EFFICIENCY AND TRANSFERABILITY OF INDUCTIVE MONDRIAN CONFORMAL PREDICTORS FOR DRUG–DRUG SYNERGY

Arushi G K Majha [1] **Ian Stott** [2] **Andreas Bender** [1]

Abstract

We propose a principled approach to quantifying prediction uncertainty in machine learning for drug-drug synergy, a burgeoning subfield within drug discovery where human decision-makers require a clear understanding of the errors associated with predictions. To address the limitations of traditional point prediction models typically outputting a single value (for regression settings) or a single label (for classification settings) without any measure of uncertainty, we introduce Mondrian inductive conformal prediction for drug-drug synergy with probabilistic guarantees on the accuracy of each prediction. By providing statistically valid prediction regions at predefined confidence levels, inductive Mondrian conformal predictors enhance the interpretability and reliability of computational drug-drug synergy models, with observed unconfidence and fuzziness scores of 0.13 ± 0.02.

1. Introduction

Conformal prediction is a machine learning framework that extends traditional prediction models by providing probabilistic guarantees on the accuracy of predictions (Vovk et al., 2005). Instead of outputting a single label or value without any indication of prediction uncertainty, conformal prediction provides prediction sets (in classification settings) or intervals (in regression settings) that contain the true target with a certain user-defined probability.

The key concept in conformal prediction is *validity*. A conformal predictor is considered *valid* if, under repeated application to different datasets, it guarantees that the true target will fall within the predicted set or interval with a predefined probability. This contrasts with traditional point prediction models that do not offer such guarantees and may produce inaccurate predictions without any measure of uncertainty associated with those predictions.

Conformal prediction offers several advantages over similar methods to estimate the reliability of predictive models in medicinal chemistry (Norinder et al., 2014; Svensson et al., 2018). Applicability domain methods that define distance between test and training instances in descriptor space to identify promising subspaces where the model can be expected to work reliably do not guarantee statistical validity; the fraction of test instances whose true value or label lies within a prediction region is not guaranteed to be proportional to a given, user-defined confidence level (Cortés-Ciriano & Bender, 2019). Methods whose outputs are well-calibrated probability distributions have a high computational cost; Gaussian process methods, for example, require the inversion of large covariance matrices during the training phase. Conformal predictors are flexible and can be integrated with any machine learning algorithm at low computational cost (Devetyarov & Nouretdinov, 2010; Svensson et al., 2018), offering ease of interpretability for the computed prediction regions in both classification and regression settings.

This work proposes a principled approach to quantifying prediction uncertainty in machine learning for drug-drug synergy, a burgeoning subfield within drug discovery where human decision-makers require a clear understanding of the errors associated with predictions. We introduce Mondrian inductive conformal prediction for drug-drug synergy with probabilistic guarantees on the accuracy of each prediction, demonstrating high conformal efficiencies (at observed unconfidence and fuzziness scores of 0.129 ± 0.022) and inter-species transferability at comparable efficiencies. By providing statistically valid prediction regions at predefined confidence levels, inductive Mondrian conformal predictors enhance the interpretability and reliability of computational drug-drug synergy models.

[1]Centre for Molecular Informatics, University of Cambridge. [2]Unilever. Correspondence to: Arushi G K Majha <ag920@cam.ac.uk>.

International Conference on Machine Learning Workshop on Artificial Intelligence for Science. Vienna, Austria, 2024.

2. Methodology

2.1. Drug Combination Screening Data

The conformal predictors in this work were trained, calibrated, and tested on a recently published, open-access database (AADB) spanning 3,035 combinations of 83 antibiotics with 226 adjuvants tested against 325 bacterial strains, compiled from 106 scientific articles by human experts (Lv et al., 2023b). Transferability experiments were conducted on Gram-negative bacterial strains with the highest number of samples in the database: *Escherichia coli*, *Pseudomonas aeruginosa*, and *Salmonella typhimurium*. Seven features extracted from AADB were used as molecular descriptors and are listed in Table 1; the first five constitute critical physicochemical paramaters in Lipinski's rule of five for evaluating drug-likeness (Lipinski, 2001; 2004).

2.2. Drug-Drug Synergy Definition

We used the synergy assignments in AADB (Lv et al., 2023b) computed according to the Bliss Independence model for the null or expected additive response of administering a drug-drug combination (Bliss, 1939). This synergy reference model assumes statistical independence between drugs (i.e., the modes of action of constituent drugs in a combination differ), symmetry in drug interactions, zero variability in responses, and continuous dose-response relationships. Mathematically, Bliss excess is defined as:

$$E_{Bliss} = E_{AB} - (E_A + E_B - E_A \times E_B) \qquad (1)$$

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.

2.3. Inductive Conformal Prediction Framework

In the conformal prediction framework (Vovk et al., 2005; 2016), a prediction set (Γ^ϵ) is deemed credible based on the training data (z_1, ..., z_l) consisting of labelled objects or examples (x_i, y_i), the non-conformity measure (A) for generating p-values, and the user-defined significance level (ϵ). More formally, a conformal prediction set is defined for a new test object (x) as the set of labels (y), for which the associated p-value (p^y) is greater than ϵ:

$$\Gamma^\epsilon(z_1, ..., z_l, x) := y \mid p^y > \epsilon \qquad (2)$$

The transductive approach to conformal prediction is online and computationally expensive as the underlying model must be retrained for each new test object. In inductive conformal prediction (ICP), training is off-line with the training set split into a proper training set for the base learner and a calibration set for generating non-conformity scores (Papadopoulos, 2008). However, standard ICP does not guarantee class-conditional *local validity*: a lower error rate for the majority class may compensate a higher error rate for the minority class, such that predictions are still statistically valid overall, satisfying *global validity*. Mondrian inductive conformal predictors (MICPs) offer class-specific calibration of validity, as well as null-class predictions where there is insufficient evidence for the MICP model to include either class in the conformal prediction set, and dual-class predictions where the threshold for significance is sufficiently low to include both classes in the conformal prediction set (Vovk, 2012).

An MICP constructs class-conditional, statistically valid prediction sets based on labelled calibration data. The non-conformity measure assesses the similarity between new test objects and training examples, creating prediction sets that contain the true label with a predefined probability (tolerated error rate) for each class locally, as well as both classes globally, in the binary classification setting. The size of the prediction region reflects the uncertainty associated with the prediction. The p^y-value of a new test object is determined by comparing its non-conformity score to the class-wise distribution of non-conformity scores for the calibration set. The class-wise sorted lists of non-conformity scores are termed *Mondrian class lists*. First, the number of calibration examples per class with non-conformity scores lower than or equal to the non-conformity score of the test object is computed. This count is then compared to the size of the Mondrian class lists to determine the size of the prediction region.

Class-conditional, off-line MICPs were constructed with Random Forests (Breiman, 2001) as the underlying machine learning method using the open-source R packages, *caret* (Kuhn, 2015) and *conformal* (Cortés-Ciriano et al., 2015). The number of decision trees was set to 100, and 5-fold cross-validation was selected as the resampling method for hyperparameter tuning and model selection. The *non-conformity measure* was defined as the fraction of decision trees in the forest voting for a given class. Isotonic regression was applied as a post-processing technique to improve calibration in the binary classification setting.

2.4. Defining Conformal Efficiency

2.4.1. ϵ-Free Criteria

The quality of MICP predictions was measured against three ϵ-free criteria (Vovk et al., 2016), under which efficiency does not depend on the user-defined significance level, ϵ. The simplest of these criteria is the *S criterion* (with "S"

Table 1. Seven features extracted from AADB (Lv et al., 2023b) used as molecular descriptors for all RF-based MICP models in this work.

FEATURE	DESCRIPTION
LOGP	OIL-WATER PARTITION COEFFICIENT
HBA	NUMBER OF HYDROGEN BOND ACCEPTORS
HBD	NUMBER OF HYDROGEN BOND DONORS
TPSA	TOTAL POLAR SURFACE AREA
ROTB	NUMBER OF ROTATABLE BONDS
AROM	NUMBER OF AROMATIC RINGS
ALERTS	NUMBER OF STRUCTURAL ALERTS

standing for "sum"), which defines efficiency as the average sum of p-values, p_i^y, generated from the Mondrian class list for each y in the label space and all k test instances:

$$S = \frac{1}{k}\sum_{i=1}^{k}\sum_{y} p_i^y \tag{3}$$

The *U criterion* (where "U" stands for "unconfidence") in Equation 4 defines efficiency as the average unconfidence across all k test instances, with *unconfidence* defined for an individual test instance as the second-largest p-value. In the case of binary classification, this criterion is equivalent to the *F criterion* (where "F" stands for "fuzziness") defined in Equation 5 as the average fuzziness across all k test instances, with *fuzziness* defined for an individual test instance as the sum of all p-values excepting the largest one. Smaller values are preferable, indicating higher conformal efficiency.

$$U = \frac{1}{k}\sum_{i=1}^{k}\min_{y}\max_{y'\neq y} p_i^{y'} \tag{4}$$

$$F = \frac{1}{k}\sum_{i=1}^{k}\left(\sum_{y} p_i^y - \max_{y} p_i^y\right) \tag{5}$$

The *OU* ("observed unconfidence") *criterion* defined in Equation 6 measures efficiency as the average observed unconfidence across all k test instances, with *observed unconfidence* for an individual test instance defined as the largest p-value for false labels $y \neq y_i$. In the case of binary classification, this criterion is equivalent to the *OF* ("observed fuzziness") *criterion* in Equation 6, which defines efficiency as the average observed fuzziness across all k test instances, with *observed fuzziness* for an individual test instance defined as the sum of all p-values for false labels $y \neq y_i$. Smaller values are preferable, indicating higher conformal efficiency.

$$OU = \frac{1}{k}\sum_{i=1}^{k}\max_{y\neq y_i} p_i^y \tag{6}$$

$$OF = \frac{1}{k}\sum_{i=1}^{k}\sum_{y\neq y_i} p_i^y \tag{7}$$

Unlike the OU and OF criteria, the S, U, and F criteria are *prior criteria* that do not depend on observed labels.

2.4.2. ϵ-DEPENDENT CRITERIA

The quality of MICP predictions was further assessed against three ϵ-dependent criteria (Vovk et al., 2016), under which efficiency is a function of the user-defined significance level, ϵ. The simplest of these criteria is the *N criterion* (with "N" standing for "number"), which defines efficiency as the average size (number of labels) of the prediction sets, $\Gamma^\epsilon := \{y \mid p^y > \epsilon\}$:

$$N = \frac{1}{k}\sum_{i=1}^{k}|\Gamma_i^\epsilon| \tag{8}$$

Values closer to 1 are preferable, indicating higher conformal efficiency.

The *M* ("multiple") *criterion* in Equation 11 defines efficiency as the percentage of k test instances with prediction sets containing more than one label. The *OM* ("observed multiple") *criterion* in Equation 12 defines efficiency as the percentage of k test instances with prediction sets containing a false label. In the case of binary classification, these criteria are equivalent to the *E* ("excess") *criterion* and *OE* ("observed excess") *criterion*, respectively (see Appendix 4). Smaller values are preferable, indicating higher conformal efficiency.

$$M = \frac{1}{k}\sum_{i=1}^{k}\mathbf{1}_{\{|\Gamma_i^\epsilon|>1\}} \tag{9}$$

$$OM = \frac{1}{k}\sum_{i=1}^{k}\mathbf{1}_{\{\Gamma_i^\epsilon\setminus\{y_i\}\neq\emptyset\}} \tag{10}$$

Unlike the OM criterion, the N and M criteria are *prior criteria* that do not depend on observed labels.

Table 2. Conformal prediction efficiencies against various ϵ-free criteria.

CRITERION	EFFICIENCY		
	GLOBAL	ANTAGONISTIC	SYNERGISTIC
S (SUM OF p-VALUES)	0.697 ± 0.026	0.675 ± 0.024	0.719 ± 0.037
U (UNCONFIDENCE) & F (FUZZINESS)	0.066 ± 0.008	0.079 ± 0.008	0.054 ± 0.013
OU & OF (OBSERVED UNCONFIDENCE & FUZZINESS)	0.129 ± 0.022	0.148 ± 0.028	0.111 ± 0.031

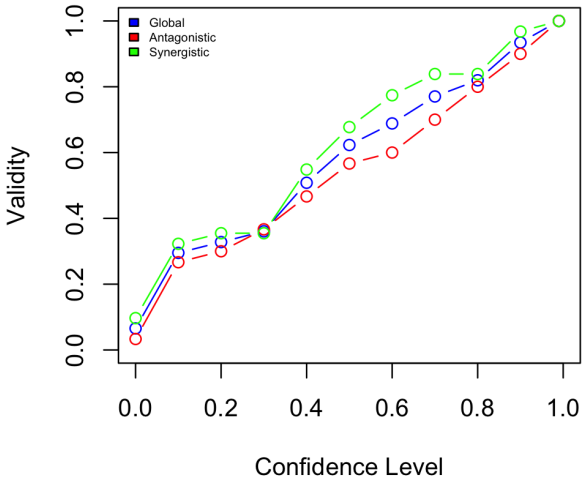

Figure 1. Calibration plot showing global validity and label-wise local validities versus the user-specified confidence level ($1 - \epsilon$).

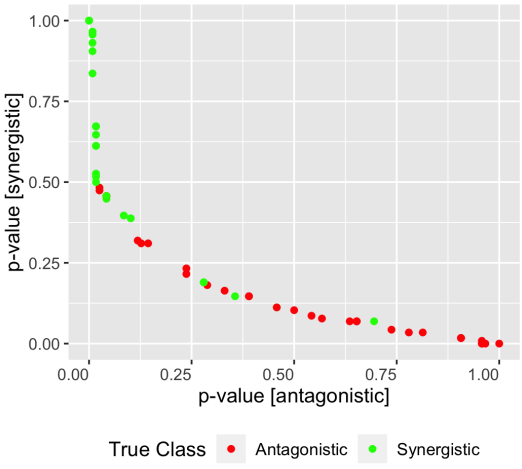

Figure 2. Conformal prediction p-values computed using Mondrian class lists.

3. Results

3.1. Conformal Validity

Calibration ensures that the prediction regions produced by MICPs have the specified coverage probability or confidence level, $1 - \epsilon$. The calibration plot in Figure 1 shows the statistical validity of MICP classification globally as well as in a class-conditional manner for both synergistic and antagonistic drug-drug combinations. At a confidence level of 80%, the property of statistical validity guarantees that 80% of the predictions deemed reliable will be correct.

3.2. Prior and Observed Conformal Efficiencies

3.2.1. ϵ-FREE CONFORMAL EFFICIENCIES

The class-wise p^y-values produced for the test set according to the MICP framework are shown in Figure 2. Synergistic drug-drug combinations tend to be assigned a higher p-value for the synergistic class label versus the antagonistic class label, and vice versa for antagonistic drug-drug combinations. This illustrates that the MICP model tends to assign higher reliability to the correct label; in ambiguous cases that are less conforming to the training examples, the MICP model tends to assign low reliability to both labels.

Table 2 lists the ϵ-free efficiency scores. According to the prior S criterion, which is the average sum of p^y-values over the test set, the constructed MICP model has only moderate efficiency, but this is unalarming as even a maximally efficient conformal predictor will produce high p-values for the true class label, thereby increasing the S score (Vovk et al., 2016). According to the observed OU and OF criteria, which are equivalent in the binary classification setting, the constructed MICP model is highly efficient with observed unconfidence and fuzziness scores of 0.129 ± 0.022 globally, indicating that the MICP can successfully assign low p-values to incorrect class labels. According to the prior U and F criteria, which do not depend on the observed labels, the constructed MICP model is extremely efficient with global unconfidence and fuzziness scores of 0.066 ± 0.008. For all ϵ-free criteria (except S), the MICP model achieves higher efficiency for synergistic drug-drug combinations, which is desirable for predictive *in silico* modelling as synergistic combination therapies offer the possibility of enhanced pharmacological efficacies (Narayan et al., 2020), with reduced effective doses and associated host toxicities (Jia et al., 2009).

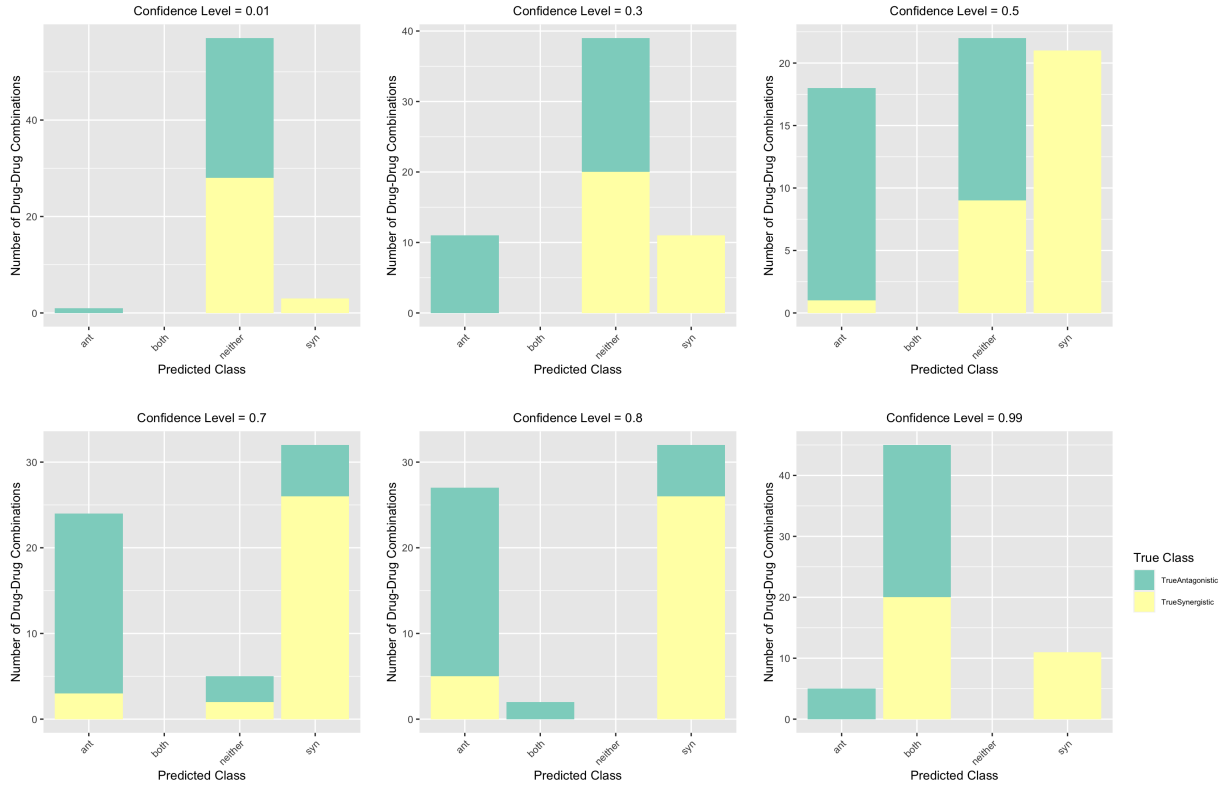

Figure 3. Distribution of class labels included in MICP prediction sets with increasing confidence $(1 - \epsilon)$.

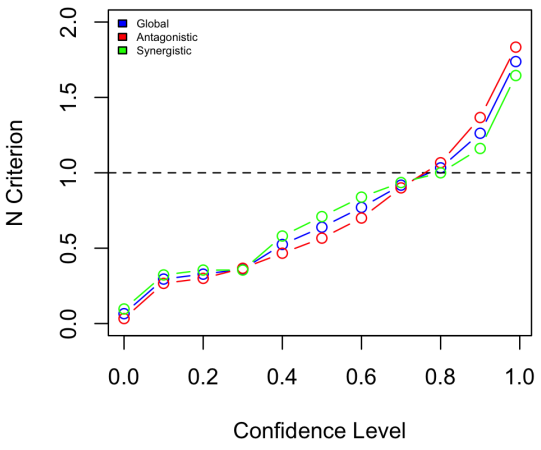

Figure 4. Conformal prediction efficiency according to the prior ϵ-dependent *N* criterion.

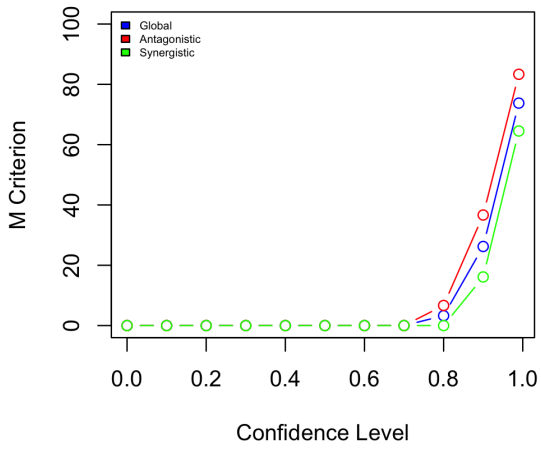

Figure 5. Conformal prediction efficiency according to the prior ϵ-dependent *M* criterion.

3.2.2. ϵ-DEPENDENT CONFORMAL EFFICIENCIES

Influence of the user-defined confidence level $(1 - \epsilon)$ on the drug-drug synergy class labels included in MICP prediction sets is shown in Figure 3. The number of drug-drug combinations in the test set predicted to belong to both synergistic and antagonistic classes increases with the confidence level used as a threshold for the class-wise p^y-values generated by comparing the non-conformity score of each test object with the Mondrian class lists constructed from the calibration set. It can be seen that there is a trade-off between the confidence level and singleton prediction rate, as previously documented in the literature on MICPs (Cortés-Ciriano &

Bender, 2019). The confidence level is negatively correlated with the number of null-class predictions; conversely, the confidence level is positively correlated with the number of dual-class predictions. Increasing the confidence level (or equivalently, decreasing the significance level) is not always the better choice as this could lead to the assignment of dual-class predictions for drug-drug combinations that were assigned singleton predictions at lower confidence levels. The size of the prediction region indicates the model's uncertainty: as the confidence level increases, or the significance level (ϵ) decreases, the conformal predictor outputs increasingly larger prediction sets to ensure the specified coverage.

Influence of the user-defined confidence level ($1-\epsilon$) on the ϵ-dependent conformal efficiency scores of the MICP model is shown in Figures 4-6. Figure 4 illustrates this ϵ-dependence according to the prior *N* criterion, which measures the average number of labels contained in prediction sets across test objects. At a confidence level of 0.8 (or $\epsilon = 0.2$), nearly all prediction sets contain a single label. A confidence level of 0.9 (or $\epsilon = 0.1$) offers increased confidence with only a modest increase in *N* score, as most prediction sets contain a single label. This trend holds globally, as well as locally for each drug-drug synergy class. Figure 5 illustrates the ϵ-dependence of conformal efficiency according to the prior *M* criterion, which measures the percentage of prediction sets containing multiple labels. The percentage of non-singleton predictions rises steeply after $1 - \epsilon \geq 0.8$ for antagonistic drug-drug combinations. For synergistic drug-drug combinations, which are of particular interest to the drug-drug synergy modelling community, the percentage of non-singleton predictions rises steeply after $1 - \epsilon \geq 0.9$. Figure 6 illustrates the ϵ-dependence of efficiency according to the observed *OM* criterion, which measures the percentage of prediction sets containing an incorrect label. While the *N* and *M* scores show that most MICP prediction sets for synergistic drug-drug combinations are singletons at $1 - \epsilon = 0.9$, roughly 20% of these predictions are incorrect.

3.3. Inter-Species Conformal Transferability

To evaluate the transferability of Mondrian inductive conformal predictors across bacterial strains, we compared the efficiencies of MICPs trained on one bacterial strain and tested on another at a significance level of $\epsilon = 0.2$, which represented a suitable trade-off between the confidence level and singleton prediction rate (see Figure 3), and is commonly used as the user-defined threshold in literature (Svensson et al., 2018). Figures 7 and 8 show the results of the various intra-species and inter-species permutations investigated against ϵ-free and ϵ-dependent criteria, respectively. MICPs trained on *Escherichia coli* and tested on *Salmonella typhimurium* show robust efficiency against all criteria, comparable to MICPs trained and tested on *Escherichia coli*.

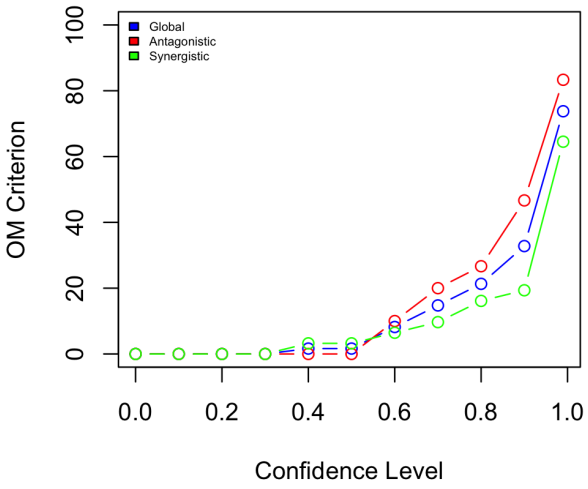

Figure 6. Conformal prediction efficiency according to the observed ϵ-dependent *OM* criterion.

This trend holds in the reverse direction for MICPs trained on *Salmonella typhimurium* and tested on *Escherichia coli*, with comparable efficiencies against all criteria for conformal efficiency. MICPs tested on *Pseudomonas aeruginosa* exhibited lower efficiencies against the ϵ-free criteria for unconfidence (*U*) or, equivalently, fuzziness (*F*), as well as the ϵ-dependent criteria for multiple (*M*) or excess (*E*) predictions, for both inter-species and intra-species experiments, indicating that this dataset may be particularly challenging to model (Majha et al., 2024). Overall, these findings corroborate the monochromatic bacterial group-group interactions reported by Lv et al. (2023a).

4. Conclusion

In conclusion, we propose a principled approach to uncertainty quantification in machine learning for drug-drug synergy. By providing statistically valid prediction regions at user-defined, tolerated error rates, Mondrian inductive conformal predictors enhance the interpretability and reliability of computational drug-drug synergy models, with observed unconfidence and fuzziness scores of 0.13 ± 0.02, and inter-species transferability at comparable conformal efficiencies. We report that MICPs are highly efficient at quantifying prediction uncertainty for synergistic drug-drug combinations according to both prior and observed criteria for ϵ-free and ϵ-dependent efficiencies. Conformal prediction can be integrated flexibly as a wrapper with any underlying machine learning model for drug-drug synergy at low computational cost.

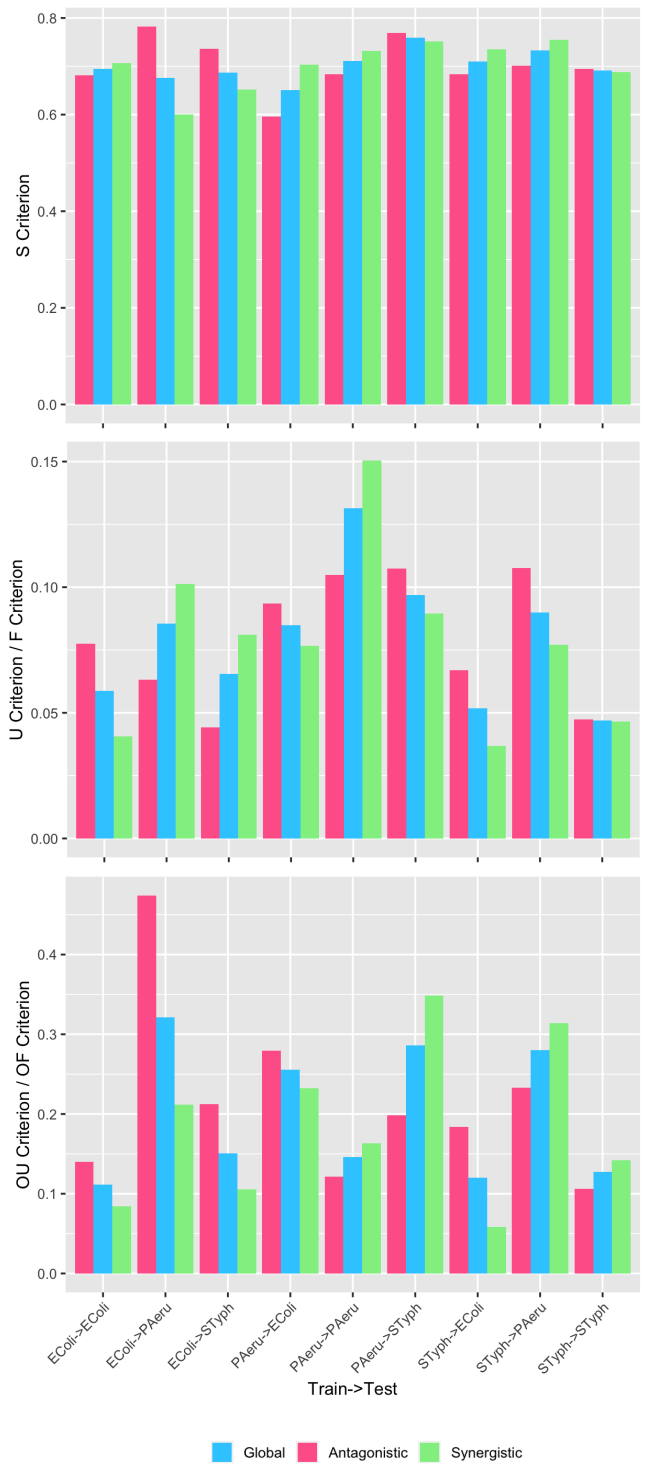

Figure 7. Intra-species and inter-species conformal prediction efficiencies against ϵ-free criteria. Upper panel: S scores; middle panel: U and F scores; lower panel: OU and OF scores.

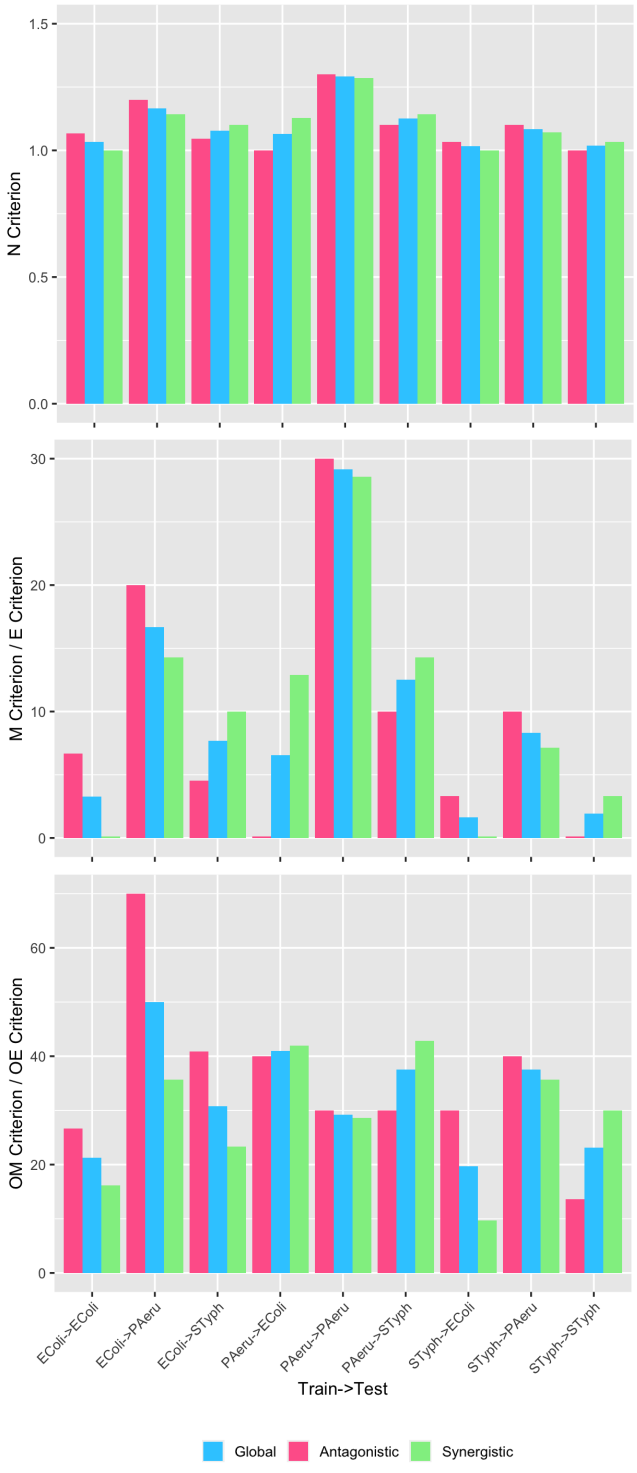

Figure 8. Intra-species and inter-species conformal prediction efficiencies against ϵ-dependent criteria at $\epsilon = 0.2$. Upper panel: N scores; middle panel: M and E scores; lower panel: OM and OE scores.

References

Bliss, C. I. The Toxicity of Poisons Applied Jointly. *Annals of Applied Biology*, 26(3):585–615, 1939.

Breiman, L. Random Forests. *Machine Learning*, 45:5–32, 2001.

Cortés-Ciriano, I. and Bender, A. Concepts and applications of conformal prediction in computational drug discovery. *arXiv:1908.03569*, 2019.

Cortés-Ciriano, I., Bender, A., and Malliavin, T. Prediction of PARP inhibition with proteochemometric modelling and conformal prediction. *Molecular Informatics*, 34 (6-7):357–366, 2015.

Devetyarov, D. and Nouretdinov, I. Prediction with confidence based on a random forest classifier. In *Artificial Intelligence Applications and Innovations: 6th IFIP WG 12.5 International Conference, AIAI 2010, Larnaca, Cyprus, October 6-7, 2010. Proceedings 6*, pp. 37–44. Springer, 2010.

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

Kuhn, M. Caret: classification and regression training. *Astrophysics Source Code Library*, pp. ascl–1505, 2015.

Lipinski, C. A. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.*, 46:3–26, 2001.

Lipinski, C. A. Lead-and drug-like compounds: the rule-of-five revolution. *Drug Discovery Today: Technologies*, 1 (4):337–341, 2004.

Lv, J., Liu, G., Ju, Y., Huang, H., Li, D., and Sun, Y. Identification of robust antibiotic subgroups by integrating multispecies drug–drug interactions. *Journal of Chemical Information and Modeling*, 63(15):4970–4978, 2023a.

Lv, J., Liu, G., Ju, Y., Huang, H., and Sun, Y. AADB: a manually collected database for combinations of antibiotics with adjuvants. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2023b.

Majha, A. G., Stott, I., and Bender, A. On Modelability and Generalizability: Are Machine Learning Models for Drug Synergy Exploiting Artefacts and Biases in Available Data? In *Machine Learning in Computational Biology*, pp. 123–134. PMLR, 2024.

Narayan, R. S., Molenaar, P., Teng, J., Cornelissen, F. M., Roelofs, I., Menezes, R., Dik, R., Lagerweij, T., Broersma, Y., Petersen, N., et al. A cancer drug atlas enables synergistic targeting of independent drug vulnerabilities. *Nature Communications*, 11(1):2935, 2020.

Norinder, U., Carlsson, L., Boyer, S., and Eklund, M. Introducing conformal prediction in predictive modeling. a transparent and flexible alternative to applicability domain determination. *Journal of Chemical Information and Modeling*, 54(6):1596–1603, 2014.

Papadopoulos, H. Inductive conformal prediction: Theory and application to neural networks. In *Tools in Artificial Intelligence*. Citeseer, 2008.

Svensson, F., Aniceto, N., Norinder, U., Cortes-Ciriano, I., Spjuth, O., Carlsson, L., and Bender, A. Conformal regression for quantitative structure–activity relationship modeling—quantifying prediction uncertainty. *Journal of Chemical Information and Modeling*, 58(5):1132–1140, 2018.

Vovk, V. Conditional validity of inductive conformal predictors. In *Asian Conference on Machine Learning*, pp. 475–490. PMLR, 2012.

Vovk, V., Gammerman, A., and Shafer, G. *Algorithmic Learning in a Random World*, volume 29. Springer, 2005.

Vovk, V., Fedorova, V., Nouretdinov, I., and Gammerman, A. Criteria of efficiency for conformal prediction. In *Conformal and Probabilistic Prediction with Applications: 5th International Symposium, COPA 2016, Madrid, Spain, April 20-22, 2016, Proceedings 5*, pp. 23–39. Springer, 2016.

Appendix

The *E criterion* (where "E" stands for "excess") is an ϵ-dependent, prior criterion for conformal efficiency and is defined as the average number of labels exceeding 1 in prediction sets across all *k* test instances:

$$E = \frac{1}{k}\sum_{i=1}^{k}(|\Gamma_i^\epsilon| - 1)^+ \tag{11}$$

where $t^+ := \max(t, 0)$.

The *OE criterion* (where "OE" stands for "observed excess") is an ϵ-dependent, observed criterion for conformal efficiency and is defined as the average number of false labels included in prediction sets across all *k* test instances:

$$OE = \frac{1}{k}\sum_{i=1}^{k}|\Gamma_i^\epsilon \setminus \{y_i\}| \tag{12}$$