lower physicochemical diversity, higher number of combinatorial tests per drug, and lower combinatorial label entropy generally increased model performance. Conversely, sub-sampling that resulted in smaller class separation, narrower synergy spread, lower number of combinatorial tests per drug, and higher combinatorial label entropy generally decreased model performance. In other words, simulating shifts in attribute distributions tended to boost model performance for originally low-performing models, and tended to degrade model performance for originally high-performing models. This suggests that the differences observed in model performance across cell lines was likely due to differences in dataset composition and not due to inherent biological variation. Of the dataset attributes identified and manipulated, combinatorial label entropy most influenced modelability,

increasing the performance of originally low-performing models by $+0.18 \pm 0.05$ $\Delta AUPRC$, which is comparable to the original difference in mean performance between high– versus low–performers (0.15 $\Delta AUPRC$). It is important to note that factors are not decoupled in these simulations as shifting one attribute distribution in isolation was not feasible; shifting one distribution simultaneously shifted other distributions to varying degrees since we must also consider how dataset attributes are correlated with each other. To contextualize these findings, we refer to improvements over state-of-the-art models reported in drug synergy literature, such as $+0.04$ $\Delta AUPRC$ by Preuer et al. [23] and Wang et al. [31].

3.5 Drug Synergicity and Lipophilicity

We then analyzed whether mechanistic insights reported in drug synergy literature, particularly the relationship between synergicity and lipophilicity [34], influence modelability. Figure 4A shows that, for the well-characterized cell line MCF7, a drug's lipophilicity (CrippenClogP) is positively correlated (Pearson's r = 0.452, p = 0.0000143) with its synergicity, particularly in the region most relevant for drug discovery, i.e., CrippenClogP interval (1,6]. Figure 4B shows the negative correlation between model performances and synergicity-lipophilicity correlation coefficients for all cell-line datasets (Pearson's $r = -0.263$, p = 0.00144): 46.6% of cell-line datasets exhibited synergicity-lipophilicity correlation coefficients ≥ 0.2, but only 13.2% of these had model performances AUPRC ≥ 0.8. High-performing models evidently do not rely on the positive correlation between synergicity and lipophilicity reported here and in literature [34] for predictions: high-performing cell-line datasets exhibited a near-zero mean synergicity-lipophilicity correlation coefficient of 0.026 PCC, 95% *CI* [-0.290,-0.164] lower than the mean correlation of 0.253 PCC for low-performing cell-line datasets (Welch's two-sample t = -7.14, df = 121, p < 0.000001).

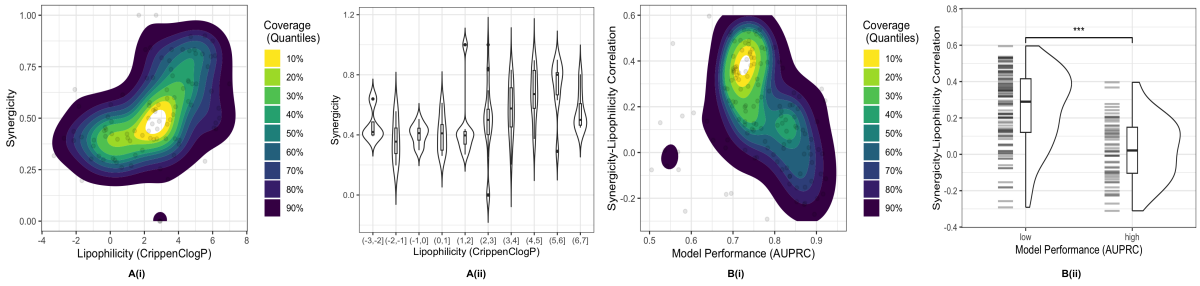

Figure 4: **Panel A.** Each dot in the density plot and each barcode line in the violin plot represents one drug. A drug's lipophilicity is postively correlated with its synergicity in the MCF7 cell-line dataset, particularly for drug-like molecules in CrippenClogP interval (1,6]. **Panel B.** Each dot in the density plot and each barcode line in the violin plot represents one cell-line model. Synergicity-lipophilicity correlation coefficients are negatively correlated with model performance across cell-line datasets.

3.6 Non-Additivity, Combinatorial Label Homogeneity, Drug Similarity

We considered the dependence of combinatorial label homogeneity, an output dataset attribute, on various input dataset attributes, such as drug similarity. It can be seen in Appendix Figure 7 that cell-line drug similarity in physicochemical (Pearson's r = 0.480) and structural (Pearson's r = 0.514) spaces correlate with combinatorial label homogeneity. A drug is more likely to behave generally synergistically or generally antagonistically, or rather elicit mostly synergistic-only or antagonistic-only labels, when combined with similar drugs, since similar drugs hit similar pathways exhibiting homogeneous synergistic *or* antagonistic effect. Different drugs hit different pathways exhibiting heterogeneous synergistic *and* antagonistic effect: synergy with some drugs and antagonism with other drugs depending on pathway hit [16]. We then considered the relationship between a drug's combinatorial label homogeneity and its tendency for non-additivity, defined in this work as median absolute distance from Bliss additivity across combinations. The correlation between

these attributes varied across cell-line models and tended to increase with dataset modelability or increasing model performance in AUPRC (Pearson's $r = 0.378$, Figure 5A). High-performing cell-line models comprised drugs exhibiting a stronger correlation between combinatorial label homogeneity and non-additivity with a 95% *CI* [0.091,0.241] higher Pearson correlation coefficient (PCC) than low-performing cell-line models (Welch's two-sample t = 4.39, df = 115, p < 0.00002). 19.4% of cell-line datasets exhibited PCCs between combinatorial label homogeneity and non-additivity ≥ 0.5. Of these, 75% had model performances AUPRC ≥ 0.8. Figure 5B shows one such cell-line dataset, namely the skin epithelial-like cell line IST-MEL1, with AUPRC ≥ 0.9 and PCC between combinatorial label homogeneity and non-additivity $r = 0.643$. In other words, drugs that elicited close-to-additive effects when combined tended to have low combinatorial label homogeneity, while drugs that elicited highly synergistic or highly antagonistic effects when combined tended to have high combinatorial label homogeneity. These findings imply that combinatorial label homogeneity could function as a crude proxy for non-additivity in some contexts, yielding greater modelability.

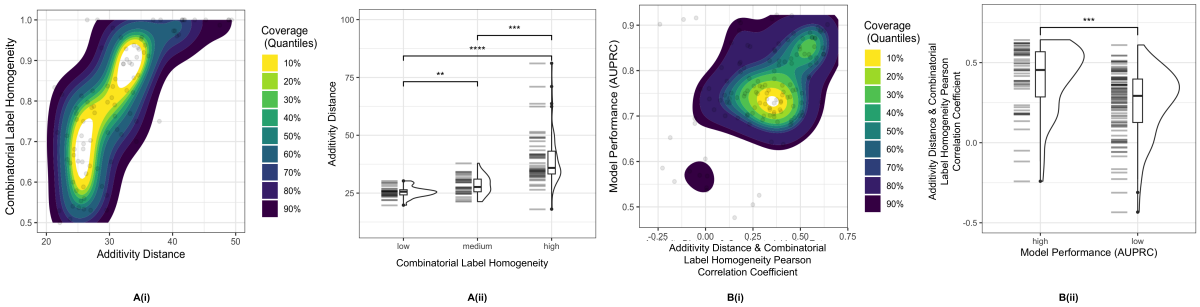

Figure 5: **Panel A.** Each dot in the density plot and each barcode line in the violin plot represents one drug. A drug's combinatorial label homogeneity is positively correlated with its non-additivity for the IST-MEL1 cell-line dataset. **Panel B.** Each dot in the density plot and each barcode line in the violin plot represents one cell-line model. Model performance tended to increase with increasing strength of Pearson's correlation between combinatorial label homogeneity and non-additivity.

4 Conclusions

In this work, we qualify and quantify various synergy dataset attributes influencing modelability: synergy spread, class separation, chemical structural diversity, physicochemical diversity, combinatorial tests per drug, and combinatorial label entropy. We simulate shifts in distributions of these attributes and report that combinatorial label entropy improved and degraded model performance most, depending on the direction of attribute shift. It is important to note that the attributes were not decoupled in our simulations as shifting one attribute distribution in isolation was not feasible; shifting one distribution simultaneously shifted other distributions to varying degrees. Overall, our findings imply that model performance is highly sensitive to distributional biases in available data. We find that distributional biases in the training-validation-test sets used for predictive modeling of drug synergy can explain up to 0.22 Δ*AUPRC* of the difference observed in model performances. For comparison, we refer to performance improvements over state-of-the-art models reported in drug synergy literature, such as 0.04 Δ*AUPRC* by Preuer et al. [23] and Wang et al. [31]. We caution that the synergy modeling community's efforts may be better expended in examining data-specific artefacts and biases rigorously prior to model building. We recommend that synergy modelers characterize the applicability domain wherein models can be expected to work reliably and report explicitly the statistical biases underlying datasets used for model generation.

References

[1] Alsherbiny, M. A., Radwan, I., Moustafa, N., Bhuyan, D. J., El-Waisi, M., Chang, D. and Li, C. G. [2023], 'Trustworthy Deep Neural Network for Inferring Anticancer Synergistic Combinations', *IEEE Journal of Biomedical and Health Informatics* **27**(4), 1691–1700.

[2] Barnett, E., Onete, D., Salekin, A. and Faraone, S. V. [2022], 'Genomic Machine Learning Meta-regression: Insights on Associations of Study Features with Reported Model Performance', *medRxiv* pp. 2022–01.

[3] Bliss, C. I. [1939], 'The Toxicity of Poisons Applied Jointly', *Annals of applied biology* **26**(3), 585–615.

[4] Breiman, L. [2001], 'Random forests', *Machine learning* **45**, 5–32.

[5] Brown, S. P., Muchmore, S. W. and Hajduk, P. J. [2009], 'Healthy skepticism: assessing realistic model performance', *Drug discovery today* **14**(7-8), 420–427.

[6] Chen, J., Wu, L., Liu, K., Xu, Y., He, S. and Bo, X. [2023], 'EDST: a decision stump based ensemble algorithm for synergistic drug combination prediction', *BMC bioinformatics* **24**(1), 1–21.

[7] Cortés-Ciriano, I. and Bender, A. [2016], 'How consistent are publicly reported cytotoxicity data? Large-scale statistical analysis of the concordance of public independent cytotoxicity measurements', *ChemMedChem* **11**(1), 57–71.

[8] Holbeck, S. L., Camalier, R., Crowell, J. A., Govindharajulu, J. P., Hollingshead, M., Anderson, L. W., Polley, E., Rubinstein, L., Srivastava, A., Wilsker, D. et al. [2017], 'The National Cancer Institute ALMANAC: a comprehensive screening resource for the detection of anticancer drug pairs with enhanced therapeutic activity', *Cancer research* **77**(13), 3564–3576.

[9] Jia, J., Zhu, F., Ma, X., Cao, Z. W., Li, Y. X. and Chen, Y. Z. [2009], 'Mechanisms of drug combinations: interaction and network perspectives', *Nature reviews Drug discovery* **8**(2), 111–128.

[10] Khetan, R., Curtis, R., Deane, C. M., Hadsund, J. T., Kar, U., Krawczyk, K., Kuroda, D., Robinson, S. A., Sormanni, P., Tsumoto, K. et al. [2022], Current advances in biopharmaceutical informatics: guidelines, impact and challenges in the computational developability assessment of antibody therapeutics, *in* 'MAbs', Vol. 14, Taylor & Francis, p. 2020082.

[11] Landrum, G., Tosco, P., Kelley, B., Sriniker, Gedeck, NadineSchneider, Vianello, R., Ric, Dalke, A., Cole, B., AlexanderSavelyev, Swain, M., Turk, S., N, D., Vaucher, A., Kawashima, E., Wójcikowski, M., Probst, D., Godin, G., Cosgrove, D., Pahl, A., JP, Francois Berenger, Strets123, JLVarjo, O'Boyle, N., Fuller, P., Jensen, J. H., Sforna, G. and DoliathGavid [2020], 'rdkit/rdkit: 2020_03_1 (q1 2020) release'.
URL: *https://zenodo.org/record/3732262*

[12] *Landscape of targeted anti-cancer drug synergies in melanoma identifies a novel BRAF-VEGFR/PDGFR combination treatment, author=Friedman, Adam A and Amzallag, Arnaud and Pruteanu-Malinici, Iulian and Baniya, Subash and Cooper, Zachary A and Piris, Adriano and Hargreaves, Leeza and Igras, Vivien and Frederick, Dennie T and Lawrence, Donald P and others* [2015], *PloS one* **10**(10), e0140310.

[13] Liaw, A. and Wiener, M. [2002], 'Classification and Regression by randomForest', *R News* **2**(3), 18–22.
URL: *https://cran.r-project.org/package=randomForest*

[14] Licciardello, M. P., Ringler, A., Markt, P., Klepsch, F., Lardeau, C.-H., Sdelci, S., Schirghuber, E., Müller, A. C., Caldera, M., Wagner, A. et al. [2017], 'A combinatorial screen of the CLOUD uncovers a synergy targeting the androgen receptor', *Nature chemical biology* **13**(7), 771–778.

[15] Lobo, J. M., Jiménez-Valverde, A. and Real, R. [2008], 'AUC: a misleading measure of the performance of predictive distribution models', *Global ecology and Biogeography* **17**(2), 145–151.

[16] Martin, Y. C., Kofron, J. L. and Traphagen, L. M. [2002], 'Do structurally similar molecules have similar biological activity?', *Journal of medicinal chemistry* **45**(19), 4350–4358.

[17] Meyer, C. T., Wooten, D. J., Lopez, C. F. and Quaranta, V. [2020], 'Charting the fragmented landscape of drug synergy', *Trends in pharmacological sciences* **41**(4), 266–280.

[18] Nair, N. U., Greninger, P., Zhang, X., Friedman, A. A., Amzallag, A., Cortez, E., Sahu, A. D., Lee, J. S., Dastur, A., Egan, R. K. et al. [2023], 'A landscape of response to drug combinations in non-small cell lung cancer', *Nature Communications* **14**(1), 3830.

[19] Narayan, R. S., Molenaar, P., Teng, J., Cornelissen, F. M., Roelofs, I., Menezes, R., Dik, R., Lagerweij, T., Broersma, Y., Petersen, N. et al. [2020], 'A cancer drug atlas enables synergistic targeting of independent drug vulnerabilities', *Nature communications* **11**(1), 2935.

[20] Niepel, M., Hafner, M., Mills, C. E., Subramanian, K., Williams, E. H., Chung, M., Gaudio, B., Barrette, A. M., Stern, A. D., Hu, B. et al. [2019], 'A multi-center study on the reproducibility of drug-response assays in mammalian cell lines', *Cell systems* **9**(1), 35–48.

[21] O'Neil, J., Benita, Y., Feldman, I., Chenard, M., Roberts, B., Liu, Y., Li, J., Kral, A., Lejnine, S., Loboda, A. et al. [2016], 'An unbiased oncology compound screen to identify novel combination strategies', *Molecular cancer therapeutics* **15**(6), 1155–1162.

[22] Pemovska, T., Bigenzahn, J. W. and Superti-Furga, G. [2018], 'Recent advances in combinatorial drug screening and synergy scoring', *Current opinion in pharmacology* **42**, 102–110.

[23] Preuer, K., Lewis, R. P., Hochreiter, S., Bender, A., Bulusu, K. C. and Klambauer, G. [2018], 'DeepSynergy: predicting anti-cancer drug synergy with Deep Learning', *Bioinformatics* **34**(9), 1538–1546.

[24] R Core Team [2023], *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/*

[25] Rani, P., Dutta, K. and Kumar, V. [2023], 'Performance evaluation of drug synergy datasets using computational intelligence approaches', *Multimedia Tools and Applications* pp. 1–27.

[26] Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., Aviles-Rivero, A. I., Etmann, C., McCague, C., Beer, L. et al. [2021], 'Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans', *Nature Machine Intelligence* **3**(3), 199–217.

[27] Scantlebury, J., Vost, L., Carbery, A., Hadfield, T. E., Turnbull, O. M., Brown, N., Chenthamarakshan, V., Das, P., Grosjean, H., von Delft, F. et al. [2022], 'A Step Towards Generalisability: Training a Machine Learning Scoring Function for Structure-Based Virtual Screening', *bioRxiv* pp. 2022–10.

[28] Shim, M., Lee, S.-H. and Hwang, H.-J. [2021], 'Inflated prediction accuracy of neuropsychiatric biomarkers caused by data leakage in feature selection', *Scientific Reports* **11**(1), 7980.

[29] Tang, J., Wennerberg, K. and Aittokallio, T. [2015], 'What is synergy? The Saariselkä agreement revisited', *Frontiers in pharmacology* **6**, 181.

[30] *The relationship between Precision-Recall and ROC curves, author=Davis, Jesse and Goadrich, Mark* [2006], *in* 'Proceedings of the 23rd International Conference on Machine Learning', pp. 233–240.

[31] Wang, T., Wang, R. and Wei, L. [2023], 'AttenSyn: An Attention-Based Deep Graph Neural Network for Anticancer Synergistic Drug Combination Prediction', *Journal of Chemical Information and Modeling* .

[32] Worthington, R. J. and Melander, C. [2013], 'Combination approaches to combat multidrug-resistant bacteria', *Trends in biotechnology* **31**(3), 177–184.

[33] Wu, L., Wen, Y., Leng, D., Zhang, Q., Dai, C., Wang, Z., Liu, Z., Yan, B., Zhang, Y., Wang, J. et al. [2022], 'Machine learning methods, databases and tools for drug combination prediction', *Briefings in bioinformatics* **23**(1), bbab355.

[34] Yilancioglu, K., Weinstein, Z. B., Meydan, C., Akhmetov, A., Toprak, I., Durmaz, A., Iossifov, I., Kazan, H., Roth, F. P. and Cokol, M. [2014], 'Target-independent prediction of drug synergies using only drug lipophilicity', *Journal of chemical information and modeling* **54**(8), 2286–2293.

[35] Zheng, S., Aldahdooh, J., Shadbahr, T., Wang, Y., Aldahdooh, D., Bao, J., Wang, W. and Tang, J. [2021], 'DrugComb update: a more comprehensive drug sensitivity data repository and analysis portal', *Nucleic acids research* **49**(W1), W174–W184.

5 Appendices

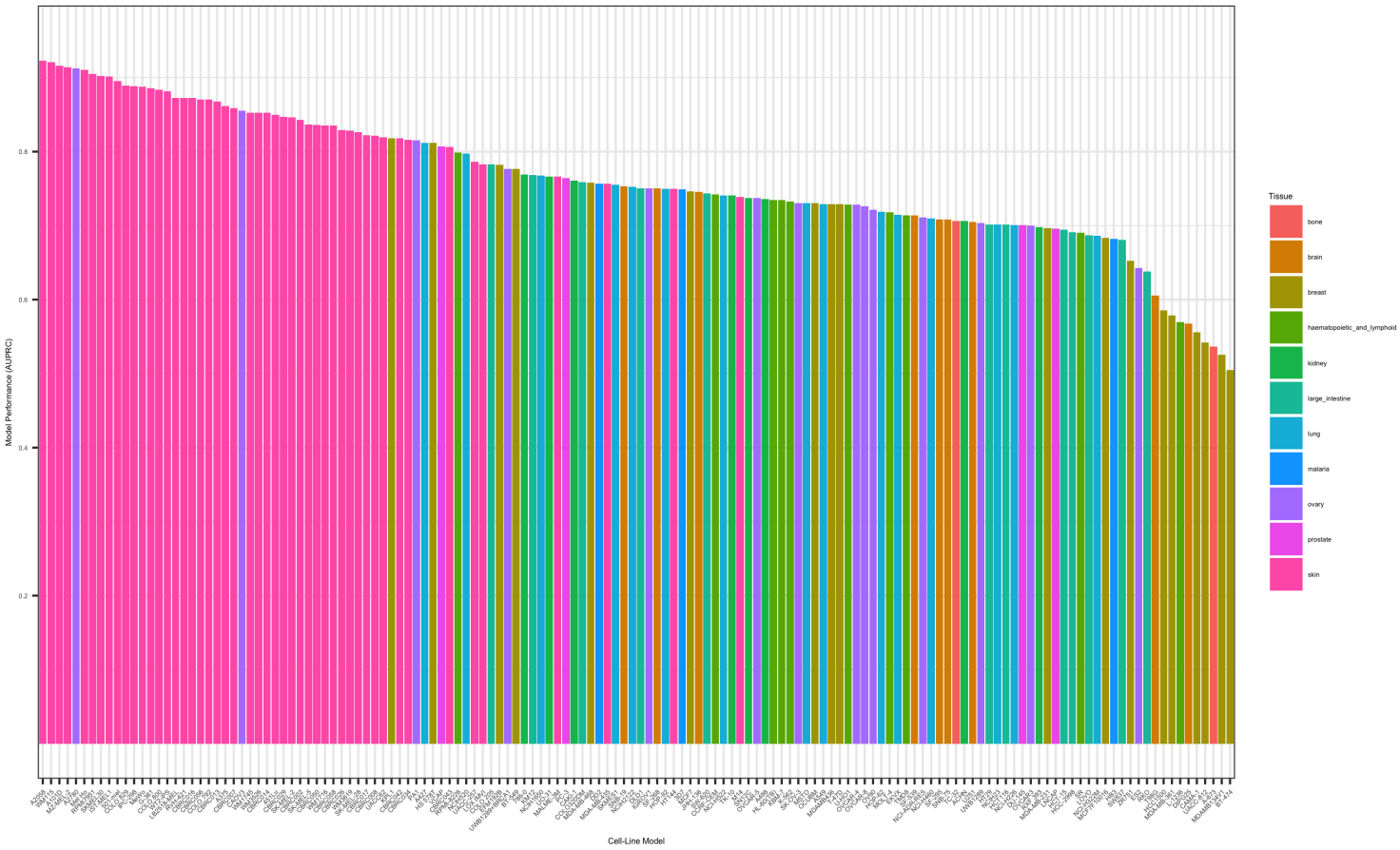

Figure 6: AUPRC performances for all cell-line models investigated in this study.

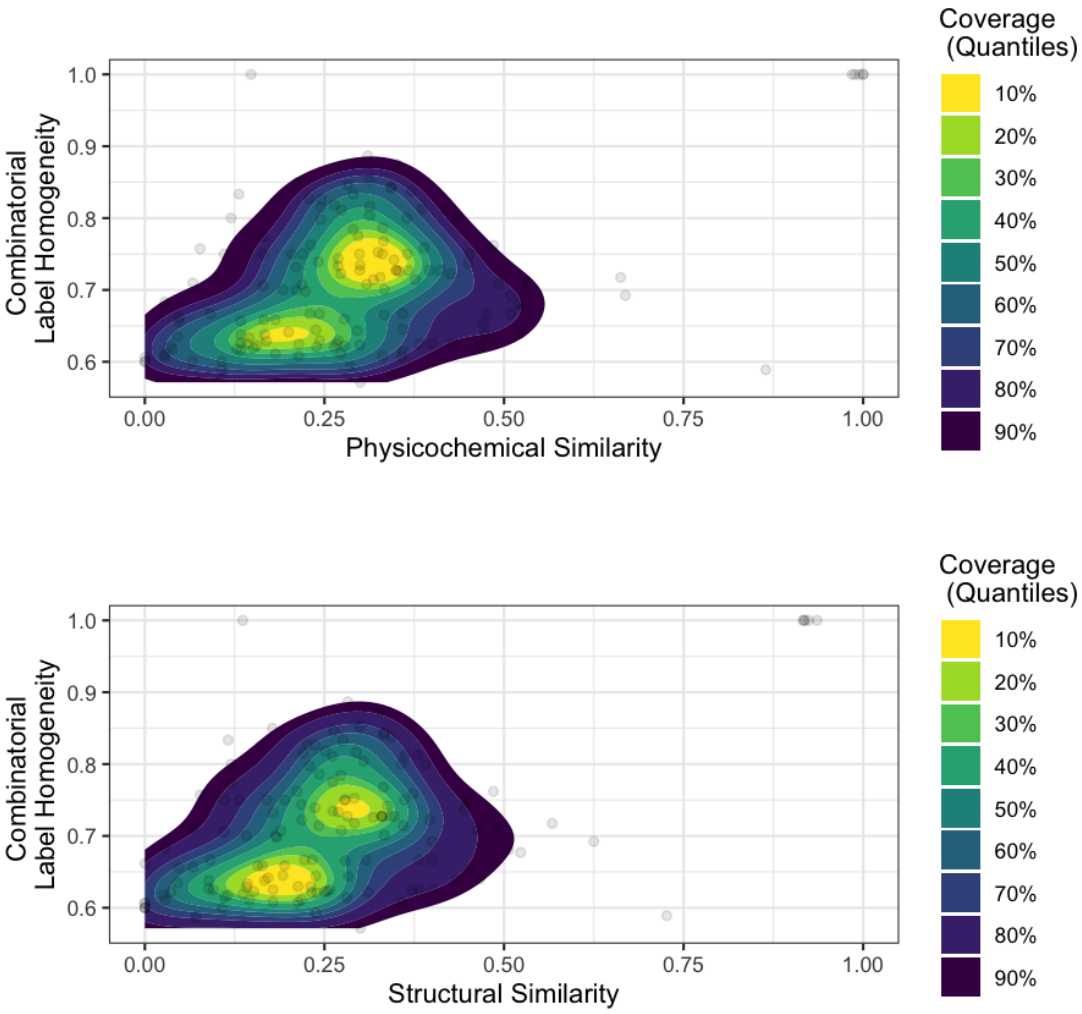

Figure 7: Drug similarity in physicochemical (upper) and structural (lower) spaces correlate with combinatorial label homogeneity.