Conformal prediction of small-molecule drug resistance in cancer cell lines

Saiveth Hernández-Hernández SAIVETH.HERNADEZ@INSERM.FR
Sachin Vishwakarma SACHIN.VISHWAKARMA@INSERM.FR
Cancer Research Center of Marseille (INSERM U1068, Institut Paoli-Calmettes, Aix-Marseille Université UM105, CNRS UMR7258), Marseille, France

Pedro J. Ballester PEDRO.BALLESTER@INSERM.FR
Department of Bioengineering, Imperial College London, London SW7 2AZ, UK.
Cancer Research Center of Marseille (INSERM U1068, Institut Paoli-Calmettes, Aix-Marseille Université UM105, CNRS UMR7258), Marseille, France

Editor: Ulf Johansson, Henrik Boström, Khuong An Nguyen, Zhiyuan Luo and Lars Carlsson

Abstract

Drug design is a critical step in the drug discovery process, where promising drug molecules are engineered to be later evaluated preclinically and perhaps clinically. Phenotypic drug design has again gained traction. Cancer cell lines, a frequently adopted *in vitro* model for phenotype drug design, can be used to evaluate the drug resistance level (lack of inhibitory activity, for example) of a large number of molecules, and discard those that are the least likely to become drug candidates. By reusing these datasets, supervised learning models have been built to predict drug resistance on cancer cell lines. Usually, these methods have assigned reliability to the whole model rather than reliability to individual predictions (molecules). In problems such as drug design, accurately achieving the latter would revolutionize decision making. Conformal prediction is a model-agnostic method to assign reliability to each model prediction. In this study, we investigated the impact of conformal prediction on the prediction of inhibitory activity of molecules on a given cancer cell line. This analysis was carried out in each of the 60 cell lines from the NCI-60 panel to understand the variability of the results across cancer types. We also discussed the implications of predicting the molecules considered most potent. In addition, we investigated how the further subdivision of the training set to build conformal prediction models may affect the results obtained. Overall, we observed that those molecules deemed most reliable by conformal prediction are substantially better predicted than those that are not. This suggest that such computational tools are promising to guide phenotypic drug design.

Keywords: Conformal Prediction, phenotypic drug design, supervised learning

1. Introduction

Drug discovery is a process that typically requires the identification of a small molecule with the potential to become a drug candidate (Hughes et al., 2011). The two main drug discovery strategies are phenotypic and target-based (Childers et al., 2020). In target-based drug discovery (TDD), the starting point is a defined molecular target that is hypothesized to have an important role in the considered disease. In contrast, phenotypic drug discovery (PDD) does not rely on knowledge about a specific drug target or a hypothesis about its role in the disease (Moffat et al., 2017). PDD evaluates observable phenotypic changes in a cell that can then be used to identify small molecules (Szabo et al., 2017). Preclinical models

such as cancer cell lines are most commonly used to study the efficacy of drugs, and has been a frequently-adopted approach for PDD screens. Cell lines are tested against drugs to assess the sensitivity/resistance determined by half-maximal inhibitory concentration (IC_{50}) value (Piyawajanusorn et al., 2021). Drug sensitivity profiling utilizing cancer cell lines is routinely performed and thus large-scale datasets are publicly available. Despite TDD being the predominant approach for the past 30 years, the majority of first-in-class drugs identified from 1999 to 2008 had originated from PDD approaches. Over the past few years, PDD has thus gained traction (Childers et al., 2020).

In parallel, computer-aided drug design has grown steadily since the late 1960s (Merz Jr et al., 2010). Many supervised learning algorithms are now employed for predictive modeling of phenotypic activities of molecules in cancer cell lines (Ballester et al., 2022). However, these predictive models usually assign a reliability to the whole model (e.g. by calculating the RMSE between predicted and observed activities of test set molecules), rather than a reliability at the instance level (e.g. a predicted activity interval where the observed activity of a given test set molecule is most likely to be). In problems such as drug design, estimating this reliability is important for decision making. For instance, to select the molecules that are not only predicted to be most potent but also those with the most reliable predictions, so as to reduce time and financial costs. Conformal Prediction, CP for short, is a mathematical framework to model the reliability of predictions in diverse tasks. The idea behind CP is that a new instance is predicted with a label that makes it similar to the old instances in some specific way. The degree to which the specific type of similarity holds within the old instances is used to estimate the confidence in the prediction (Vovk et al., 2005). CP have successfully applied to a range of drug design problems (Norinder et al., 2014; Eklund et al., 2015; Cortés-Ciriano et al., 2016; Bosc et al., 2019; Alvarsson et al., 2021).

In this study, we investigated if CP can enhance the prediction of the inhibitory activity of molecules on a given cancer cell line. We tested our hypothesis on 60 cell lines from the NCI-60 panel by building one model per cell line. Class-imbalance data have been rarely explored in the context of regression problems, where the minority class instances are the most valuable data instances. In this study, we investigate whether CP generates robust predictions in molecules with submicromolar potency (these molecules constitute a minority class in the NCI-60 data). We also look at how different training data partitions impact CP performance at this task.

2. Experimental design

2.1. Dataset

We modelled the pGI_{50} of molecule-cell line pairs, defined as the negative logarithm of the half-maximal inhibitory concentration of the molecule on the cell line. We used data with such measures of activities for 50,555 compounds on 60 cancer cell lines from the National Cancer Institute (NCI-60) data. These 60 cell lines comprises 9 cancer types: leukemia, melanoma, non-small-cell lung, colon, central nervous system, ovarian, renal, prostate, and breast. Each molecule submitted to the NCI-60 for testing and evaluation is identified with a unique registration number called the National Service Center (NSC) ID.

The preprocessing of the NCI-60 dataset had two main stages: data cleaning and data representation. In the data cleaning stage, we remove low-activity molecules ($pGI_{50} <$

4) as they do not have therapeutic potential. Some NSC–cell line pairs were tested at different concentrations, resulting in multiple pGI_{50} measurements. Thus, we calculated the mean when more than one pGI_{50} measurements were available for the same NSC-cell line pair. In the data representation stage, chemical structures were curated from the Chem2D_Jun2016.sdf file using the openbabel library. The salts were removed from the chemical structure and SMILES were standardized using the Molvs package. Using the SMILES of a molecule, we generated its Morgan circular fingerprint (Rogers and Hahn, 2010) with radius 2 and 256 bits. This fingerprint size works well on similar phenotypic drug design problems (Sidorov et al., 2019). Therefore, each molecule is represented by a 256-bit Morgan fingerprint that corresponds to the presence or absence of a particular substructure in the molecule. These will be the features employed to build each of the 60 models, one per cell line.

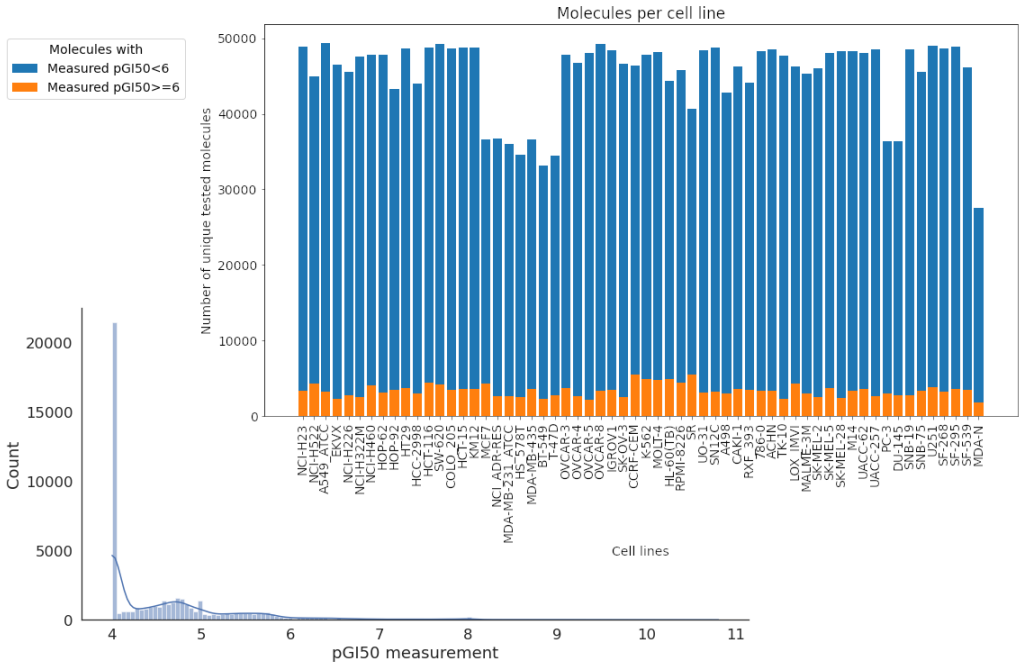


Figure 1: **Each cell line has abundant data, although potent molecules are rather scarce**. Distribution of pGI_{50} measurements in the 50,846 unique NSC IDs (bottom). Distribution of the number of unique molecules tested per cell line (top). The most potent molecules ($pGI_{50} \geq 6$) for each cell line are in orange color.

These preprocessing steps led to 50,846 unique NSC IDs and 2,707,434 pGI_{50} measurements that correspond to 50,555 unique molecules. Figure 1 shows NCI-60 data after the preprocessing stage. Here, the pGI_{50} values are represented by a right (or positive) skewed distribution with a long tail form by the most potent molecules ($pGI_{50} \geq 6$ or molecules with submicromolar potency). The peak at pGI_{50}=4 is originated because if the inhibitory activity is not reached or is exceeded, the pGI_{50} value is expressed as greater or less than

the maximum or minimum concentration tested (NCI-60 screening methodology). This study includes molecules with pGI_{50}=4 as preliminary results showed better performance in model evaluation when these molecules were present in the data set.

2.2. Models

We employed the Python implementations of the random forest or RF (Pedregosa et al., 2011) and the extreme gradient boosting (XGB) algorithms to build a regression model for each cell line. RF is a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest (Breiman, 2001). The term "Gradient Boosting" in XGB originates from the paper (Friedman, 2001). XGB is an implementation of gradient boosted decision trees designed for speed and performance (Sheridan et al., 2016).

Hyperparameter tuning is a way to enhance the performance of machine learning (ML) algorithms. Previous studies have found sets of hyperparameters that work well in similar problems, using RF or XGB. For example, Svetnik et al. (2003) found that in RF, the number of trees in the forest ($n\_estimator$) and the number of features to consider when looking for the best split ($max\_features$) are the most important hyperparameters to tune. For XGB, Sheridan et al. (2016) found that the most important hyperparameters to tune are the number of gradient boosted trees ($n\_estimator$), the maximum tree depth for base learners (max_depth), and the subsample ratio of columns when constructing each tree ($cosample\_bytree$). To tune these hyperparameters, a grid search was carried out on each of the 60 training sets to find their best values for predicting the inhibitory activity of molecules on that cancer cell line. To do this, we used an 80/10/10 scheme. This means that 80% of the data was used as a training set, 10% as a validation set, and 10% as a test set. After identifying the best values for these hyperparameters, the algorithm used them to re-train the model on 90% of the data. Therefore, the test set was not used in any way to train or select the corresponding underlying model (the same is true for the error model, and thus, CP models). The hyperparameter values used for the 60 cell lines were selected based on the value at which most cell lines achieved their best performance. The best hyperparameter values for RF were $n\_estimator$=1000 and $max\_features$=0.4; and for XGB were $n\_estimator$=1000, max_depth=9 and $cosample\_bytree$=0.4.

2.3. Conformal prediction

For each of these ML regression problems, we consider the example space defined as $Z \equiv X \times Y$. The elements in Z are usually expressed as $z_i = (x_i, y_i)$, where x_i is a vector of features and y_i is the real valued variable to predict (Vovk et al., 2005). Note that in this basic scenario, the prediction of y_i, $h(x_i) = \hat{y}_i$, is usually given without a degree of confidence. To enhance this, CP provides reliability for a single instance by predicting an interval for y_i instead of a point prediction. The standard ML model $h(x)$ is called the underlying model in CP.

The inductive CP (ICP) framework – a subtype of CP – requires a user-specified confidence level, which refers to the minimum fraction of predictions whose true value will lie within the predicted confidence interval. A calibration function, which is used to assess to which extent a new instance conforms to the data that the model has been built upon.

In particular, a calibration function can be defined as a non-conformity function, in which case, this function measures how different a new example is from the examples on which the model was built. If a conformal predictor is well calibrated, the error rate $\epsilon \in [0, 1]$ should not be larger than $\epsilon = 1-$ confidence level (Balasubramanian et al., 2014).

In the standard ICP for regression problems, the interval size is given by the non-conformity score α_s obtained at the specified confidence level. Here, s is the index of the $(1-\epsilon)^{th}$ percentile nonconformity score. If the cardinality of the calibration set is q, then $s = \lfloor (1-\epsilon)q \rfloor$. For example, if the list of 10 sorted $\alpha-$values is 0.1, 1.2, ..., 8.5, 8.8, 9.3, the non-conformity score at a confidence level of 90%, i.e $\epsilon = 0.1$, is the value at the index $s = 9$, that is, $\alpha_9 = 8.8$. Therefore, intervals of the same size are obtained for each prediction, which is a suboptimal approximation. The normalized ICP is a variant that includes the error model to control the size of the predicted intervals (Norinder et al., 2014). Mathematically, we get a second example space $Z_e \equiv X \times E$, where E is the set of errors. Thus, a second supervised learning model, $g(x)$, is trained on the same features, but to predict the error made by the $h(x)$ model rather than the actual y_i. The error model $g(x)$ is trained on the same training set as $h(x)$ but to predict the error in the calibration set and in the test set. The information of the error model is then included in the non-conformity function, which is evaluated for the z_i instance giving rise to its non-conformity score,

$$\alpha^{z_i} = \frac{|y_i - h(x_i)|}{g(x_i) + \delta} \tag{1}$$

and in the predicted interval (Γ_i^ϵ) for the z_i instance:

$$\Gamma_i^\epsilon = h(x_i) \pm \alpha_s(g(x_i) + \delta) \tag{2}$$

where δ is a user-supplied sensitivity parameter to control the ratio of the observed error to the predicted error.

To measure the quality of the predicted intervals, CP introduces the validity and efficiency metrics for model evaluation. Validity measures how reliable the predictions are, and efficiency quantifies how specific these are. For example, validity implies that for a confidence level of 90% (0.9), the predictor will include the true y_i value within its prediction intervals $\Gamma_i^{0.1}$ in at least 90% of all predictions if CP is well calibrated. Efficiency is how small the predicted interval is. Usually, requesting a higher confidence level results in reduced efficiency, i.e. larger predicted intervals (Vovk et al., 2005; Balasubramanian et al., 2014).

2.4. Model building and evaluation

We built one model per cell line using the molecular features of the molecules as features and the measured pGI_{50}s on the cell line as the real-valued variable to predict. A 90-10 random partition was applied to each of the 60 cell lines, where 10% of the molecules tested on that cell line are kept aside as the test set. The remaining 90% are used as the training set and the training set was further randomly subdivided into the proper training set and the calibration set, as required for ICP. The purpose of the calibration set is to estimate the confidence, through the non-conformity function, in the new predictions based on the previous predictions. To explore the impact of different training data partitions on each

regression problem (i.e. cell line), we employed three training set partitions: 70-30, 80-20, and 90-10.

The normalized ICP for regression, implemented in the NonConformist python package, was used to run all the experiments. NonConformist sets the k-nearest neighbors (kNN) algorithm as the default error model, so we evaluated this option in addition to the RF and XGB algorithms. In this package, the error target is defined as:

$$error = \log(|y_i - h(x_i)| + 0.00001) \tag{3}$$

where a small value (0.00001) was added to the $|y_i - h(x_i)|$ term to avoid the singularity $y_i = h(x_i)$. We kept the default sensitivity parameter value ($\delta = 0$), so the predicted interval in Equation (2) is now more directly computed as $\Gamma_i^\epsilon = h(x_i) \pm \alpha_s g(x_i)$. For the underlying model, we employed either the RF algorithm or the XGB algorithm. Table 1 shows the combinations of the considered CP models, each of them evaluated at four confidence levels: 80%, 85%, 90%, and 95%. A total of 1080 models (60 cell lines $\times$ 6 CP models $\times$ 3 training data partitions) were built and evaluated in this study.

Table 1: CP models trained. Each CP model is specified by its $h(x)$-$g(x)$ combination

Underlying model h(x)	**Error model g(x)**		
	kNN	**RF**	**XGB**
RF	RF-kNN	RF-RF	RF-XGB
XGB	XGB-kNN	XGB-RF	XGB-XGB

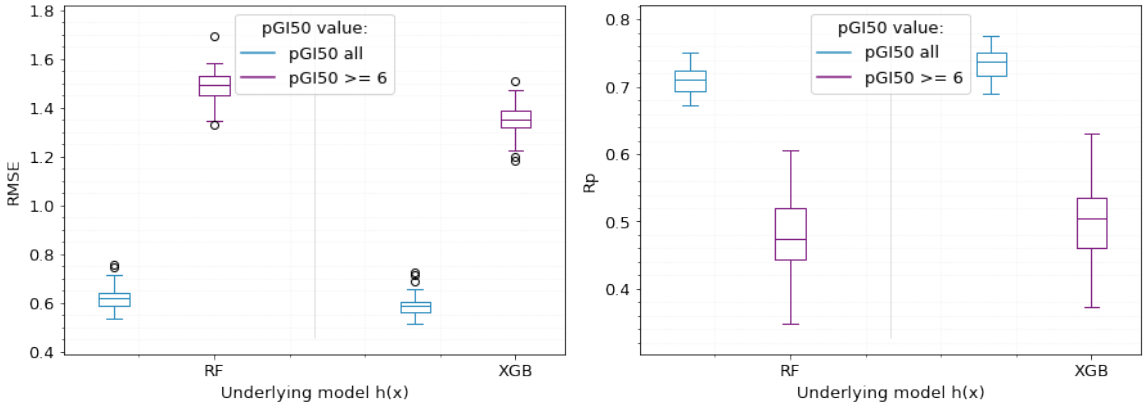
The performance of the regression models was evaluated for the 60 test sets. Each CP model in Table 1 was evaluated using the validity and efficiency. The underlying model $h(x)$ and the error model $g(x)$ are standard supervised learning models (Subsection 2.3), for which we compute the root mean square error (RMSE) and the Pearson correlation coefficient (Rp). A lower RMSE (error close to 0) and a higher Rp (correlation close to 1) are hence better values.

3. Results and Discussion

3.1. Underlying models

For each cell line, we employed the corresponding trained RF and XGB models to predict the pGI_{50} value of a given test set molecule from its molecular features. Then, we computed the RMSE and Rp metrics for each of the 60 sets (one per cell line). These metrics were additionally computed for the most potent test molecules (those with measured $pGI_{50} \geq 6$, or submicromolar potency). Since the underlying model $h(x)$ is a regression model, i.e. without CP, confidence levels are not applicable at this stage.

Figure 2 shows RMSE and Rp distributions across the 60 test sets, for either RF or XGB models. All models are well above the random Rp level of zero. The XGB model provides slightly better performance (median RMSE=0.58 and median Rp=0.74) than the RF model (median RMSE=0.62 and median Rp=0.71). This trend remains the same when we look



(*a*) RMSE (*b*) Rp

Figure 2: **The underlying models can predict the pGI_{50} of test set molecules, although this is worse for potent molecules.** The boxplots summarize the (a) RMSE and (b) Rp distributions across 60 test sets (one per cell line). RMSE and Rp values were computed between the observed and predicted pGI_{50} values using either RF or XGB models. Color code refers to all molecules (pGI_{50} all) or most potent molecules ($pGI_{50} \geq 6$) in the test sets.

at the subset of test sets containing only the most potent molecules (XGB with median RMSE=1.35 and median Rp=0.5; RF with median RMSE=1.5 and median Rp=0.48). These results may be influenced by the few examples of most potent molecules ($pGI_{50} \geq 6$) present in each cell line of the NCI-60 panel (Figure 1), which may make these types of molecules more difficult to predict.

We are now looking for the best/worst predicted cell line in terms of RMSE, i.e. the cell line with the lowest/largest error between the observed and predicted pGI_{50} value. Table 2 shows the best and worst predicted cell line (from top to bottom) for each underlying model. For the RF model, the best predicted cell line is OVCAR-5, an ovarian carcinoma cell line. For the XGB model, the best predicted cell line is SNB-19, a central nervous system cancer cell line. The worst predicted cell line, in both models, is the SR, a leukemia cancer cell line. We will discuss these results in Subsection 3.2.

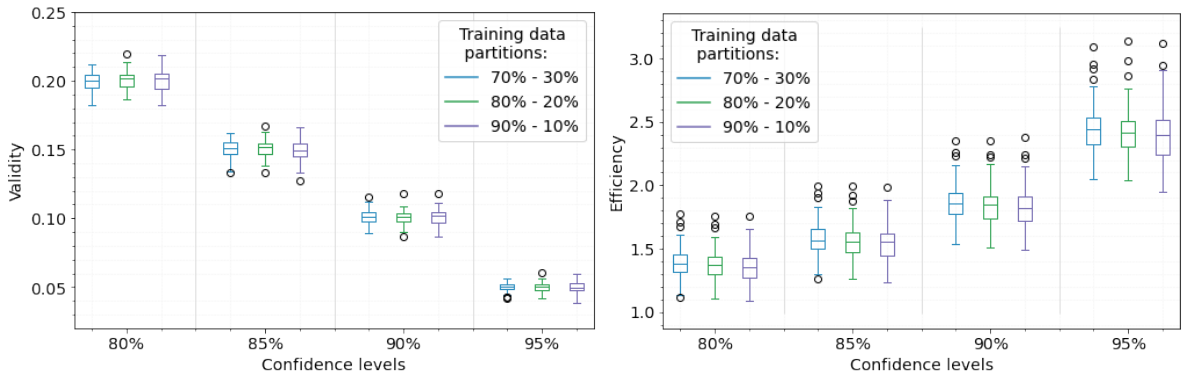
3.2. Conformal prediction

We computed the validity and efficiency for the six CP models shown in Table 1. The validity is computed as the average error rate. This is the fraction of molecules whose observed pGI_{50} values lie outside the predicted interval. The efficiency is computed as the average prediction interval size of the test set molecules.

3.2.1. Training data partitions

In supervised learning, the size of the training and test set is likely to have an impact on the output predictions (Singh et al., 2021; Rácz et al., 2021). We investigated whether this holds true in the training set when ICP is used in the context of this problem. That is, we investigated if the further subdivision of the training set, into the proper training set and calibration set, has an impact in terms of validity and efficiency in the predictions made using ICP.

Given the three training data partitions, similar validity and efficiency results were obtained for all CP models described in Table 1. Consequently, as an example, we explain the training data partition results only for the RF-RF model (Figure 3).



(*a*) Validity (*b*) Efficiency

Figure 3: **Different training data partitions have indistinguishable validity and efficiency.** The boxplots summarize the validity and efficiency distributions, at each confidence level, across 60 test sets (one per cell line) using the RF-RF CP model. Color code refers to the proper training and calibration data partitions evaluated.

The median validity, Figure 3(*a*), is similar in all three training data partitions, a trend that is repeated at each confidence level. The median efficiency, Figure 3(*b*), independent of the used training data partition, is also similar at each confidence level. Furthermore, the validity and efficiency values, exemplified for the RF-RF model, show that this model remains consistent with respect to these values independent of the data size used for calibration of the predictions. These results suggest that, at least for these datasets, varying proper training and calibration data partitions do not affect the obtained results. Consequently, the rest of the study employs 90-10 training data partitions without loss of generality. The validity and efficiency results for all CP models will be further discussed in Subsection 3.2.2.

3.2.2. Validity and efficiency

The median validity and efficiency values were computed in the 60 test sets for each CP model (Table 1) at four confidence levels. Figure 4 shows that, at each confidence level,

the median validity is close to the required error in all CP models ($\epsilon \pm 0.002$). Moreover, on average 4 out of 6 CP models are valid (well calibrated), i.e. the error rate ϵ does not exceeded at each confidence level.

However, achieving valid and near-valid models has a cost in terms of worsened efficiency as shown in Figure 4. Here, the efficiency decreases (i.e. the average size of the predicted intervals increases) as we increase the confidence level, which is a usual behaviour in CP. The median efficiency (interval size) in CP models such as RF-kNN (blue markers) or XGB-kNN (red markers) increases rapidly as we increase the confidence level, reaching values that are not informative for pGI_{50} prediction. For example, a median efficiency around 9 pGI_{50} units is obtained by the RF-kNN CP model at a confidence level of 95%. This behavior will be explained later in this subsection. For the remaining CP models, a better efficiency is obtained when RF is used as the underlying model (median efficiency ranges from 1.43 to 2.74 pGI_{50} units in RF-RF and RF-XGB CP models) rather than the XGB model (median efficiency ranges from 1.64 to 5.92 pGI_{50} units in XGB-RF and XGB-XGB CP models).

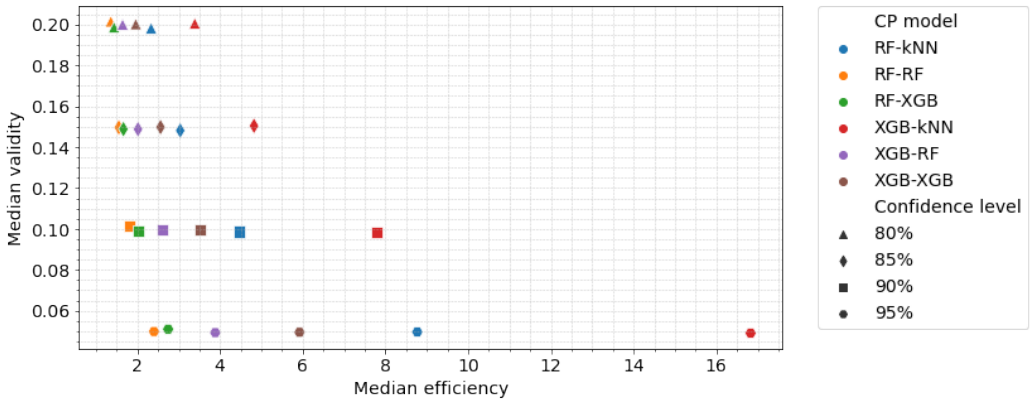
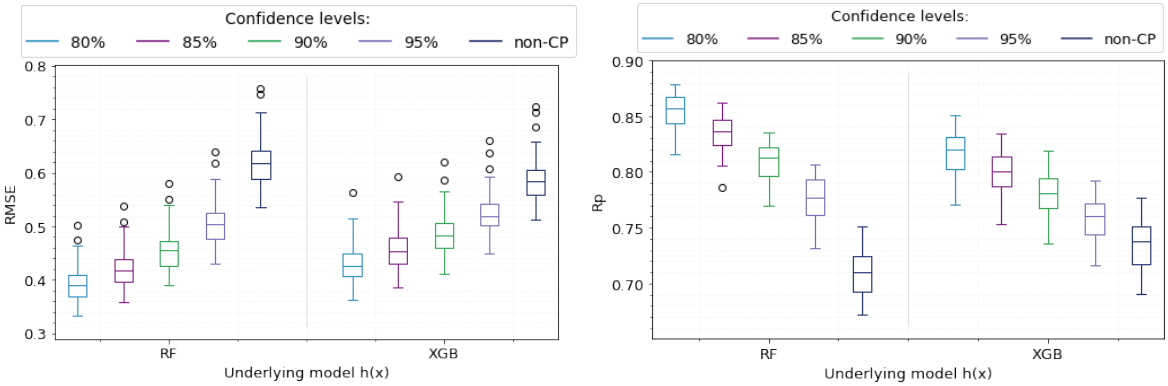


Figure 4: **CP models have substantially different efficiency within a given confidence level.** Median efficiency vs median validity across 60 test sets (one per cell line), at four confidence levels. Color code refers to the six CP models ($h(x)$-$g(x)$) employed. Markers refers to the requested confidence level.

As already observed when we analyzed the efficiency, the confidence level influences the pGI_{50} interval size, becoming more pronounced in the RF-kNN and XBG-kNN CP models. Here, pGI_{50} interval size is given by the multiplication of the non-conformity score and the error predicted. The kNN error model obtains the highest non-conformity scores across 60 test sets (data not shown), which suggests that these predictions are dissimilar compared to the examples in the calibration set. Hence, the RF-kNN and XGB-kNN CP models are less confident about the predictions made. This highlights the importance of choosing an appropriate error model when using the normalized ICP, as well as a confidence level which should be linked to the specific application.

3.2.3. Prediction performance using CP

To quantify whether there is an improvement in prediction performance using CP, we are focusing on the performance of those instances that are valid at a given confidence level. Thus, we call a prediction CP-valid if the observed pGI_{50} value is within the predicted interval. The predictions made by a regression model, without using CP, are called non-CP predictions (non-CP = CP-valid ∪ CP-invalid). To make a holistic comparison between CP-valid predictions and non-CP predictions, we analyzed the RMSE and Rp obtained in the underlying models and in the error models.



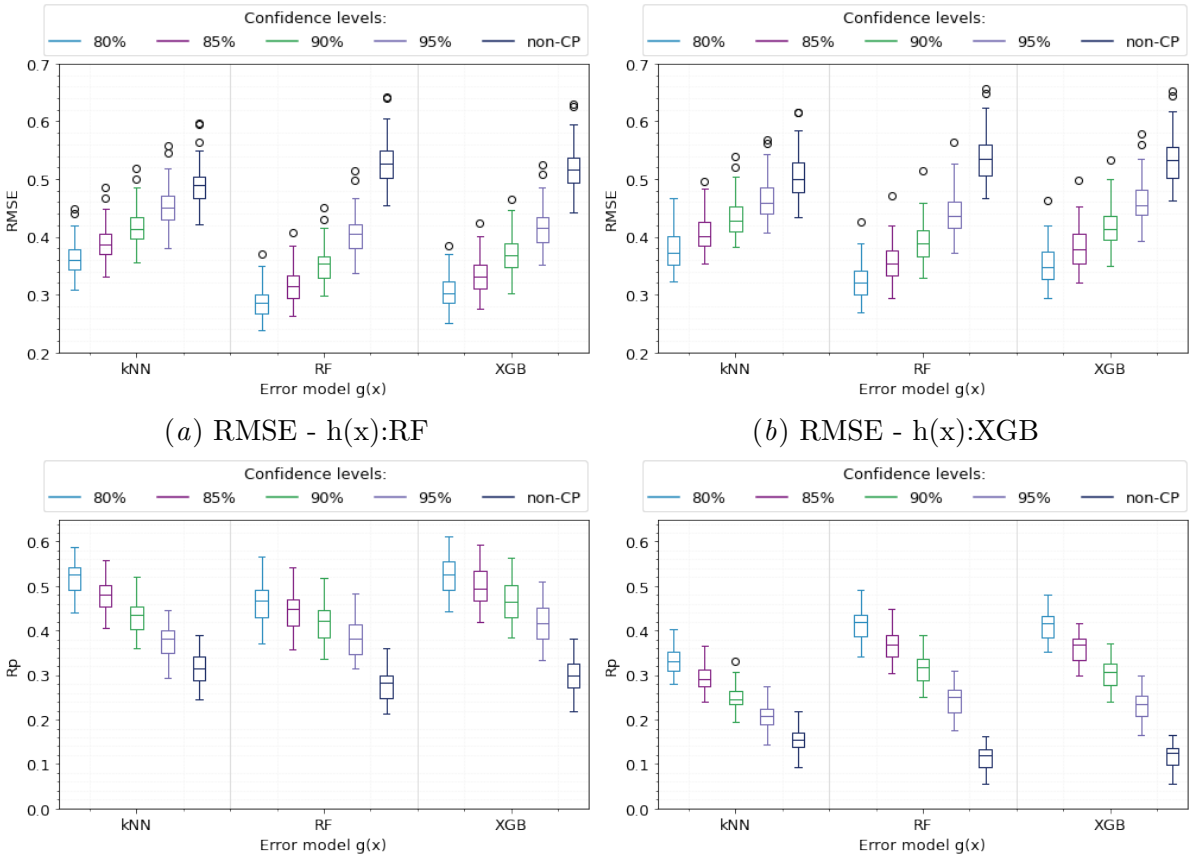
(*a*) RMSE (*b*) Rp

Figure 5: **Relaxing the requested confidence level leads to more accurate CP-valid predictions.** The boxplots summarize the (a) RMSE and (b) Rp distributions across the 60 test sets (one per cell line). RMSE and Rp values were computed between the observed and predicted pGI_{50} value using either RF or XGB as the underlying models. Color code refers to either CP-valid test set molecules, at each confidence level, or non-CP test set molecules.

In addition to the improvement we obtain in predictions when CP is used, results from Figure 5(*a*) suggest that the CP-invalid predictions concentrate a large part of the error obtained. Moreover, at each confidence level, using the RF-based $h(x)$ model leads to a better RMSE distribution across the 60 test sets (one per cell line). Indeed, the median RMSE in the XGB-based $h(x)$ model is approximately 0.02 pGI_{50} units higher than that of the RF model at each confidence level. A similar behavior occurs in terms of Rp, Figure 5(*b*), where we have a better correlation between the observed and predicted pGI_{50} value when RF is used. Here, the median Rp in the XGB model is approximately 0.04 pGI_{50} units lower at each confidence level. Note that the median RMSE for the RF model ranges from 0.39 to 0.5, while the median for the non-CP case is 0.62. For the XGB, the median RMSE ranges from 0.41 to 0.52, while the median for the non-CP case is 0.58.

Figure 6 suggests that in molecules where pGI_{50} error is well predicted, CP-valid predictions are more likely. Overall, we can observe an improvement in terms of RMSE and Rp values, since the best values are obtained with the CP-valid predictions. Moreover,

the RF- and XGB-based $g(x)$ models obtain similar results and show better RMSE and Rp performance, at each confidence level, compared to the kNN-based $g(x)$ model. This suggests that the RF and XGB error models better capture the relationship between the predicted and observed pGI_{50} error.



(*a*) RMSE - h(x):RF

(*b*) RMSE - h(x):XGB

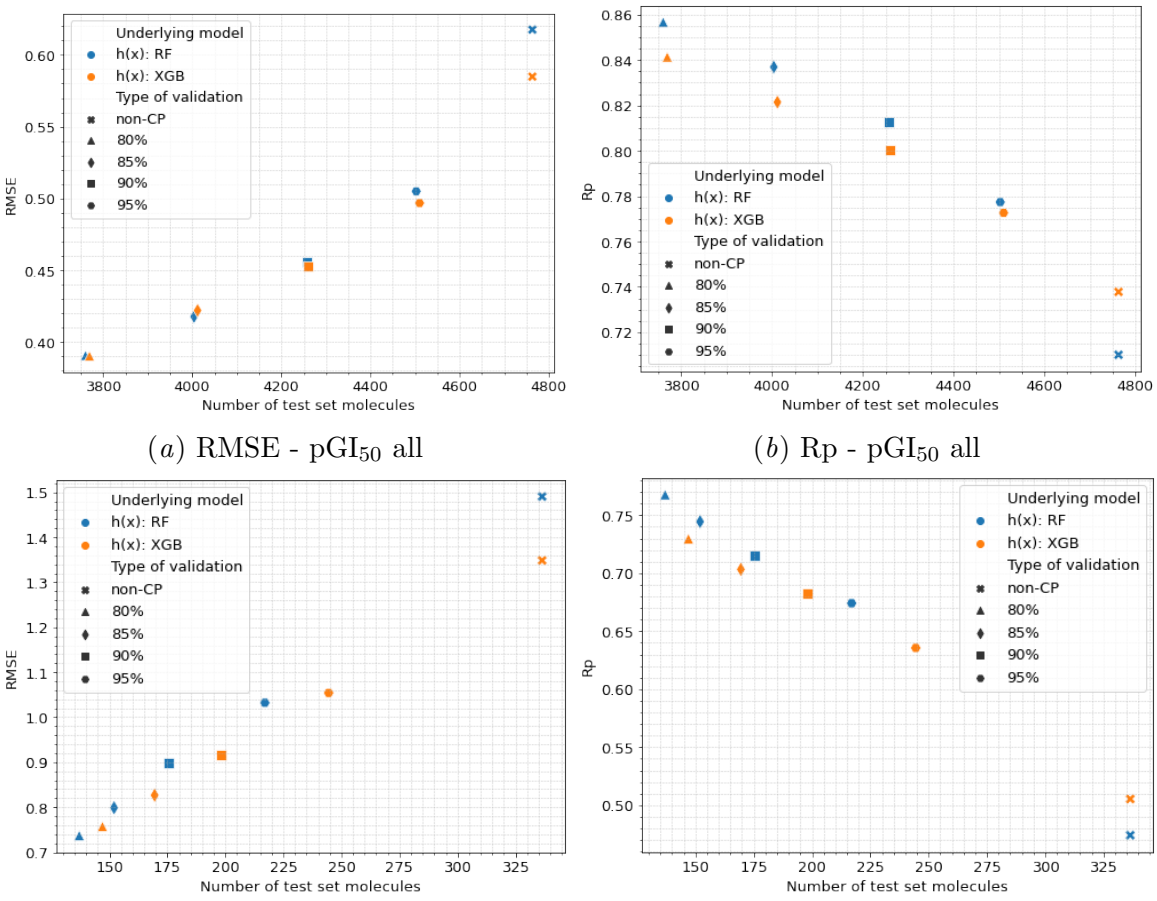
(*c*) Rp - h(x):RF

(*d*) Rp - h(x):XGB

Figure 6: **Molecules that are CP-valid have a better prediction of the pGI_{50} error.** The boxplots summarize the RMSE (top) and Rp (bottom) in predicting pGI_{50} errors across the 60 test sets (one per cell line). RMSE and Rp values were computed between the observed and predicted pGI_{50} error, using either kNN-, RF- or XGB-based $g(x)$ models. Color code refers to either CP-valid test set molecules, at each confidence level, or the non-CP test set molecules.

To analyze the trade-off between the confidence level and the number of CP-valid molecules, we calculated the median RMSE and Rp over the 60 test sets. Figure 7 shows a monotonic improvement of the predictions as the requested confidence level is decreased. While higher confidence levels are related to larger predicted intervals, the choice of the confidence level should be guided by the specific task to be predicted. Here, pGI_{50} predictions at a lower confidence level are more reliable, i.e., the predicted interval is smaller. Note

that the standard prediction based on using the underlying model only, or non-CP, can be regarded as CP with a requested 100% confidence level, which requires an arbitrarily large predicted interval and hence corresponds to a prediction without uncertainty assigned.



(a) RMSE - pGI_{50} all

(b) Rp - pGI_{50} all

(c) RMSE - $pGI_{50} \geq 6$

(d) Rp - $pGI_{50} \geq 6$

Figure 7: **Trade-off between requested confidence level and number of test molecules at that level.** Median RMSE (left) and Rp (right) values in the 60 test sets (one per cell line). Color code refers to either the RF- or XGB-based $h(x)$ model. Markers refers to the type of validation. RF is the error model used in the case of CP-valid predictions. X-axis shows the number of test set molecules without restriction in their pGI_{50} value (pGI_{50} all), and the most potent molecules ($pGI_{50} \geq 6$).

The number of test molecules that are CP-valid predicted and obtained by the RF-RF CP model is slightly lower than in the XGB-RF CP model (Figure 7). This behavior is more evident when dealing with the most potent molecules (Figures 7(c) and 7(d)). This could impact the performance of the models, as the most potent molecules tend to be more

difficult to predict (Figure 2), therefore a larger number of them may impact the RMSE and Rp values obtained. Moreover, the results obtained by the RF-RF CP model are better than those obtained by the XGB-RF CP model.

3.2.4. Best and worst cell line model

In Subsection 3.1 we showed the best/worst predicted cell line in terms of RMSE. Now we looked at which cell lines come out when we apply CP and select them in terms of their best/worst efficiency. Based on previous results, we chose the error model RF, and a confidence level of 80% when CP was used.

Table 2 shows that the RF model keeps the trend of best and worst predicted cell lines with and without CP validation, suggesting that in addition to a small error, the prediction uncertainty (interval size) is small. For the XGB model, the worst predicted cell line is the same with and without CP. However, the best predicted cell line without CP is the SNB-19 cell line, while the best predicted cell line with CP is the NCI-H322M, a non-small cell lung cancer cell line. This suggests that even if we obtain a small error, the uncertainty of the prediction may be large, highlighting the importance of assigning uncertainty to the predictions made.

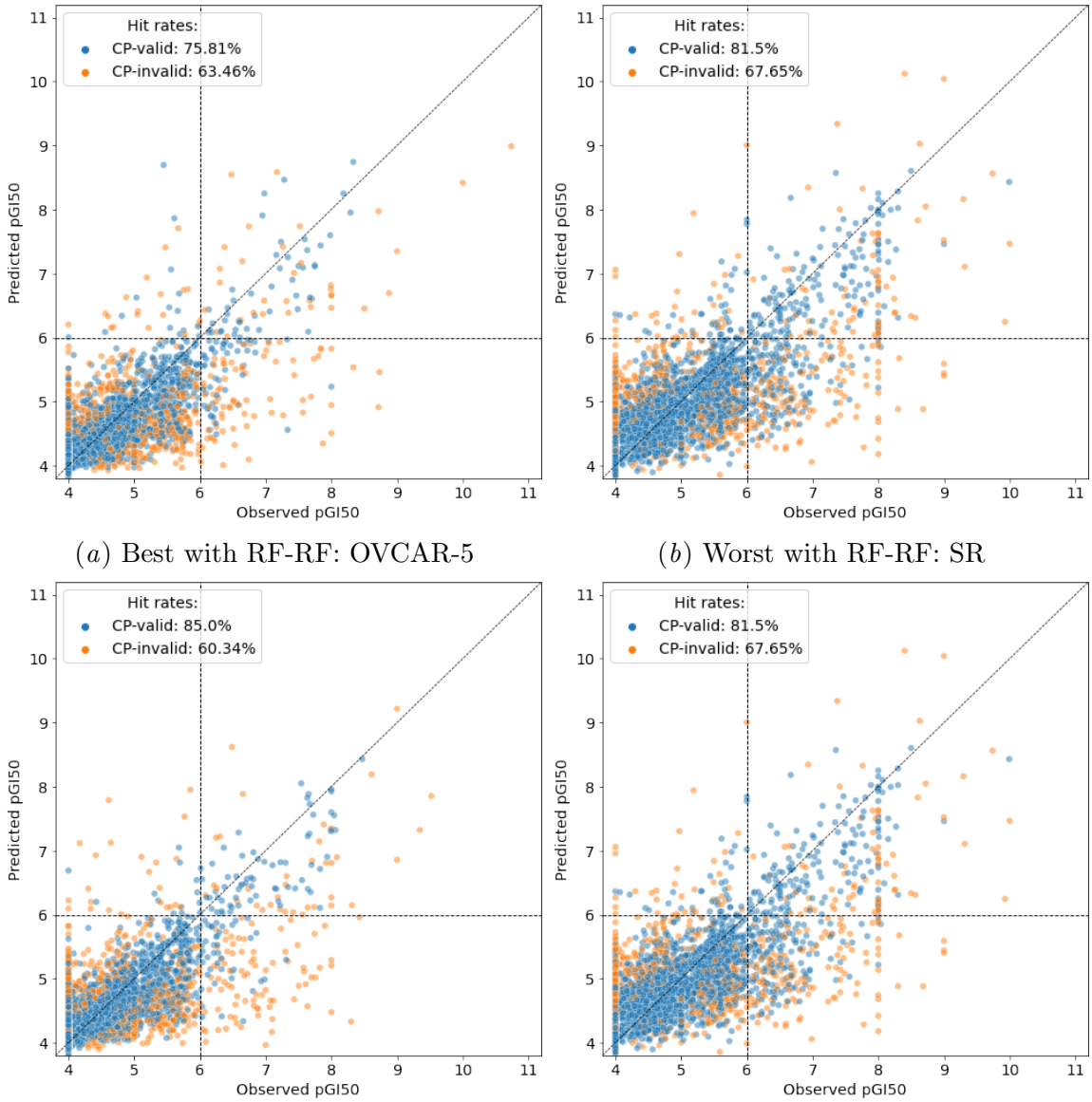
Table 2: Best and worst predicted cell line for each underlying model, for predictions that are non-CP and CP-valid. The confidence level in the CP is 80%. The total column represents the number of test set molecules used to calculate these performance metrics.

Model	Cell	All test set					CP-valid test set		
h(x)	line	RMSE	Rp	Validity	Efficiency	Total	RMSE	Rp	Total
RF									
	OVCAR-5	0.5360	0.6969	0.2102	1.0868	4800	0.3350	0.8443	3791
	SR	0.7575	0.6910	0.1981	1.7550	4064	0.5027	0.8371	3259
XGB									
	SNB-19	0.5125	0.7646	0.1875	1.7448	4852	0.3626	0.8382	3942
	NCI-H322M	0.5328	0.7164	0.2035	1.3503	4757	0.3400	0.8438	3789
	SR	0.7231	0.7137	0.2084	2.1834	4064	0.5147	0.8189	3217

The trend of our results is summarized, as a particular example, in Table 2. That is, an improvement in terms of lower RMSE and higher Rp was obtained using CP, with a cost in the number of CP-valid molecules bounded by 1 - Validity. Indeed, Figure 8 shows that the subset of molecules that are CP-valid predicted (blue dots) have lower errors and higher correlations than those that are not CP-valid predicted (orange dots).

4. Conclusions

- The primary goal of this study was to investigate the improvement introduced by CP when predicting the inhibitory activity of molecules on a given cancer cell line. We



(a) Best with RF-RF: OVCAR-5

(b) Worst with RF-RF: SR

(c) Best with XGB-RF:NCI-H322M

(d) Worst with XGB-RF: SR

Figure 8: **The prediction of the pGI_{50} of molecule-cell line pairs improves when CP is used.** Observed and predicted pGI_{50} value in the best (left) and worst (right) predicted cell line. Color code refers to test set molecules with (blue) and without (orange) CP validation. The vertical and horizontal dotted lines show the threshold for molecules with $pGI_{50} \geq 6$. See Table 2 for performance metrics.

conducted the same analysis on each of the 60 cell lines to understand how results vary across cancer types.

- CP models were better at each selected confidence level, with a cost in terms of worsened efficiency (higher uncertainty associated to the pGI_{50} prediction) at higher confidence levels. This was expected as CP does not alter the predictions of the underlying model in any way. Instead, it anticipates which of these are the most reliable.
- CP models were also better when trying to predict the most potent molecules, which constitutes a minority class within NCI-60 data.
- CP-valid predictions at lower confidence levels are more reliable. However, the choice of the confidence level should be guided by the specific task to be predicted. Here, higher confidence levels needs to be balanced against the uncertainty in the prediction of the pGI_{50} value.
- The results from different training data splits showed that the chosen proper training set and the calibration set split do not affect the efficiency and validity results in each of the 60 test sets.
- CP-valid predictions in each of the 60 test sets have lower errors and higher correlations than those that are non-CP (for each test set, these predictions come from the same underlying model, thus ensuring a fair comparison). Therefore, the CP model should improve hit rates in prospective virtual screening, by not only testing *in vitro* those molecules likely to be potent (predicted $pGI_{50} \geq 6$), but also requesting that are CP-valid.
- We are not aware of any previous study that demonstrated that CP improves the retrieval of molecules with high potency on NCI-60 cell lines (Figure 8). These results strongly suggest that selecting compounds for *in vitro* validation will result in higher hit rates when restricting to those predicted to be CP-valid at the chosen confidence level, rather than the most common approach of merely using the underlying model prediction (non-CP).
- In the future, we plan to investigate the application of CP to other scenarios such as those where test set present a higher proportion of chemotypes not seen on the training set.

Supplementary information

The code and data for the reproduction of the results shown are available at: https://doi.org/10.5281/zenodo.6504995

Acknowledgments

S. Hernández-Hernández acknowledges the National Council of Sciences and Technology of Mexico (CONACYT) for her PhD fellowship. S. Vishwakarma acknowledges the Institut Paoli-Calmettes.

References

Jonathan Alvarsson, Staffan Arvidsson McShane, Ulf Norinder, and Ola Spjuth. Predicting with confidence: using conformal prediction in drug discovery. *Journal of Pharmaceutical Sciences*, 110(1):42–49, 2021.

Vineeth Balasubramanian, Shen-Shyang Ho, and Vladimir Vovk. *Conformal prediction for reliable machine learning: theory, adaptations and applications*. Newnes, 2014.

Pedro J Ballester, Rick Stevens, Benjamin Haibe-Kains, R Stephanie Huang, and Tero Aittokallio. Artificial intelligence for drug response prediction in disease models. *Briefings in Bioinformatics*, 23(1):1–3, 2022.

Nicolas Bosc, Francis Atkinson, Eloy Felix, Anna Gaulton, Anne Hersey, and Andrew R Leach. Large scale comparison of QSAR and conformal prediction methods and their applications in drug discovery. *Journal of cheminformatics*, 11(1):1–16, 2019.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Wayne E Childers, Khaled M Elokely, and Magid Abou-Gharbia. The resurrection of phenotypic drug discovery. *ACS Medicinal Chemistry Letters*, 11(10):1820–1828, 2020.

Isidro Cortés-Ciriano, Gerard JP van Westen, Guillaume Bouvier, Michael Nilges, John P Overington, Andreas Bender, and Thérèse E Malliavin. Improved large-scale prediction of growth inhibition patterns using the NCI60 cancer cell line panel. *Bioinformatics*, 32 (1):85–95, 2016.

Martin Eklund, Ulf Norinder, Scott Boyer, and Lars Carlsson. The application of conformal prediction to the drug discovery process. *Annals of Mathematics and Artificial Intelligence*, 74(1):117–132, 2015.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

James P Hughes, Stephen Rees, S Barrett Kalindjian, and Karen L Philpott. Principles of early drug discovery. *British journal of pharmacology*, 162(6):1239–1249, 2011.

Kenneth M Merz Jr, Dagmar Ringe, and Charles H Reynolds. *Drug design: structure-and ligand-based approaches*. Cambridge University Press, 2010.

Screening methodology. National Cancer Institute. NCI-60 screening methodology (accessed: 01.04.2022). URL https://dtp.cancer.gov/discovery_development/nci-60/methodology.htm.

John G Moffat, Fabien Vincent, Jonathan A Lee, Jörg Eder, and Marco Prunotto. Opportunities and challenges in phenotypic drug discovery: an industry perspective. *Nature reviews Drug discovery*, 16(8):531–543, 2017.

NCI-60. National Cancer Institute. NCI-60 human tumor cell lines screen (accessed: 01.04.2022). URL https://wiki.nci.nih.gov/display/NCIDTPdata/NCI-60+Growth+Inhibition+Data.

NonConformist. Nonconformist v2.1.0 python package (accessed: 01.04.2022). URL http://donlnz.github.io/nonconformist/.

Ulf Norinder, Lars Carlsson, Scott Boyer, and Martin Eklund. 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.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

Chayanit Piyawajanusorn, Linh C Nguyen, Ghita Ghislat, and Pedro J Ballester. A gentle introduction to understanding preclinical data for cancer pharmaco-omic modeling. *Briefings in Bioinformatics*, 22(6):bbab312, 2021.

Anita Rácz, Dávid Bajusz, and Károly Héberger. Effect of dataset size and train/test split ratios in qsar/qspr multiclass classification. *Molecules*, 26(4):1111, 2021.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010.

Robert P Sheridan, Wei Min Wang, Andy Liaw, Junshui Ma, and Eric M Gifford. Extreme gradient boosting as a method for quantitative structure–activity relationships. *Journal of chemical information and modeling*, 56(12):2353–2360, 2016.

Pavel Sidorov, Stefan Naulaerts, Jérémy Ariey-Bonnet, Eddy Pasquier, and Pedro J Ballester. Predicting synergism of cancer drug combinations using NCI-ALMANAC data. *Frontiers in chemistry*, 7(509):1–13, 2019.

Vikash Singh, Michael Pencina, Andrew J Einstein, Joanna X Liang, Daniel S Berman, and Piotr Slomka. Impact of train/test sample regimen on performance estimate stability of machine learning in cardiovascular imaging. *Scientific reports*, 11(1):1–8, 2021.

Vladimir Svetnik, Andy Liaw, Christopher Tong, J Christopher Culberson, Robert P Sheridan, and Bradley P Feuston. Random forest: a classification and regression tool for compound classification and QSAR modeling. *Journal of chemical information and computer sciences*, 43(6):1947–1958, 2003.

Mihaly Szabo, Sara Svensson Akusjärvi, Ankur Saxena, Jianping Liu, Gayathri Chandrasekar, and Satish S Kitambi. Cell and small animal models for phenotypic drug discovery. *Drug Design, Development and Therapy*, 11:1957–1967, 2017.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005.

XGB. XGBoost developers, v1.3.3 (accessed: 01.04.2022). URL https://xgboost.readthedocs.io/en/stable/index.html.