11th Symposium on Conformal and Probabilistic Prediction with Applications

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

Saiveth HERNANDEZ HERNANDEZ[1]
Sachin VISHWAKARMA[1]
Pedro J. BALLESTER[1] [2]

[1]Cancer Research Center of Marseille (INSERM U1068, Institut Paoli-Calmettes, Aix-Marseille Université UM105, CNRS UMR7258), Marseille, France
[2]Department of Bioengineering, Imperial College London, London SW7 2AZ, UK

Contents

1. **Introduction**

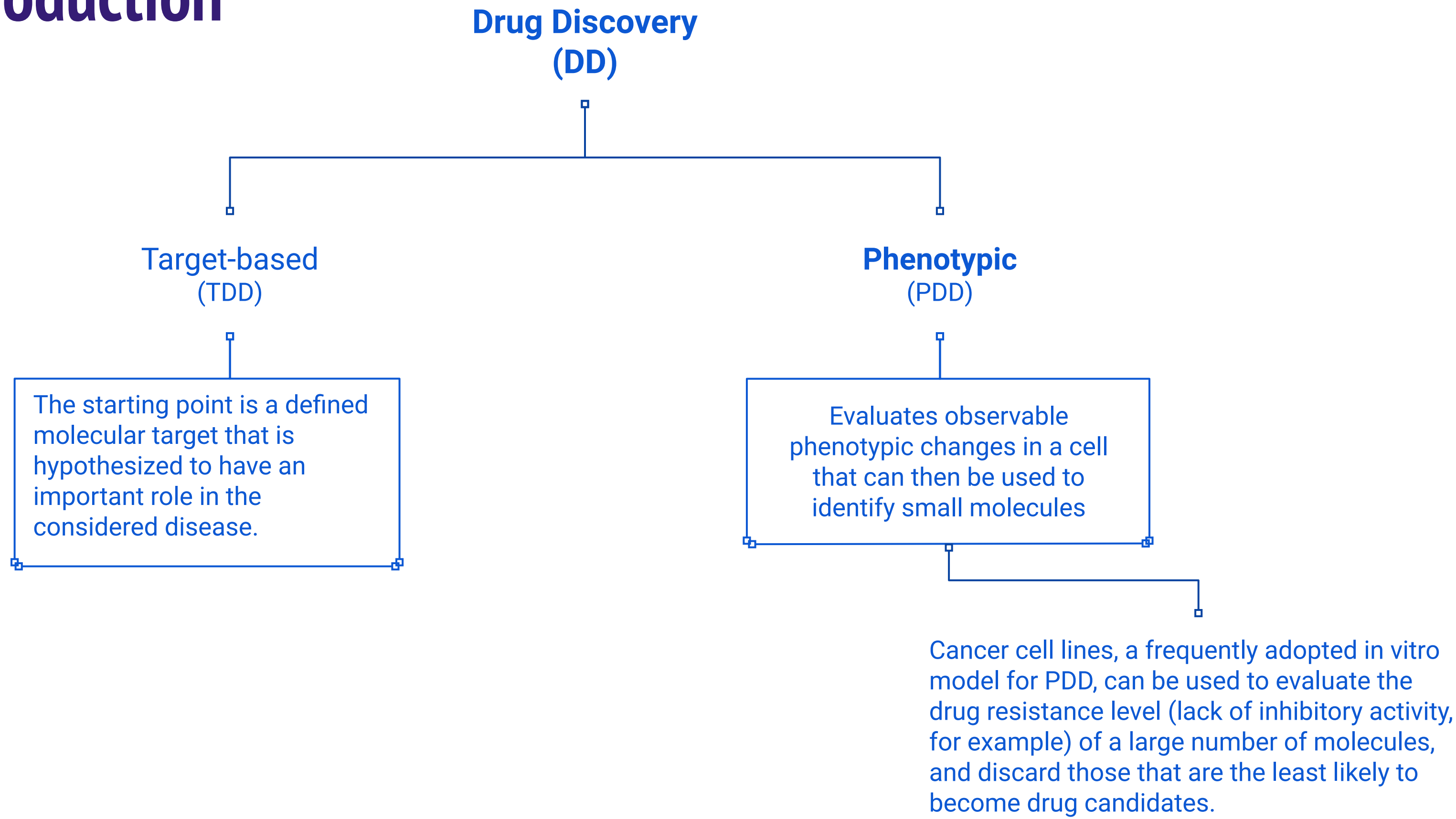
2. **Experimental design**

3. **Results and discussion**

4. **Conclusions**

Introduction

Introduction



Drug Discovery (DD)

Target-based (TDD)

The starting point is a defined molecular target that is hypothesized to have an important role in the considered disease.

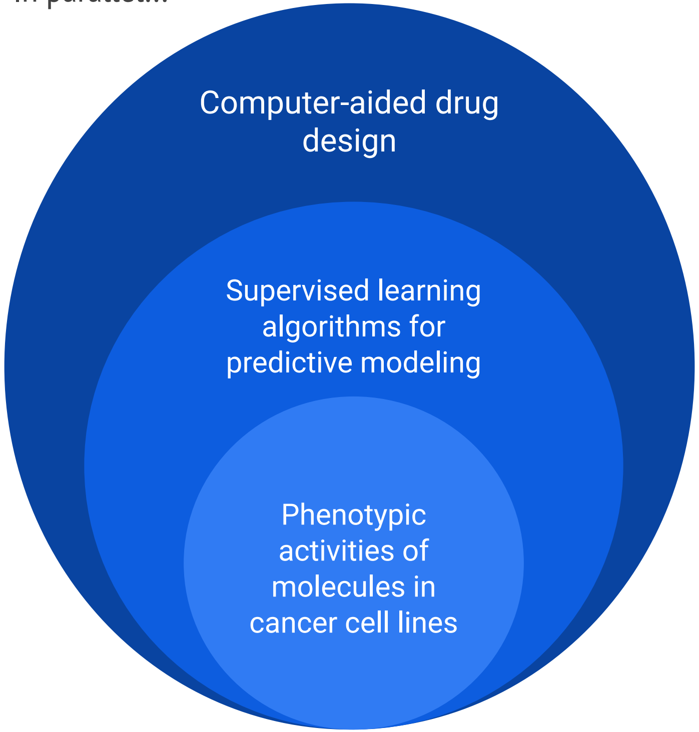
Phenotypic (PDD)

Evaluates observable phenotypic changes in a cell that can then be used to identify small molecules

Cancer cell lines, a frequently adopted in vitro model for PDD, 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.

Introduction

In parallel...



Computer-aided drug design
Supervised learning algorithms for predictive modeling
Phenotypic activities of molecules in cancer cell lines

Predictive models

- Assign a reliability to the whole model (e.g. by calculating the RMSE between predicted and observed activities of test set molecules
- 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)

Reliability is important for decision making

- 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.

Introduction

In this study,

- We investigated if CP can enhance the prediction of the inhibitory activity of molecules on a given cancer cell line.
- We investigated 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.

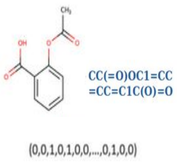
Experimental design

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 downloaded the data from NCI-60, which contains ~53K unique NSC IDs and a panel of 159 cell lines with ~3M pGI_{50}s measurements.

Data cleaning

1. Remove low-activity molecules ($pGI_{50} < 4$).
2. Compute the mean when more than one pGI_{50} measurements were available for the same NSC-Cell line.



Data representation

1. We generated the SMILES from the Chem2D_Jun2016.sdf file using the openbabel library.
2. Using the SMILES information and the RDKit library, we generated the Morgan circular fingerprints in bit format, with radius 2 and 256 bits.

There remained ~2.7M data points with measured pGI50 that corresponds to 50,555 total unique molecules for 50,846 unique NSC IDs and 60 cell lines.

NCI-60 data after preprocessing stage

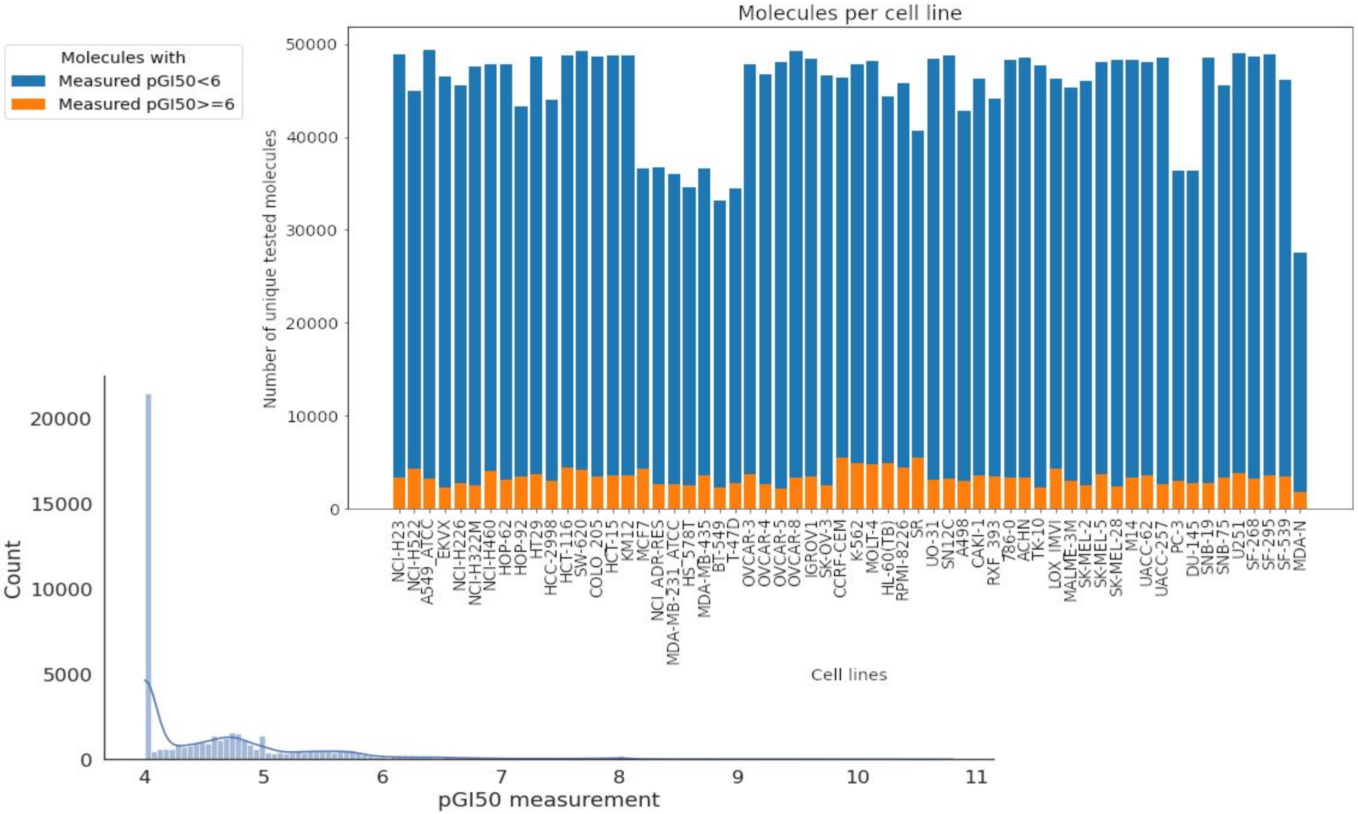


Figure 1: Each cell line has abundant data, although potent molecules are rather scarce. Distribution of pGI50 measurements in the 50,846 unique NSC IDs (bottom). Distribution of the number of unique molecules tested per cell line (top). The most potent molecules (pGI50 ≥ 6) for each cell line are in orange color.

Models

- Random forest (RF, Scikit-learn v1.0.2)
  - 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.
- Extreme gradient boosting (XGB, v1.3.3)
  - Is an implementation of gradient boosted decision trees designed for speed and performance

Hyperparameter tuning

- 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.
  - 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).

Conformal prediction

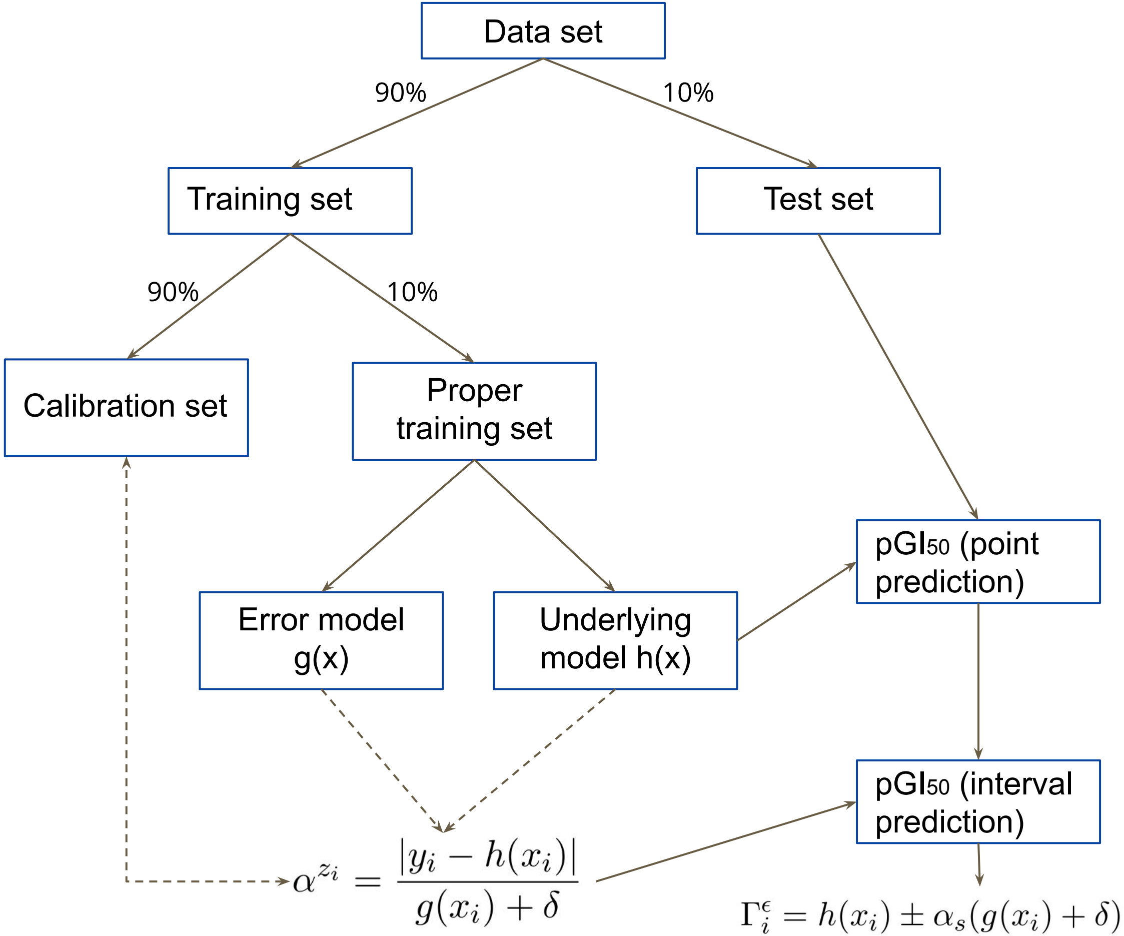
- We predict that a new instance will have a label that makes it similar to the old instances in some specific way and we use the degree to which the specific type of similarity holds within the old instances to estimate our confidence in the prediction.

- CP returns prediction regions, i.e, interval for regression problems.

Additional comments

- Nonconformist v2.1.0 python package

- We employed three training set partitions: 70-30, 80-20, and 90-10



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

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

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 random partition of the dataset, as explained in the previous slide, was applied to each of the 60 cell lines
- 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.
- Table 1 shows the combinations of the considered CP models, each of them evaluated at four confidence levels: 80, 85, 90, and 95%.

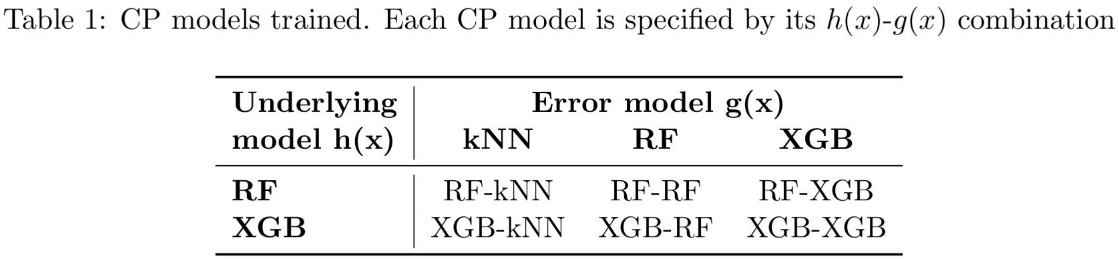


Table 1: CP models trained. Each CP model is specified by its $h(x)$-$g(x)$ combination

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

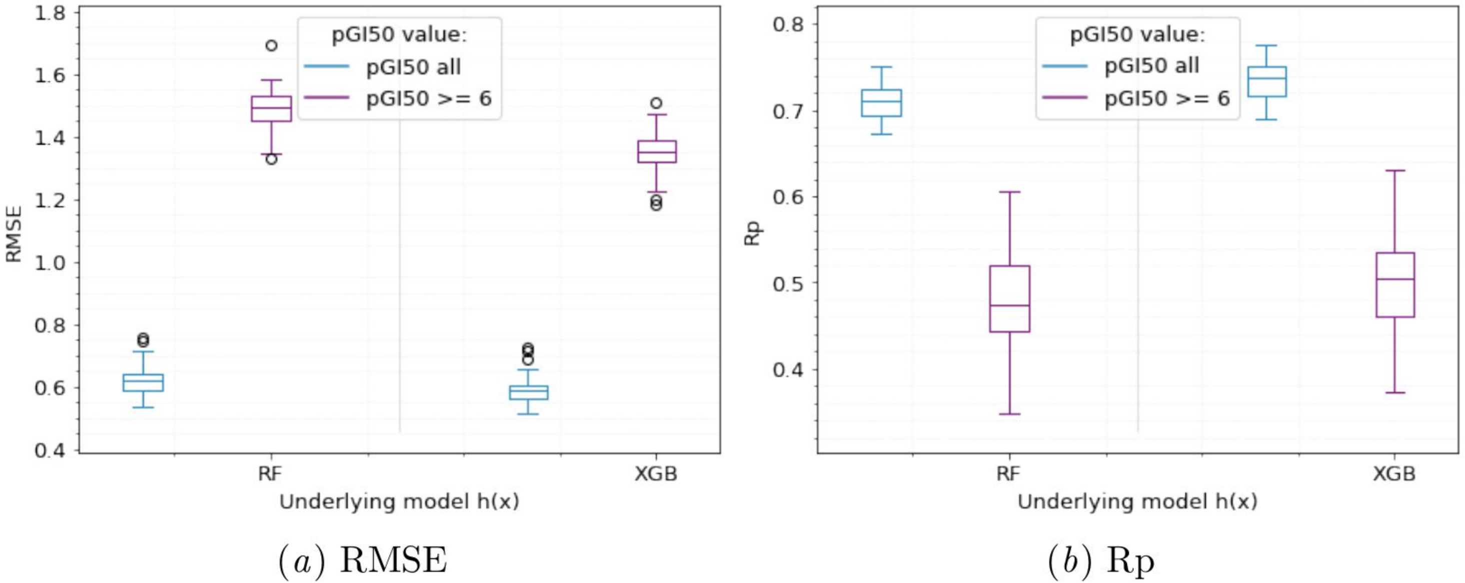
- Evaluation: Validity, efficiency, root mean squared error (RMSE), and Pearson's correlation coefficient (Rp)
  - RMSE and Rp were additionally computed for the most potent test molecules (those with measured pGI50 ⩾ 6)

A total of 1080 models (60 cell lines × 6 CP models × 3 training data partitions) were built and evaluated in this study.

Results and discussion

Underlying models

For each cell line, we employed the corresponding trained RF and XGB models to predict the pGI50 value of a given test set molecule from its molecular features.



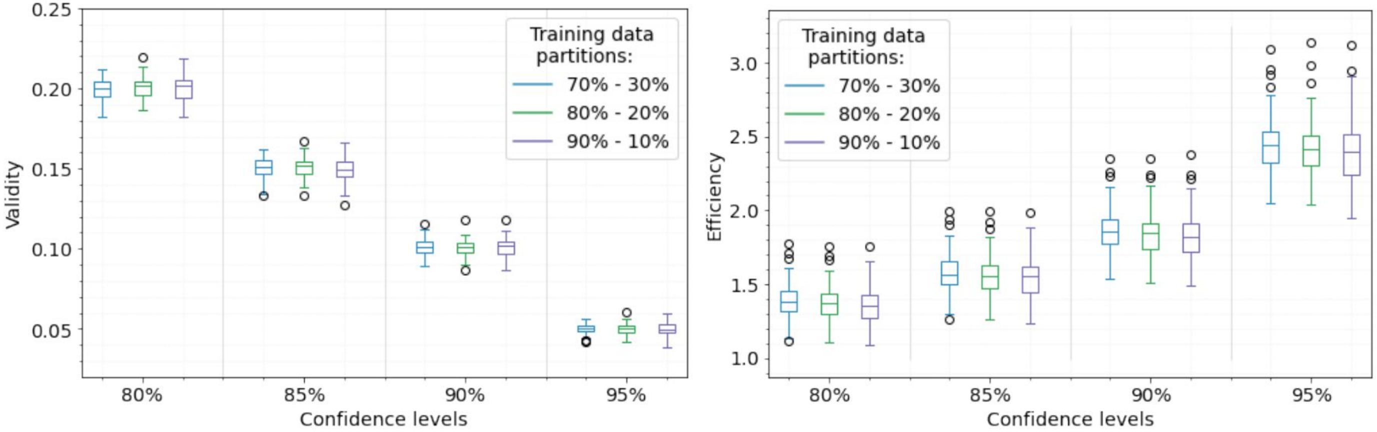
(a) RMSE

(b) Rp

Figure 2: The underlying models can predict the pGI50 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 pGI50 values using either RF or XGB models. Color code refers to all molecules (pGI50 all) or most potent molecules ($pGI_{50} \geq 6$) in the test sets.

Conformal prediction - Training data partitions

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.



(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.

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.

CP - Validity and Efficiency

- At each confidence level, the median validity is close to the required error in all CP models (ε± 0.002).
- Achieving valid and near-valid models has a cost in terms of worsened efficiency.
- 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 pGI50 prediction.

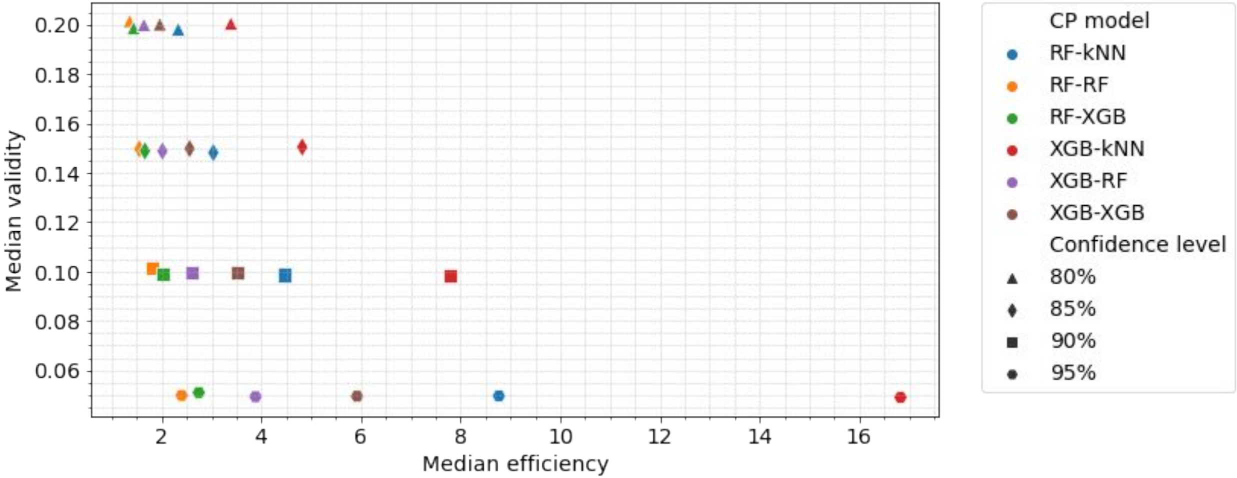
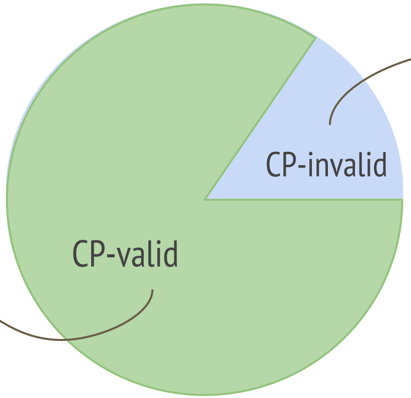


Figure 4. CP models have substantially different efficiency within a given confidence level. Median efficiency vs median validity across 60 test sets (one per cell line), at four confidence levels. Color code refers to the six CP models (h(x)-g(x)) employed. Markers refers to the requested confidence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

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.

Test set (non-CP)



The interval predicted does not contain the observed value

The interval predicted contains the observed value

CP-valid at 80, 85, 90 or 95% confidence level

non-CP = CP-valid U CP-invalid

Prediction performance using CP

We evaluate the performance of the **underlying models** and the **error models**

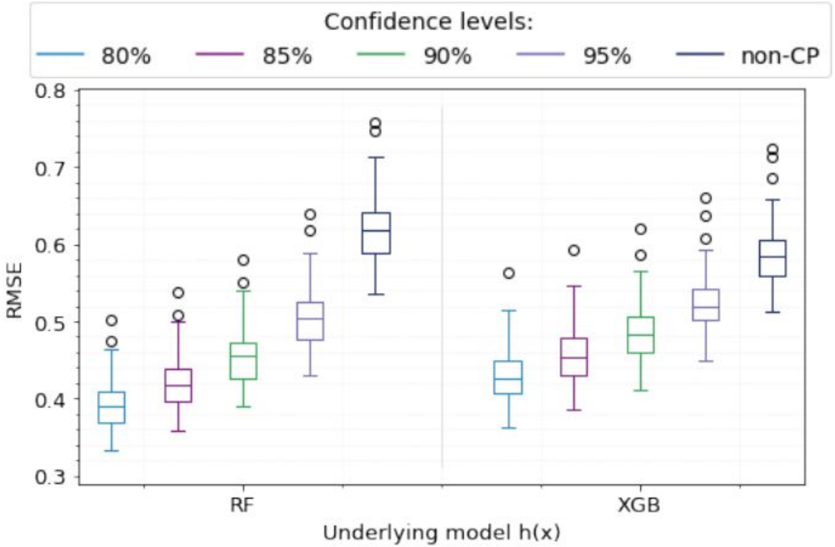


Figure 5. Relaxing the requested confidence level leads to more accurate CP-valid predictions. The boxplots summarize the RMSE distribution across the 60 test sets (one per cell line). RMSE values were computed between the observed and predicted pGI50 value using either RF or XGB as the underlying models. Color code refers to either CP-valid test set molecules, at each confidence level, or non-CP test set molecules.

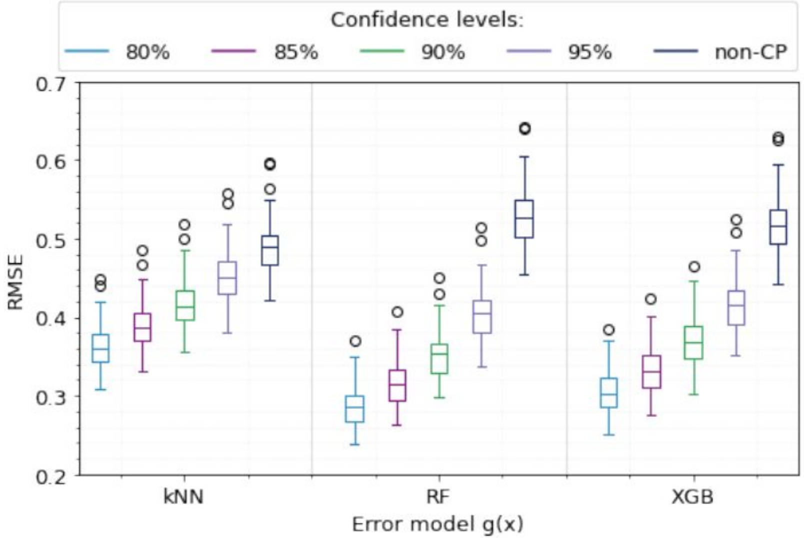
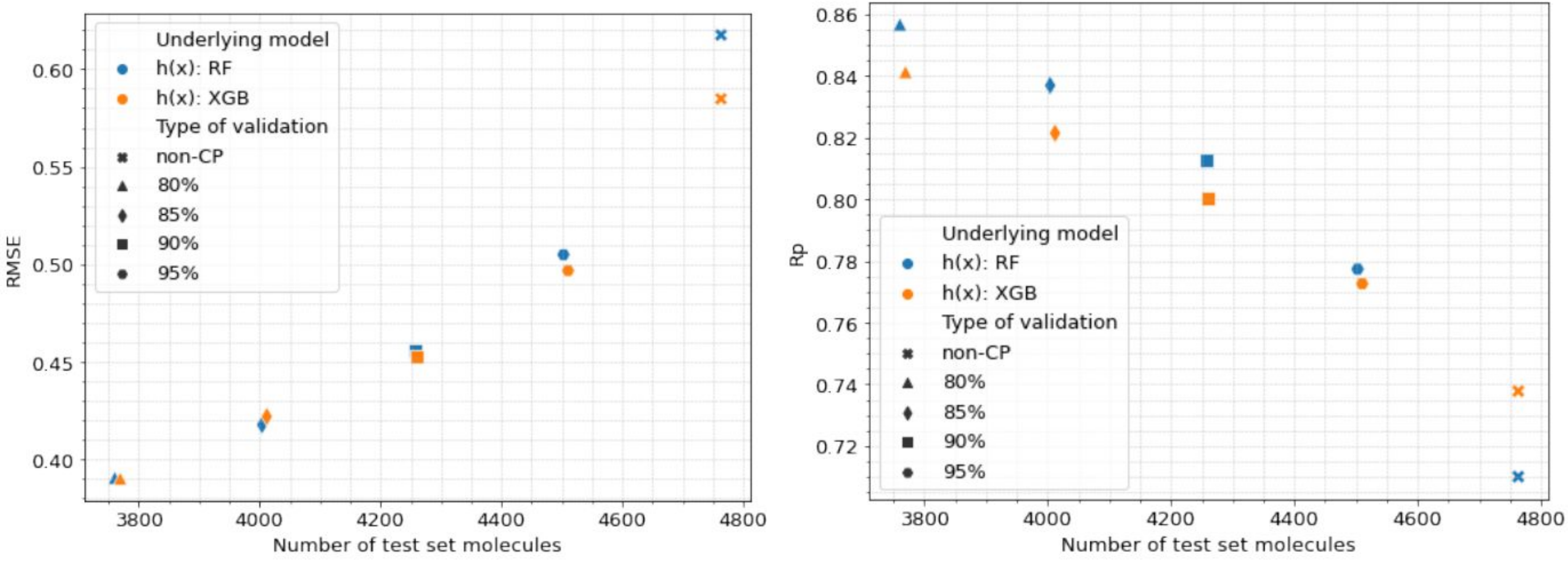


Figure 6. Molecules that are CP-valid have a better prediction of the pGI50 error. The boxplots summarize the RMSE in predicting pGI50 errors across the 60 test sets (one per cell line). RMSE values were computed between the observed and predicted pGI50 error, using either kNN-, RF- or XGB-based g(x) models. Color code refers to either CP-valid test set molecules, at each confidence level, or the non-CP test set molecules. Underlying model: RF.

Prediction performance using CP –All molecules

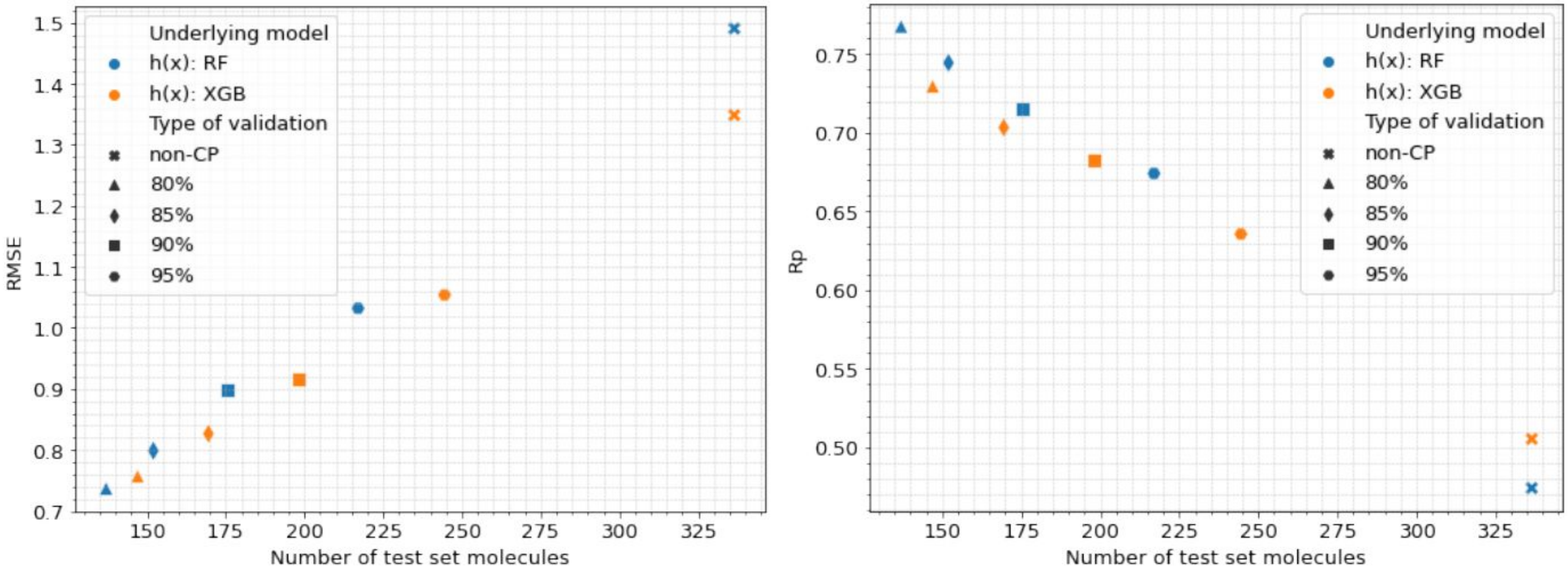


(a) RMSE - pGI_{50} all

(b) Rp - pGI_{50} all

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 pGI50 value (pGI50 all).

Prediction performance using CP - Most potent molecules



(c) RMSE - $\mathrm{pGI}_{50} \geq 6$

(d) Rp - $\mathrm{pGI}_{50} \geq 6$

Figure 7 (continued). 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 the most potent test set molecules (pGI50 ≥ 6).

Prediction performance using CP - Best and worst predicted cell line

- We look for the best/worst predicted cell lines 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.
- 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.

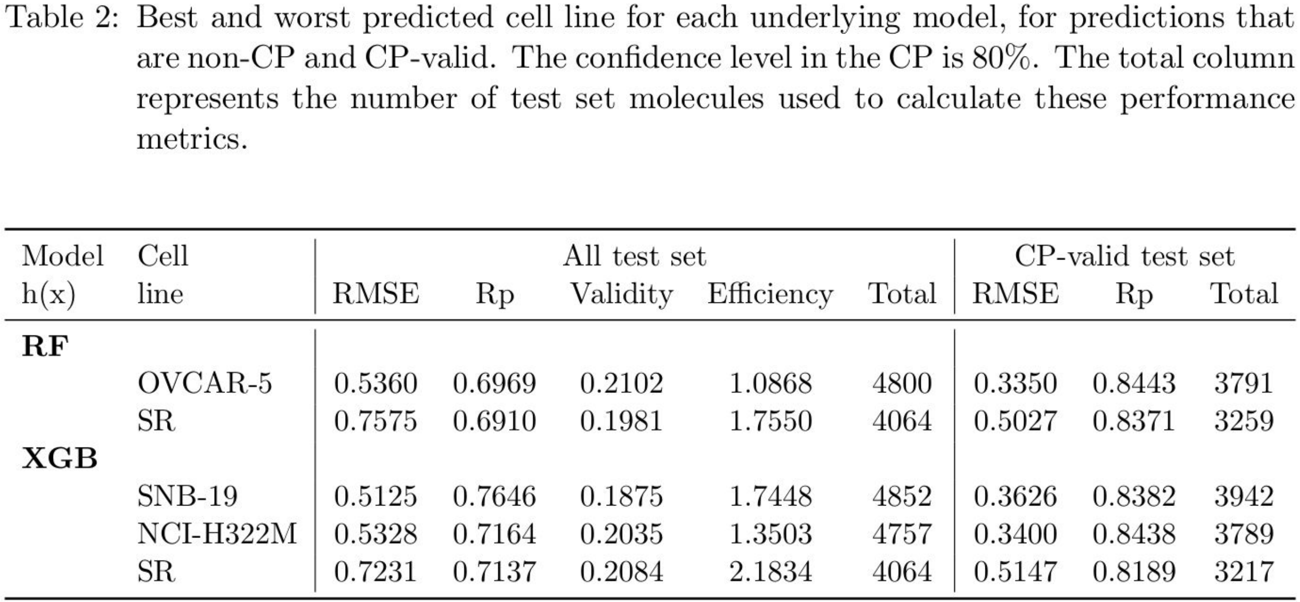
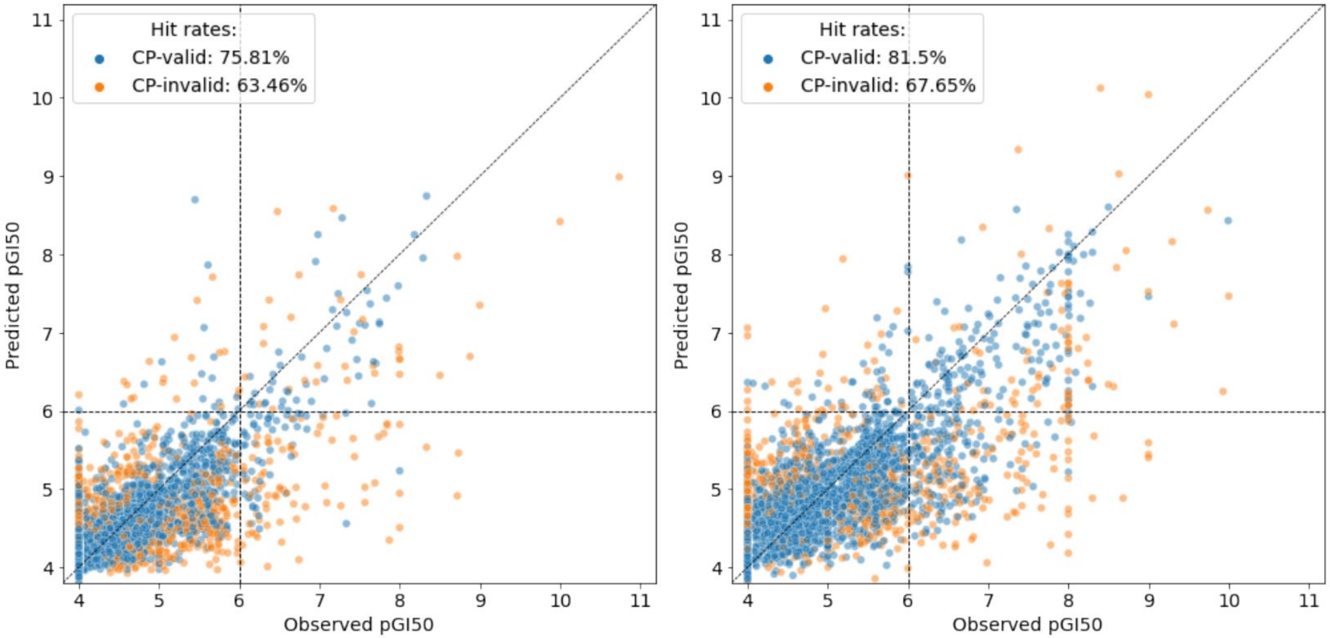


Table 2: Best and worst predicted cell line for each underlying model, for predictions that are non-CP and CP-valid. The confidence level in the CP is 80%. The total column represents the number of test set molecules used to calculate these performance metrics.

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

Prediction performance using CP - Best and worst predicted cell line



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

(b) Worst with RF-RF: SR

Figure 8. The prediction of the pGI50 of molecule-cell line pairs improves when CP is used. Observed and predicted pGI50 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 pGI50 ≥ 6.

Conclusions

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 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 pGI50 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 pGI50 value.

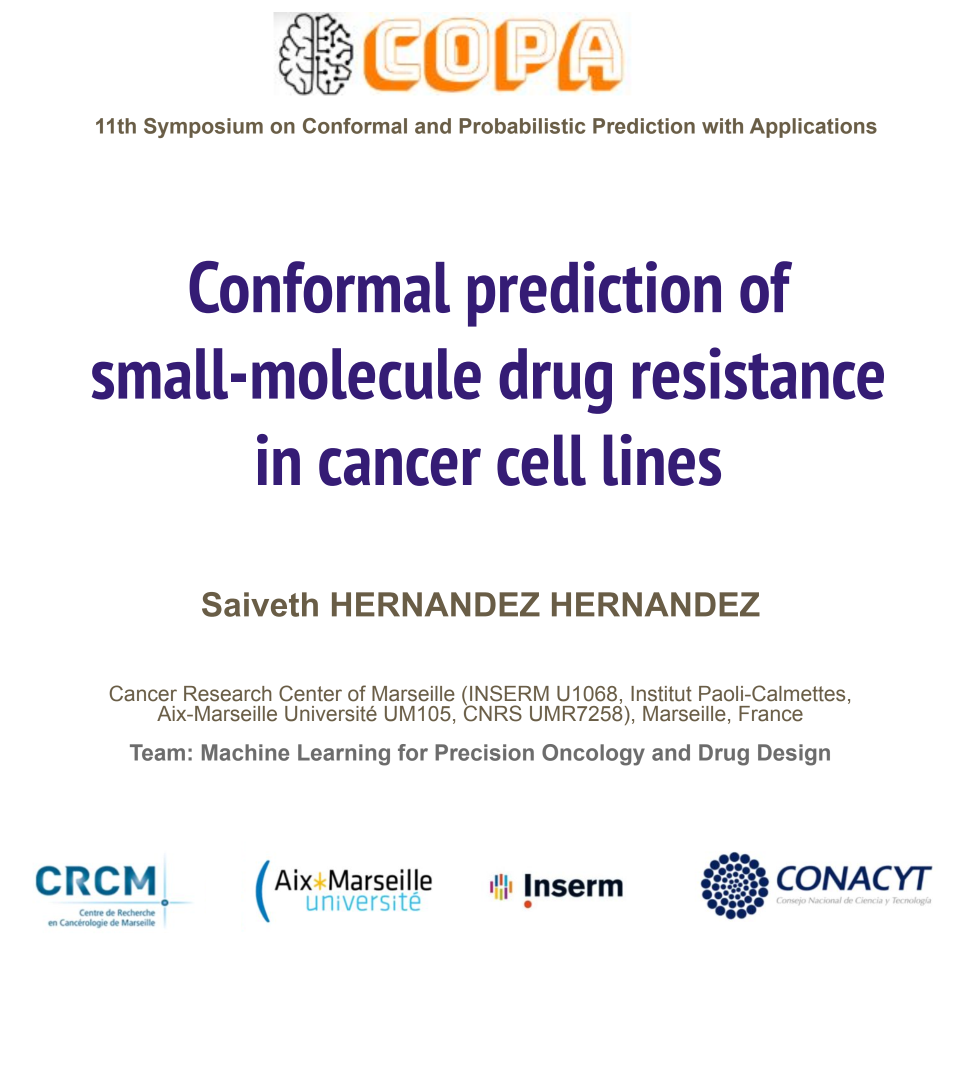
- The results from different training data splits showed that the chosen proper training set and the calibration set split do not affect the efficiency and validity results in each of the 60 test sets.

Conclusions

- 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 pGI50 ⩾ 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.



11th Symposium on Conformal and Probabilistic Prediction with Applications

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

Saiveth HERNANDEZ HERNANDEZ

Cancer Research Center of Marseille (INSERM U1068, Institut Paoli-Calmettes,
Aix-Marseille Université UM105, CNRS UMR7258), Marseille, France

Team: Machine Learning for Precision Oncology and Drug Design

THANK YOU