Conformal prediction of molecule-induced cancer cell growth inhibition challenged by strong distribution shifts

Saiveth Hernandez-Hernandez[1], Qianrong Guo[2], Pedro J. Ballester[2*]

[1] Centre de Recherche en Cancérologie de Marseille, Marseille 13009, France

[2] Department of Bioengineering, Imperial College London, London SW7 2AZ, UK

*** Corresponding author:** Pedro J. Ballester (p.ballester@imperial.ac.uk)

Abstract

The drug discovery process often employs phenotypic and target-based virtual screening to identify potential drug candidates. Despite the longstanding dominance of target-based approaches, phenotypic virtual screening is undergoing a resurgence due to its potential being now better understood. In the context of cancer cell lines, a well-established experimental system for phenotypic screens, molecules are tested to identify their whole-cell activity, as summarized by their half-maximal inhibitory concentrations. Machine learning has emerged as a potent tool for computationally guiding such screens, yet important research gaps persist. Consequently, this study focuses on the application of Conformal Prediction (CP) to predict the activities of novel molecules on specific cancer cell lines. Two CP models were constructed and evaluated on each cell line, resulting in a total of 120 performance evaluations (60 cell lines x 2 CP models) per training-test partition. From this comprehensive evaluation, we concluded that, regardless of the cell line or model, novel molecules with smaller CP-calculated confidence intervals tend to have smaller predicted errors once measured activities are revealed. It was also possible to anticipate the activities of dissimilar test molecules across 50 or more cell lines. These outcomes demonstrate the robust efficacy that CP models can achieve in realistic and challenging scenarios, thereby providing valuable insights for enhancing decision-making processes in drug discovery.

Keywords Conformal prediction, phenotypic virtual screening, benchmarking, cancer cell lines

Introduction

The drug discovery process typically starts with the identification of a small organic molecule with the potential to become a drug candidate(Hughes et al. 2011). There are two main strategies: phenotypic virtual screening (VS) and target-based VS. In target-based VS, there is a defined molecular target that has been validated for the considered disease and is leveraged to find molecules with activity for that therapeutic target(Quancard et al. 2020). For instance, crystal structures of the target can be exploited to identify such active molecules(Fresnais and Ballester 2020; Tran-Nguyen et al. 2023). In contrast, phenotypic drug discovery (PDD) does not rely on a specific molecular target or its mechanism of action in treating the disease(Moffat et al. 2017; Quancard et al. 2021). For instance, PDD leverages molecule-induced phenotypic changes in a cell(Diehl et al. 2005). Despite TDD being the predominant approach for the past 30 years, PDD has been shown to be more effective than has been given credit for(Lin et al. 2019; Sadri 2023; Vincent et al. 2022) and hence has been undergoing a resurgence(Moffat et al. 2017; Wagner 2016). Cancer cell lines is a well-established experimental system for PDD screens(Shoemaker 2006). Molecules are tested on a cell line to identify how sensitive or resistance they are, as determined by their half-maximal inhibitory concentrations(Piyawajanusorn et al. 2021). These concentrations are denoted as IC_{50}s or GI_{50}s or their logarithmic versions pIC_{50}s or pGI_{50}s.

Machine Learning (ML) has become a powerful approach to build models for phenotypic VS(Liu et al. 2023; Smer-Barreto et al. 2023; Stokes et al. 2020; Wong et al. 2023). Investigating the most suitable learning algorithms for this and related purposes is an active research area(Ballester 2023; Ballester et al. 2021; De Niz et al. 2016). In the context of cancer cell lines, research has been intense on the precision oncology side(Ballester and Carmona 2021; Chiu et al. 2019; Ianevski et al. 2021; Menden et al. 2013; Naulaerts et al. 2020; Niepel et al. 2019; Partin et al. 2023) , where

the activities of cell lines when treated with a given drug molecule are predicted. However, studies for phenotypic VS(Ammad-ud-din et al. 2014; Kumar et al. 2014), where the activities of molecules on a given cell line are predicted instead, are still scarce. Furthermore, reliability is usually assigned to the whole model (e.g. as the RMSE between predicted and measured activities of test set molecules), rather than a reliability at the instance level (e.g. as a predicted activity interval where the measured activity of the corresponding test set molecule is most likely to be). Estimating this reliability is important to support decision making, e.g. to test *in vitro* those molecules that are not only predicted to be most potent, but also whose predictions are the most reliable. In Conformal Prediction (CP) for regression, the activity of a molecule is predicted to be within a confidence interval, which can be employed to improve model performance.

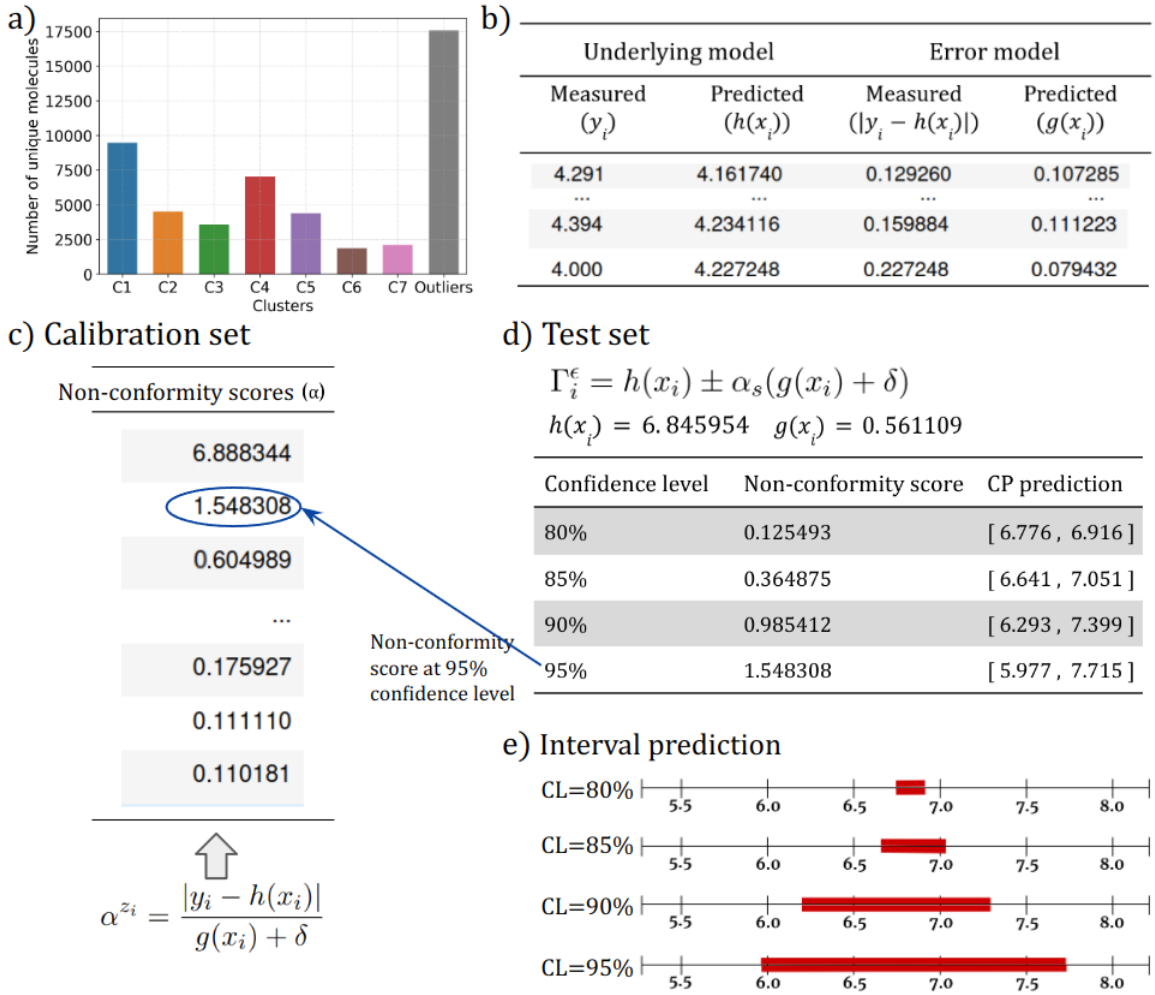

Fig. 1 Conformal Prediction enhances an underlying model with reliability estimates at a given confidence level. *A) Clusters used to build a CP model (LDMO-CV) per cell line. At each fold, 5 clusters are used as a proper training set, 1 cluster as a calibration set, and the remaining 1 cluster is used as a test set. B) At each fold, two supervised learning models are trained on the proper training set: the underlying model, $h(x)$, and the error model, g(x). $h(x)$ is trained on y, here pGI$_{50}$s, and $g(x)$ is trained on the resulting error residuals ($|y - h(x)|$) from the proper training set. C) The non-conformity scores are calculated for each sample, $z_i = (x_i, y_i)$, in the calibration set, using the underlying and error models trained on the proper training set samples. Each score represents how far was the predicted error to match the actual error for that calibration set sample. D) Given an unseen instance x_i for which a prediction is desired, the underlying model generates a point prediction, such as $h(x_i) = 6.846$ in this example. Additionally, the error prediction is computed for this instance, in this example, $g(x_i) = 0.561$. The non-conformity score is selected based on a user-defined confidence level. For instance, at a 95% confidence level, the non-conformity score is 1.548 because 95% of the scores in the calibration set are lower than this value. The interval prediction is formed by combining the predictions from both the underlying model $h(x)$ and the error model $g(x)$, along with the non-conformity score. Thus, by constructions, with a confidence level of 95% the true value will be within the predicted confidence intervals in at least 95% of the cases. In retrospect, if the interval contains the true value, these predictions are called CP-valid, otherwise, they are called CP-invalid. E) Graphical representation of the interval predicted (red) at each confidence level. Note that $uncertainty = 2\alpha_s g(x_i)$*

The application of CP to several drug design problems has been investigated(Alvarsson et al. 2021; Bosc et al. 2019; Cortés-Ciriano et al. 2016; Hernández-Hernández et al. 2022; Norinder et al. 2014). In particular, we published a proof-of-concept study showing that CP improves phenotypic VS on cancer cell lines on random training-test dataset splits(Hernández-Hernández et al. 2022). Here we will study how well these CP models perform on the much more realistic and challenging scenario where the test molecules are not only unseen, but also dissimilar to those in the training set. The evaluation will be carried out per cell line with two CP models and hence the results will be robust in that these will come from analyzing 2x60 independent cases. We will investigate how different are the chemical structures of the most potent cell-active molecules that were correctly predicted. We will also assess the potential of these models to anticipate the activities of a dissimilar test molecule across cell lines.

Results and discussion

Strong distribution shifts challenge non-CP model predictions in all 60 cell lines

We first compared the performance of the underlying LR, RF, and XGB models (h(x)) to identify the most informative molecular features for predicting pGI_{50}. The two sets of features under investigation were Morgan Fingerprints descriptors (MFPs) of 1024 bits and a combined set of Morgan Fingerprints and 7 Physicochemical Descriptors (MFPs+PCDs). The objective was to understand how the inclusion of PCD contributes to model accuracy, as measured by their pGI_{50} prediction error (RMSE) and correlation (Rp and Rs). To ensure the accuracy of the results, we used the LDMO-CV approach (clusters 1 to 7) to train and evaluate each of the 60 non-outlier datasets (one per cell line). The results of the LDMO CV run are summarized with the median of the model evaluation scores. Since the underlying model h(x) is a regression model, i.e., without CP, confidence levels are not applicable at this stage. Therefore, such predictions are referred to as non-CP predictions.

Fig. 2 compares how well the two sets of features, MFPs, and MFPs+PCDs, perform in terms of RMSE, Rp, and Rs across 60 cell lines. We performed a statistical test to assess the significance of the difference in performance between the models trained and evaluated with these two sets of features. The results indicate that including PCDs as input features (MFPs+PCDs) leads to a statistically significant improvement in the correlation between observed and predicted pGI50 values, as indicated by the Rs and Rp metrics. This suggests that including the PCDs enhances the model's robustness, allowing it to adapt more effectively to data distribution shifts. However, regarding RMSE, results indicate no significant difference between the two feature sets, regardless of the model used. The comparable RMSE values suggest that the model's predictive accuracy in terms of absolute errors remains consistent across both feature sets. As a result, the MFPs+PCDs feature set was chosen for subsequent analysis.

In addition to this, results from Fig. 2 demonstrate that non-linear models exhibited superior performance compared to the linear regression model. Based on this finding, the subsequent analyses were carried out exclusively with non-linear models.

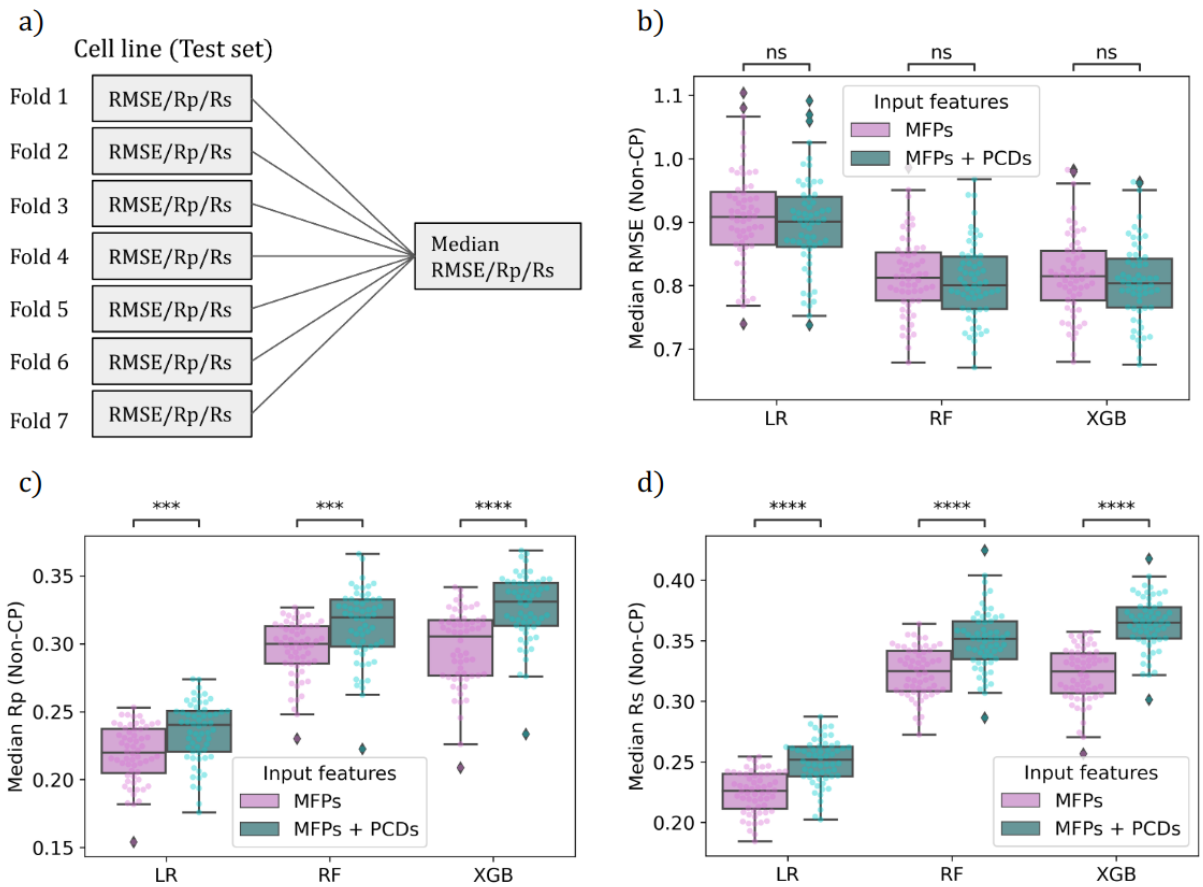

Fig. 2 The three underlying models improve when the PCDs are included as part of the input features (MFPs+PCDs). *Each molecule is encoded with an MFP of 1024 bits and 7 physicochemical descriptors. LR, RF, and XGB were used to evaluate the performance of each feature set (MFPs and MFPs+PCDs, identified with a color code), with each algorithm trained and evaluated on non-outliers molecules using an LDMO-CV approach (each fold being a different cluster of molecules). a) For each cell line, the evaluation score of the model is the median of the evaluation score of each fold, i.e. a 7-fold-median score. From b) to c), each boxplot represents, respectively, the distribution of the median RMSE, Rp, and Rs in predicting* pGI_{50} *values across the 60 cell lines. The results of the statistical test performed for the underlying models have been placed at the top of the boxplots. p-value annotation legend: ns: 5.00e-02 < p <= 1.00e+00; ***: 1.00e-04 < p <= 1.00e-03; ****: p <= 1.00e-04*

CP models are substantially more accurate than non-CP models in this scenario

Based on our previous study(Hernández-Hernández et al. 2022), we explored the performance of CP models at an 80% confidence level since this level balances reliability and interpretability in the pGI_{50} predictions. CP models are compared to the non-CP models (i.e., a standard machine learning model). This comparison shows that, across all evaluated metrics (RMSE, Rp, and Rs), CP models consistently outperformed the non-CP models (Fig. 3). More specifically, the median RMSE in the CP-valid predictions is about 0.3 pGI_{50} units lower than that of the non-CP predictions. This trend is also observed in correlation (Rp and Rs), with CP-valid predictions displaying a better alignment between the predicted and observed pGI_{50} values. One of the reasons for the superior performance of the CP models is their ability to assign uncertainty to their predictions. This is accomplished by generating prediction intervals associated with specific confidence levels. By doing so, the CP models inherently consider the uncertainties arising from distribution shifts, thus providing more reliable and accurate predictions.

Another significant result derived from Fig. 3 is related to the CP-invalid predictions. Since CP-valid predictions exhibited superior results across all evaluated metrics (RMSE, RP, and Rs), compared to the broader non-CP predictions (non-CP = CP-valid U CP-invalid), it suggests that the errors in the CP-invalid predictions disproportionately contribute to the overall model

inaccuracies. The concentration of errors within the CP-invalid predictions indicates potential challenges in accurately modeling uncertainties where the observed pGI_{50} values fall outside the predicted intervals.

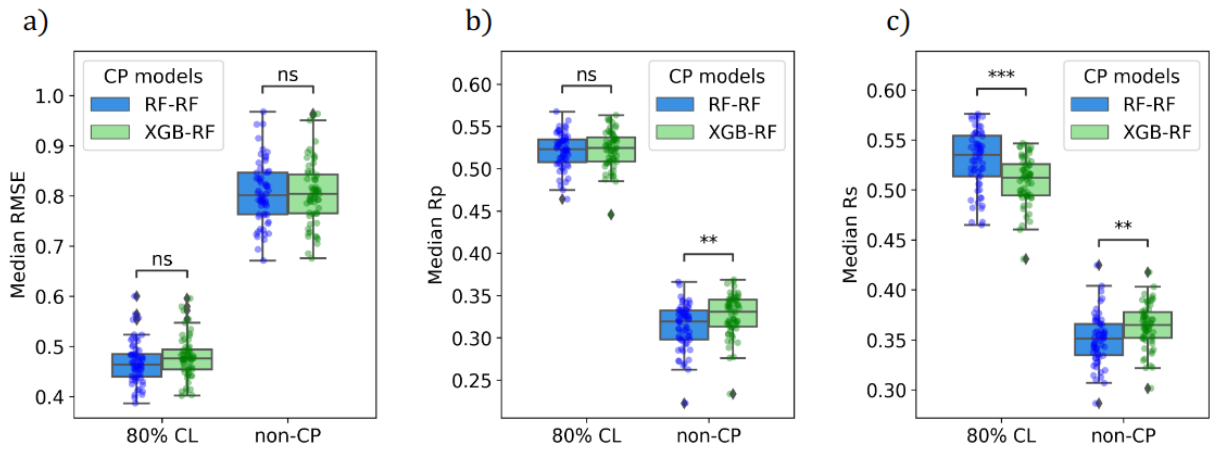

Fig. 3 CP-valid predictions exhibit lower errors and higher correlations than non-CP predictions. *RMSE, Rp, and Rs values were computed between the observed and predicted pGI_{50} values using either RF-RF or XGB-RF CP models. Each boxplot contains 60 points that represent the 7-fold-median a) RMSE, b) Rp, and c) Rs, summarizing their distributions across cell lines. Color code refers to the CP model used. The results of the statistical test performed for the models have been placed at the top of the boxplots. p-value annotation legend: ns: 5.00e-02 < p <= 1.00e+00; **: 1.00e-03 < p <= 1.00e-02; ***: 1.00e-04 < p <= 1.00e-03*

Table 1 Performance summary of best- and worst-performing cell line (CP models) in terms of the out-of-sample hit rate. The predictions come from either the RF-RF or XGB-RF modes, using MFPs+PCDs as the input feature set. The models were trained and evaluated using LDMO CV, with the reported values representing the median scores obtained across the folds. The evaluation metrics include median RMSE/Rp/Rs, median uncertainty, and hit rate. These metrics are computed only for the CP-valid molecules at an 80% confidence level. The hit rate is obtained by applying a threshold of 6 pGI_{50} units that distinguishes between active and inactive molecules. The "Rank" column indicates the ranking of the different models based on their performance in terms of the hit rate. The last column indicates the total number of CP-valid molecules in the dataset used for the calculation of the performance metrics. In bold are highlighted the best-performing values for each metric

Rank	CP Model	Cell line and Tissue	Median RMSE (CP-valid)	Median Rp (CP-valid)	Median Rs (CP-valid)	Median uncertainty (CP-valid)	Hit rate (CP-valid)	Number of CP-valid molecules
Best	RF-RF	SK-OV-3 (OV)	**0.398**	0.480	0.487	**1.41**	**100.0%**	24,153
Best	XGB-RF	TK-10 (RE)	0.410	**0.533**	0.515	1.67	**100.0%**	25,226
Worst	RF-RF	HCT-15 (CO)	0.464	0.507	**0.521**	1.66	25.0%	25,235
Worst	XGB-RF	SF-268 (CNS)	0.454	0.507	0.500	1.83	22.2%	25,852

Although there are no statistically significant differences in RMSE and Rp between the two CP models, a statistically significant difference exists in Rs. These differences are further explored to identify the strengths of each CP model. The RF-RF and XGB-RF models show comparable results in RMSE and Rp, indicating proficiency in reducing absolute errors and establishing linear relationships between predicted and observed pGI_{50} values. In addition, the RF-RF performs better

in capturing non-linear dependencies, as shown by their Rs. These findings suggest that while the CP models may have similar performance in terms of overall prediction accuracy (RMSE) and linear correlations (Rp), they may differ in their ability to accurately capture the underlying patterns and non-linear relationships (Rs) among the 60 cell lines.

Among the 60 cells, we identified both the best-performing and worst-performing cells based on their hit rates computed only over the CP-valid molecules. In cases where multiple cells had identical hit rates, we utilized uncertainty as a tie-breaking measure. Table 1 shows these results. When considering only CP-valid molecules, both the RF-RF model and the XGB-RF model achieve a 100% hit rate, which highlights the importance of using CP to calibrate the predictions. The RF-RF model identifies cell line SK-OV-3 as the optimal choice, while the XGB-RF model selects cell line TK-10. Based on the uncertainty, the RF-RF assigns more confidence in the predictions for the molecules tested against the SK-OV-3 cell line.

Useful predictions are achieved despite training on highly dissimilar molecules

To simulate a real screening scenario, we focused on the best-performing cell lines predicted by the two CP models, SK-OV-3 and TK-10 (see Table 1 for performance metrics). We first identify the 100 CP-valid molecules with the highest predicted pGI_{50} values for these cell lines. Subsequently, we narrowed our search to the three molecules with the highest measured pGI_{50} values from this subset. This mimics the real screening scenario, where a large pool of candidate molecules is initially considered, followed by a refinement process to identify the most promising candidates for further analysis. Our goal is to measure how dissimilar the best CP-valid molecules are from the molecules the CP model already saw during the training stage. To accomplish this, we search for its closest counterpart for each of the three best-performing CP-valid molecules by computing the Tanimoto similarity within the training clusters. This search yields six candidates (one per cluster, as the test cluster is not considered during this search), from where we identify the training molecule that exhibits the highest degree of similarity to the best-selected CP-valid molecule. This selection is made as the other five candidates exhibit a higher dissimilarity to this particular molecule.

Table 2 Similarity analysis between the best-performing CP-valid molecules and the training molecules (non-outlier dataset). For the best-performing cell lines, SK-OV-3 and TK-10 (see Table 1 for performance metrics), we mimic the real screening scenario, where a large pool of candidate molecules is initially considered, followed by a refinement process to identify the most promising candidates for further analysis. The three molecules with the highest measured pGI_{50}s among the top 100 CP-valid predictions in the test set are shown in this table (thus simulating the results of *in vitro* testing these 100 molecules). For each of these three CP-valid molecules, we search for their respective 6 closest counterparts within the training clusters, from where we identify the training molecule that exhibits the highest degree of similarity to the best-selected CP-valid molecule. The cell line, NSC-ID, measured pGI_{50}, and predicted pGI_{50} are included in this table

CP Model	Cell line	Test molecule			Most similar training molecule		
		NSC ID	Measured pGI_{50}	Predicted pGI_{50}	NSC ID	Measured pGI_{50}	Similarity
RF-RF	SK-OV-3	777407	8.000	5.481	765805	5.240	0.422
RF-RF	SK-OV-3	677598	7.061	6.815	662190	5.000	0.317
RF-RF	SK-OV-3	360702	6.711	5.326	630602	5.036	**0.290**
XGB-RF	TK-10	683692	7.695	7.026	662190	5.000	0.345
XGB-RF	TK-10	696078	7.597	6.904	662190	5.000	0.339
XGB-RF	TK-10	677598	7.294	7.109	662190	5.000	0.317

Table 2 shows that regardless of the CP model used, either RF-RF or XGB-RF, all the similarities observed between the molecules were relatively small, ranging from 0.29 to 0.42. These modest similarity values indicate subtle structural similarities or shared features among the molecules (Fig. 4). Using the XGB-RF model, the three most promising molecules identified were predicted with pGI_{50} values greater than 6; therefore, these molecules are considered active. Notably, the most similar molecule found in the training set for all three promising molecules was the same (NSC-662190), highlighting a consistent pattern in their structural similarities. In contrast, among the three promising molecules obtained using the RF-RF model, only one was considered an active molecule (predicted pGI_{50}=6.8).

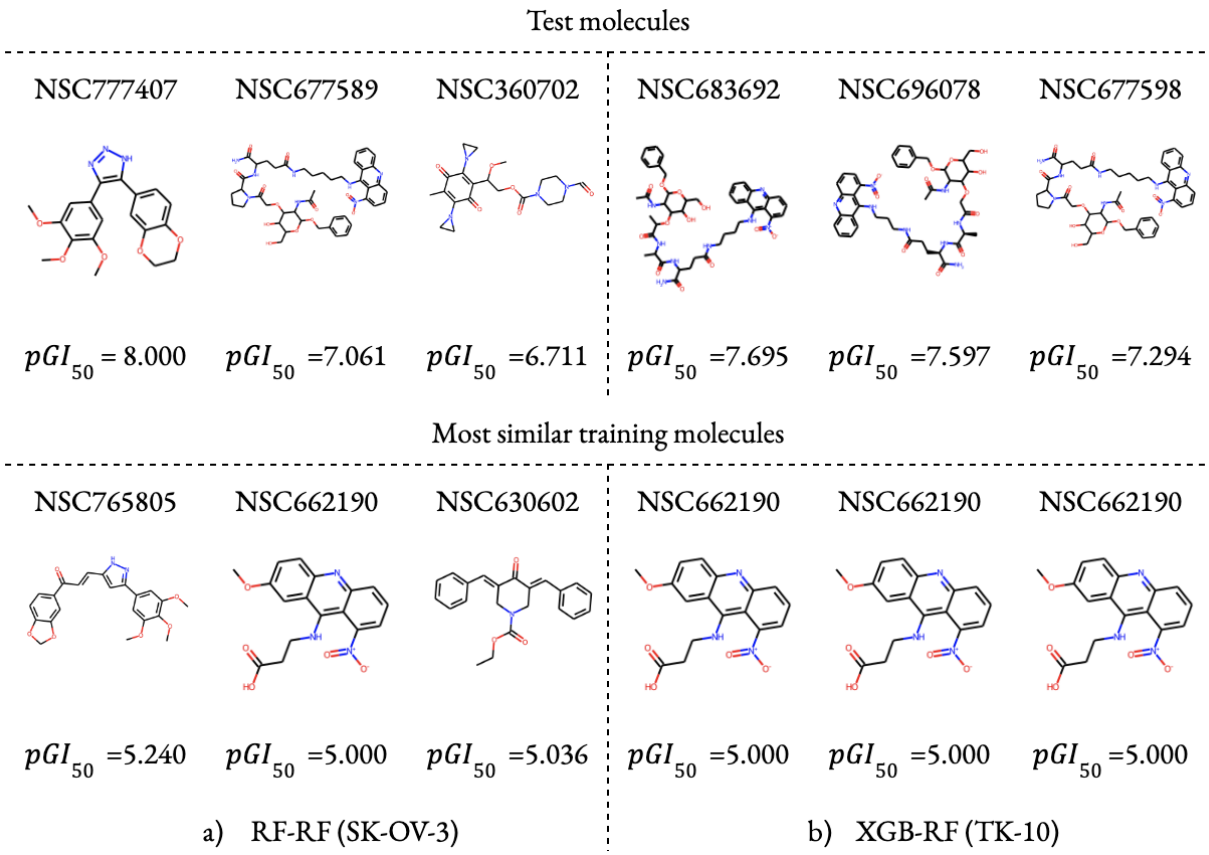

Fig. 4 ***Chemical structures of the three molecules with the highest measured pGI50s among the top 100 CP-valid predictions in the test set along with their most similar training molecules (non-outlier dataset).*** *Test molecules are shown at the top, while their most similar training molecules are shown at the bottom. The measured pGI_{50} is shown at the bottom of each molecule. Regardless of the CP model used, either a) RF-RF or b) XGB-RF, all the similarities observed between the molecules were relatively small, ranging from 0.29 to 0.42. See Table 2 for metrics performance*

CP predictions with low uncertainty are more accurate than those with high uncertainty

The CP models detailed in the previous sections underwent training and evaluation within a relatively controlled environment represented by the non-outlier dataset. However, a question remains concerning the ability of a CP model to effectively generalize to molecules with dissimilar chemical structures from those in its training set. To address this, we retrain a CP model on the non-outlier datasets and test it on the outlier dataset for each of the 60 cell lines. Thus, this experimental design encompasses a comprehensive evaluation of the adaptability and performance of a CP model in handling different scenarios to which it was exposed during training. The two best CP models are re-trained per cell line using the complete set of non-outlier

molecules, for which 6 of the 7 clusters are used as the proper training set, and the remaining cluster is used as the calibration set (Fig. 1-a).

The results from the XGB-RF and RF-RF models agree in identifying NCI-H322M as the cell with the highest predicted hit rate, demonstrating consensus between the two models. In the subsequent analysis, we compared the results of the 5% of molecules with the lowest uncertainty to the 5% with the highest uncertainty. It is important to note that by considering only these predicted uncertainties, this analysis does not employ any information whatsoever from the measured pGI_{50} values in the independent test sets (outlier molecules tested on each cell line). This is suitable for prospective studies, where we are interested in making predictions or drawing conclusions about future outcomes based on data that has not yet been generated.

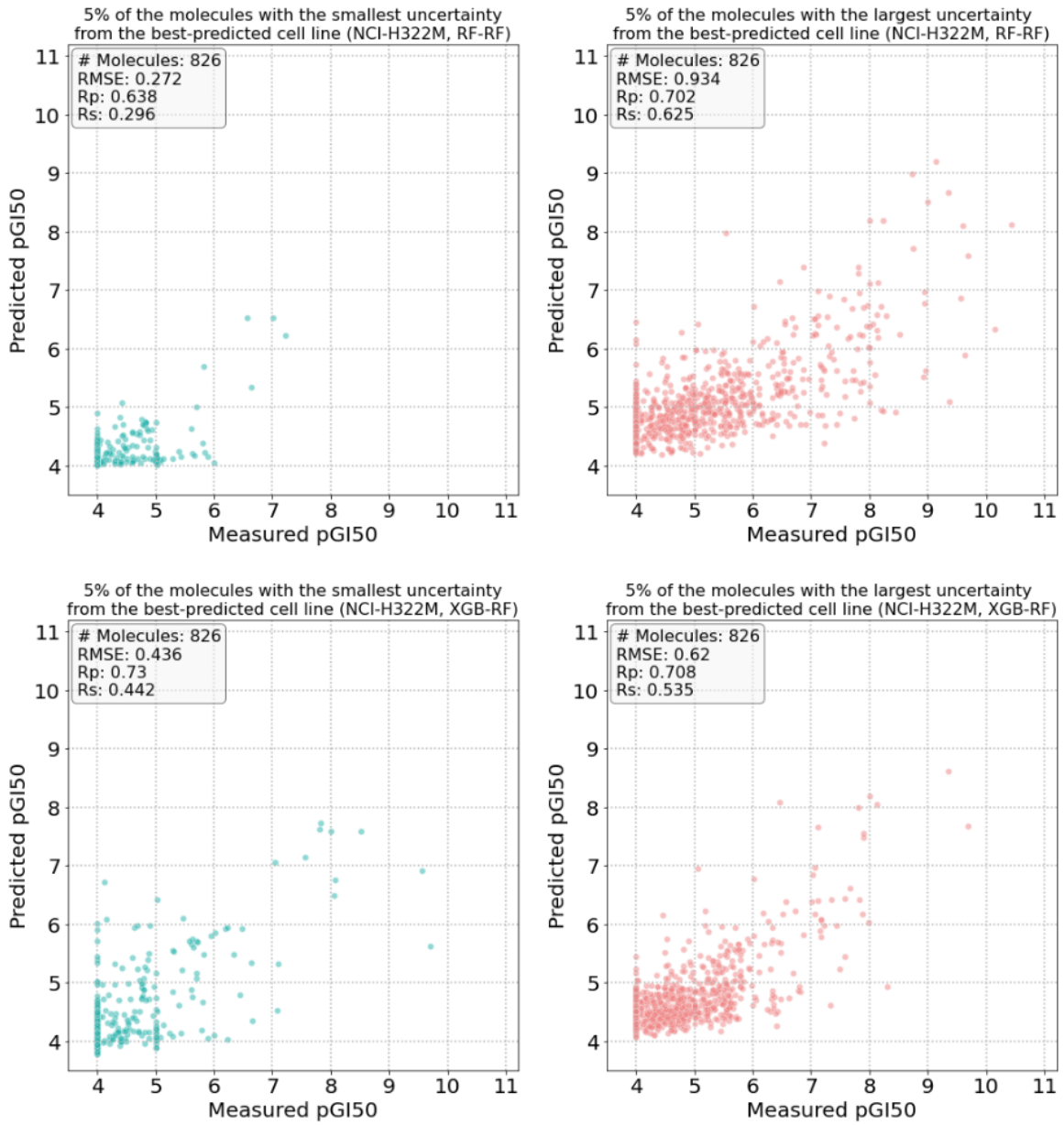

Fig. 5 The prediction of the pGI_{50} of molecule-cell line pairs improves when CP is used. *Observed and predicted pGI_{50} value in the best-predicted cell line, obtained by the RF-RF (top) and the XGB-RF (bottom) CP models. Predictions are obtained at an 80% confidence level. The uncertainty is calculated by multiplying the predicted error for a given molecule by the non-conformity score chosen at an 80% confidence level. Each plot contains the 5% (826 molecules) of the test set molecules, either those with the smallest (left) or largest uncertainty (right). See Table 6 for performance metrics. The anomalous density of molecules at pGI_{50}=4 is due to the inhibitory activity of weaker molecules being rounded up at the closest considered pGI_{50} of 4.*

Fig. 5 shows that both the RF-RF and XGB-RF models benefit when considering only the 5% of molecules with the lowest uncertainty. In this scenario, the RF-RF model achieves lower RMSE while the XGB-RF achieves higher Rp and Rs. In contrast, for the 5% of molecules with the highest uncertainty, the RF-RF model shows better Rs but a higher RMSE compared to those obtained by the XGB-RF CP model. These findings underscore the significance of considering uncertainty levels in model predictions, as lower uncertainty is consistently associated with enhanced predictive accuracy and reliability across both models.

The ensemble of CP models can predict the NCI-60 cell line selectivity of a test molecule

An inherent advantage of the CP framework is its ability to assign a degree of uncertainty at an instance level. We leverage this characteristic to identify the best-performing molecules among the 60 cell lines based on their respective degrees of uncertainty for both CP models. In this instance-level evaluation, molecules with lower uncertainty are considered the best-performing. The uncertainty is calculated by multiplying the predicted error for a given molecule by the non-conformity score chosen at a confidence level. In essence, the uncertainty corresponds to the length of the predicted interval and reflects the range within which the observed value of the molecule is likely to fall.

To achieve this task, we first merge the data for each molecule across all 60 cell lines and define additional criteria to keep only the CP-valid and most potent molecules ($pGI_{50} \geq 6$). After this filter, the RF-RF model retained 5,252 NSC-Cell points for 301 small molecules screened against the 60 cancer cell lines, resulting in a matrix completeness of 29.08%. We further narrowed our search to only those NSCs where data is available for at least 50 cell lines. As a result, we identified 37 NSC, from which we selected the best-performing CP-valid molecule based on their uncertainty. Similarly, for the XGB-RF model, the filtering process yielded 6,075 NSC-Cell points for 567 small molecules, representing a matrix completeness of 17.86%. Subsequently, we identified 20 NSCs for which the data is available for at least 50 cell lines, and from which we selected the best-performing CP-valid molecule based on their uncertainty.

The molecule NSC-ID: 747457, identified by the RF-RF CP model, demonstrates a relatively confident prediction (uncertainty of 1.8 units) and high potency across 52 cell lines, with pGI_{50} values higher than 6. In contrast, the XGB-RF CP model identified the molecule NSC-7521 (proscillaridin), which demonstrates a relatively confident prediction given by an uncertainty of 1.3 units. Moreover, this molecule demonstrates a high potency across an even broader spectrum of 58 cell lines, with pGI_{50} values higher than 8. Therefore, in terms of uncertainty and potent activity, the XGB-RF CP model, with the proscillaridin molecule, outperforms the results from the RF-RF CP model.

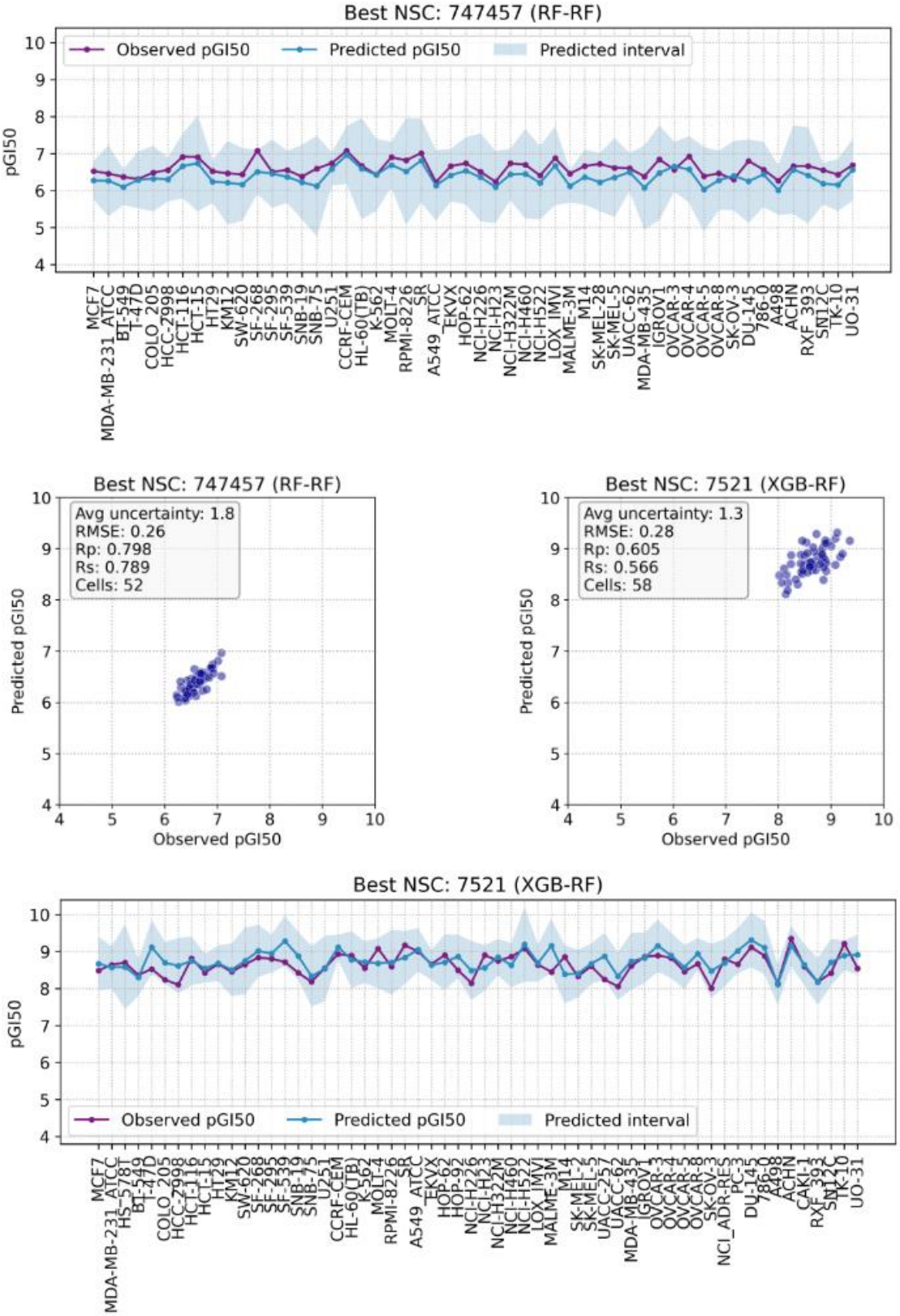

Fig. 6 The ensemble of CP models can predict the selectivity of a dissimilar molecule across NCI-60 cell lines. *Relationship between the measured and predicted pGI_{50} values across different cells, for the best-performing 80% CP-valid predictions. The analysis prioritizes the identification of molecules predicted reliable (CP-valid) and potent $pGI_{50} \geq 6$, either by the (top) RF-RF or (bottom) XGB-RF model. The best CP-valid molecule (NSC:747457) is selected from a pool of 37 molecules for the RF-RF, while for the XGB-RF, the best CP-valid molecule (Proscillaridin; NSC:7521) is selected from a pool of 20 molecules tested on at least 50 cell lines. The best-performing molecule is based on its uncertainty (length of the interval predicted). The uncertainty is calculated by multiplying the predicted error for a given molecule by the non-conformity score chosen at an 80% confidence level. The evaluation metrics also include RMSE, Rp, and Rs. Each dot in the plot corresponds to a cell line treated with the corresponding molecule*

Conclusions

We have developed CP models for out-of-sample prediction and confidence intervals of the pGI_{50}s of molecules on each NCI-60 cell line. This panel contains data from 60 distinct human tumor cell lines and is a comprehensive resource for identifying and characterizing novel compounds with growth inhibition or killing of tumor cell lines on various cancer types.

The LDMO-CV approach used to build and evaluate these CP models on clusters of non-outlier molecules leads to models that generally perform well on unseen molecules dissimilar to those in the training. This is the most common scenario as real-world compound libraries contain many more molecules with higher diversity than those already tested on the cell line. We observed that the higher the requested confidence level, the higher the prediction uncertainty becomes, which in turn results in worse RMSE when predicting such dissimilar molecules. CP models trained on all non-outlier molecules and tested on outlier molecules further establish the ability of these models to handle training-test partitions introduction strong distribution shifts.

Overall, we showed that predictions with associated confidence intervals enhance decision making in that novel molecules with smaller prediction error can be selected for further exploration and experimental validation. Furthermore, these CP models could predict the selectivity of dissimilar molecules across NCI-60 cell lines. Focusing on the 80% CP-valid predictions, we identified molecules exhibiting both reliability and potency. Notably, NSC:747457 and NSC:7521 (Proscillaridin) emerged as the CP-valid molecules whose cell line selectivity could be predicted best, highlighting the potential of this approach for strategic molecule prioritization in drug discovery efforts.

Methods

Datasets

The availability of preclinical data enables the development and validation of ML-based models in drug design. Such is the case of the National Cancer Institute (NCI-60) human tumor cell lines screen, which since 1990 has been used by the cancer research community to find compounds with potential anticancer activity. The NCI-60 panel utilizes 60 different human tumor cell lines to identify and characterize novel compounds with growth inhibition or killing of tumor cell lines(Shoemaker 2006). These 60 different human tumor cell lines comprise 9 broad cancer types: leukemia (LE), melanoma (MEL), lung (NSCL), colon (CO), brain (CNS), ovary (OV), breast (BR), prostate (PR), and kidney cancers (RE). All compounds screened are initially tested in a one-dose assay on the full NCI-60 panel. Compounds showing significant growth inhibition at this assay are then evaluated against the NCI-60 panel in a 5-dose assay(Shoemaker 2006). Each compound screened by the NCI-60 is identified with a unique registration number called the National Service Center (NSC) ID. In this study, we modeled 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.

Data quality plays a key role in developing AI/ML models. Therefore, data cleansing is necessary to ensure the use of high-quality data in generating these models. To achieve this, we follow the preprocessing protocol described by some of us(Hernández-Hernández et al. 2022) and comprising the removal of low-activity molecules ($pGI_{50} < 4$) as they do not have therapeutic potential and the calculation of the mean when more than one pGI_{50} measurements were available for the same NSC-cell line pair. Next, each molecule is represented by its 1024-bit MFP with a radius of 2, corresponding to the presence or absence of a particular substructure in the molecule. The 1024-bit MFP size has been successfully used in retrospective studies, e.g.(Siramshetty et al. 2020). While the choice of fingerprint and metric typically has little effect on downstream predictions. For example, in target prediction based on molecular similarity(Peón et al. 2016), this choice led to practically the same performance as other fingerprints and metrics, suggesting that it should be a near-optimal choice here. This process obtained 2,707,434 data points (50,555 small

molecules screened against 60 cancer cell lines), representing a matrix completeness of 89.25%. For these 50,555 small molecules, we also computed seven PCDs to encompass both the pharmacokinetic and pharmacodynamic aspects of drug action. These PCDs include molecular weight, logP, number of hydrogen bond donors (HBD), number of hydrogen bond acceptors (HBA), total polar surface area, number of aliphatic rings, and number of aromatic rings. These descriptors provide insights into the molecular characteristics influencing the absorption, distribution, metabolism, and excretion or ADME processes (pharmacokinetics), as well as the interaction of the drug with its molecular targets (pharmacodynamics). These features were calculated with the RDKit software suite(Landrum 2016).

UMAP split for each of the 60 datasets

Molecular clustering involves grouping similar molecules based on their chemical properties or structural features. In a previous study(Hernández-Hernández and Ballester 2023), we identified an UMAP-based clustering method as the best way to cluster the molecules screened by the NCI-60 panel. The molecules are encoded using MFPs of size 1024 and radius 2. UMAP-based clustering refers to first applying the UMAP dimensionality reduction algorithm and subsequently performing agglomerative clustering within the resulting embedded space. In that study, we first defined an outlier as the molecule whose Tanimoto similarity to its most similar molecule is lower than or equal to a given cutoff. Thus, an outlier molecule differs from any other molecule in the set at the predefined outlier cutoff. As a result, 32,971 non-outlier and 17,584 outlier molecules were retrieved at an outlier cutoff of 0.5. Utilizing the UMAP-based clustering method, 7 clusters were derived for non-outlier molecules, with the outlier set forming its distinct cluster. It is important to note that the outlier detection method and clustering algorithm were applied to the entire set of molecules in the NCI-60 panel and not on a per-cell line basis. The clustering output includes NSC ID, SMILES representations, and assigned clusters, facilitating mapping to individual cell lines. Each of the 50,555 molecules receives a cluster ID from 1 to 7 or is marked as Outlier. Combining NSC-ID with the cluster ID allows mapping clustering results per-cell line, resulting in eight distinct clusters for each cell line.

Finally, clusters 1 to 7 (non-outlier set) are used to build and evaluate a CP model using the LDMO cross-validation strategy. In this stage, at each fold, 5 clusters are used as the proper training set, 1 as the calibration set, and 1 as the test set. In the second phase, the non-outlier set is used to retrain the models (6 clusters as the proper training set and 1 as the calibration set), which are subsequently evaluated on the outlier set.

Performance metrics

The evaluation of the CP models involves the use of regression machine learning metrics, which include the Root Mean Squared Error (RMSE), Pearson correlation (Rp), and Spearman correlation (Rs). The RMSE measures the average magnitude of errors between predicted and actual values, where lower RMSE values indicate better model accuracy, with 0 being a perfect fit. The Rp quantifies the linear relationship between predicted and actual values. It ranges from -1 to 1, indicating perfect negative and positive correlations. The Rs assesses the monotonic relationship between predicted and actual values. It ranges from -1 to 1, with values close to 1 or -1 implying a strong monotonic correlation, while 0 indicates no monotonic relationship. In addition, we employ efficiency and validity metrics from the CP framework to complement our evaluation(Bosc et al. 2019; Norinder et al. 2014). Validity measures the reliability of the predictions, indicating that at a confidence level of 80% (0.8), the predictor will include the actual yi value within its prediction intervals at least 80% of the time, assuming CP is well-calibrated. Efficiency, on the other hand, measures the specificity of the predictions and is related to the interval size (uncertainty). A higher confidence level reduces efficiency, resulting in larger predicted intervals, i.e., predictions with higher uncertainty.

Learning algorithms

Linear regression (LR) is here used as a performance baseline. LR fits a linear model to minimize the difference between observed and predicted pGI_{50} values across training instances using the features as independent variables(Su et al. 2012).

Random Forest (RF) is an algorithm that generates an ensemble of diverse tree predictors which collectively constitute a nonlinear model(Breiman 2001). In regression, RF predicts the label of an instance as the average of the individual tree predictions.

The eXtreme Gradient Boosting algorithm(Chen and Guestrin 2016), XGBoost or XGB for short, weights the prediction of each generated tree by how well it individually match the data. XGB was designed to optimize training speed and generalization performance.

Here we employed the scikit-learn implementations(Pedregosa et al. 2011) of the LR, RF and XGB algorithms to build regression models for each training dataset.

Hyperparameter tuning

Hyperparameter tuning is a pivotal strategy to augment the efficacy of ML algorithms, especially when applied to complex problem domains such as the prediction of inhibitory activity of molecules on a cancer cell line. In our previous work(Hernández-Hernández et al. 2022), we performed hyperparameter tuning to find the best combination of the most critical hyperparameters for the RF and XGB algorithms using a systematic grid search approach. More concretely, we did this for each of the 60 datasets of labelled molecules, one per cell line, adhering to an 80/10/10 partitioning scheme for training, validation, and test sets. In our current work, we are utilizing the optimal hyperparameter values that we have identified in our previous work, as for each cell line multiple models are trained as a part of the LDMO CV and outlier-based partitions.

Acknowledgments

This work was supported by CONAHCYT [S.H-H.] and the Wolfson Foundation and the Royal Society for a Royal Society Wolfson Fellowship [P.J.B.].

Author information

Contributions

P.J.B. conceived and led the study. S.H-H. wrote the code to run the experiments. All authors analysed the results from the experiments. P.J.B. wrote the paper with assistance from S.H-H. and Q.G.

Ethics declarations

Competing interests

The authors declare no competing interests.

Data availability

The data used in our study are publicly available in NCI-60 (https://wiki.nci.nih.gov/display/NCIDTPdata/NCI-60+Growth+Inhibition+Data), whereas all the partitions made for this study can be found at GitHub (https://github.com/Sahet11/nci60_clustering).

Code availability

The source code to allow replication of the results of this study is freely available at GitHub (https://github.com/Sahet11/STL_LDMO_project).

References

Alvarsson, J., Arvidsson McShane, S., Norinder, U., & Spjuth, O. (2021, January 1). Predicting With Confidence: Using Conformal Prediction in Drug Discovery. *Journal of Pharmaceutical Sciences*. Elsevier B.V. https://doi.org/10.1016/j.xphs.2020.09.055

Ammad-ud-din, M., Georgii, E., Gönen, M., Laitinen, T., Kallioniemi, O., Wennerberg, K., et

al. (2014). Integrative and personalized QSAR analysis in cancer by kernelized Bayesian matrix factorization. *Journal of chemical information and modeling*, *54*(8), 2347–59. https://doi.org/10.1021/ci500152b

Ballester, P. J. (2023). The AI revolution in chemistry is not that far away. *Nature*, *624*(7991), 252–252. https://doi.org/10.1038/d41586-023-03948-w

Ballester, P. J., & Carmona, J. (2021). Artificial intelligence for the next generation of precision oncology. *npj Precision Oncology*, *5*(1), 1–3. https://doi.org/10.1038/s41698-021-00216-w

Ballester, P. J., Stevens, R., Haibe-Kains, B., Huang, R. S., & Aittokallio, T. (2021). Artificial intelligence for drug response prediction in disease models. *Briefings In Bioinformatics*, bbab450. https://doi.org/10.1093/bib/bbab450

Bosc, N., Atkinson, F., Felix, E., Gaulton, A., Hersey, A., & Leach, A. R. (2019). Large scale comparison of QSAR and conformal prediction methods and their applications in drug discovery. *Journal of Cheminformatics*, *11*(1). https://doi.org/10.1186/s13321-018-0325-4

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16* (pp. 785–794). New York, New York, USA: ACM Press. https://doi.org/10.1145/2939672.2939785

Chiu, Y.-C., Chen, H.-I. H., Gorthi, A., Mostavi, M., Zheng, S., Huang, Y., & Chen, Y. (2019). Deep learning of pharmacogenomics resources: moving towards precision oncology. *Briefings in bioinformatics*. https://doi.org/10.1093/bib/bbz144

Cortés-Ciriano, I., van Westen, G. J. P., Bouvier, G., Nilges, M., Overington, J. P., Bender, A., & Malliavin, T. E. (2016). Improved large-scale prediction of growth inhibition patterns using the NCI60 cancer cell line panel. *Bioinformatics (Oxford, England)*, *32*(1), 85–95. https://doi.org/10.1093/bioinformatics/btv529

De Niz, C., Rahman, R., Zhao, X., & Pal, R. (2016). Algorithms for Drug Sensitivity Prediction. *Algorithms*, *9*(4), 77. https://doi.org/10.3390/a9040077

Diehl, F., Li, M., Dressman, D., He, Y., Shen, D., Szabo, S., et al. (2005). Detection and quantification of mutations in the plasma of patients with colorectal tumors. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(45), 16368–73. https://doi.org/10.1073/pnas.0507904102

Freedman, D. A. (2009). A simple regression equation has on the right hand side an intercept and an explanatory variable with a slope coefficient. A multiple regression equation has two or more explanatory variables on the right hand side, each with its own slope coefficient. *Statistical Models Theory and Practice*, *26*.

Fresnais, L., & Ballester, P. J. (2020). The impact of compound library size on the performance of scoring functions for structure-based virtual screening. *Briefings in Bioinformatics*, bbaa095. https://doi.org/10.1093/bib/bbaa095

Hernández-Hernández, S., & Ballester, P. J. (2023). On the Best Way to Cluster NCI-60 Molecules. *Biomolecules*, *13*(3), 498. https://doi.org/10.3390/biom13030498

Hernández-Hernández, S., Vishwakarma, S., & Ballester, P. J. (2022). Conformal prediction of small-molecule drug resistance in cancer cell lines. *Proceedings of Machine Learning Research*, *179*, 1–17. https://proceedings.mlr.press/v179/hernandez-hernandez22a.html. Accessed 25 June 2023

Ho, T. K. (1995). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition* (Vol. 1, pp. 278–282).

Hughes, J. P., Rees, S. S., Kalindjian, S. B., & Philpott, K. L. (2011, March). Principles of early drug discovery. *British Journal of Pharmacology*. Wiley-Blackwell. https://doi.org/10.1111/j.1476-5381.2010.01127.x

Ianevski, A., Lahtela, J., Javarappa, K. K., Sergeev, P., Ghimire, B. R., Gautam, P., et al. (2021). Patient-tailored design for selective co-inhibition of leukemic cell subpopulations. *Science Advances*, *7*(8), eabe4038. https://doi.org/10.1126/sciadv.abe4038

Kumar, R., Chaudhary, K., Singla, D., Gautam, A., & Raghava, G. P. S. (2014). Designing of

promiscuous inhibitors against pancreatic cancer cell lines. *Scientific reports*, *4*, 4668. https://doi.org/10.1038/srep04668

Landrum, G. (2016). RDKit: Open-Source Cheminformatics Software.

Lin, A., Giuliano, C. J., Palladino, A., John, K. M., Abramowicz, C., Yuan, M. Lou, et al. (2019). Off-target toxicity is a common mechanism of action of cancer drugs undergoing clinical trials. *Science Translational Medicine*, *11*(509), 8412. https://doi.org/10.1126/scitranslmed.aaw8412

Liu, G., Catacutan, D. B., Rathod, K., Swanson, K., Jin, W., Mohammed, J. C., et al. (2023). Deep learning-guided discovery of an antibiotic targeting Acinetobacter baumannii. *Nature Chemical Biology*, 1–9. https://doi.org/10.1038/s41589-023-01349-8

Menden, M. P., Iorio, F., Garnett, M., McDermott, U., Benes, C. H., Ballester, P. J., & Saez-Rodriguez, J. (2013). Machine Learning Prediction of Cancer Cell Sensitivity to Drugs Based on Genomic and Chemical Properties. *PLoS ONE*, *8*(4), e61318. https://doi.org/10.1371/journal.pone.0061318

Moffat, J. G., Vincent, F., Lee, J. A., Eder, J., & Prunotto, M. (2017, July 31). Opportunities and challenges in phenotypic drug discovery: An industry perspective. *Nature Reviews Drug Discovery*. Nature Publishing Group. https://doi.org/10.1038/nrd.2017.111

Naulaerts, S., Menden, M. P., & Ballester, P. J. (2020). Concise polygenic models for cancer-specific identification of drug-sensitive tumors from their multi-omics profiles. *Biomolecules*, *10*(6), 1–21. https://doi.org/10.3390/BIOM10060963

Niepel, M., Hafner, M., Mills, C. E., Subramanian, K., Williams, E. H., Chung, M., et al. (2019). A Multi-center Study on the Reproducibility of Drug-Response Assays in Mammalian Cell Lines. *Cell Systems*, *9*(1), 35-48.e5. https://doi.org/10.1016/j.cels.2019.06.005

Norinder, U., Carlsson, L., Boyer, S., & Eklund, M. (2014). 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. https://doi.org/10.1021/ci5001168

Partin, A., Brettin, T. S., Zhu, Y., Narykov, O., Clyde, A., Overbeek, J., & Stevens, R. L. (2023, February 15). Deep learning methods for drug response prediction in cancer: Predominant and emerging trends. *Frontiers in Medicine*. Frontiers Media S.A. https://doi.org/10.3389/fmed.2023.1086097

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(Oct), 2825–2830.

Peón, A., Dang, C. C., & Ballester, P. J. (2016). How Reliable Are Ligand-Centric Methods for Target Fishing? *Frontiers in Chemistry*, *4*, 15. https://doi.org/10.3389/fchem.2016.00015

Piyawajanusorn, C., Nguyen, L. C., Ghislat, G., & Ballester, P. J. (2021). A gentle introduction to understanding preclinical data for cancer pharmaco-omic modeling. *Briefings In Bioinformatics*, bbab312. https://doi.org/10.1093/bib/bbab312

Quancard, J., Bach, A., Cox, B., Craft, R., Finsinger, D., Guéret, S. M., et al. (2021). The European Federation for Medicinal Chemistry and Chemical Biology (EFMC) Best Practice Initiative: Phenotypic Drug Discovery. *ChemMedChem*, *16*(11), 1737–1740. https://doi.org/10.1002/cmdc.202100041

Quancard, J., Cox, B., Finsinger, D., Guéret, S. M., Hartung, I. V., Koolman, H. F., et al. (2020). The European Federation for Medicinal Chemistry (EFMC) Best Practice Initiative: Validating Chemical Probes. *ChemMedChem*, *15*(24), 2388–2390. https://doi.org/10.1002/cmdc.202000597

Sadri, A. (2023). Is Target-Based Drug Discovery Efficient? Discovery and “Off-Target” Mechanisms of All Drugs. *Journal of Medicinal Chemistry*. https://doi.org/10.1021/acs.jmedchem.2c01737

Shoemaker, R. H. (2006). The NCI60 human tumour cell line anticancer drug screen. In *Nature Reviews Cancer* (Vol. 6, pp. 813–823). Nat Rev Cancer. https://doi.org/10.1038/nrc1951

Siramshetty, V. B., Shah, P., Kerns, E., Nguyen, K., Yu, K. R., Kabir, M., et al. (2020).

Retrospective assessment of rat liver microsomal stability at NCATS: data and QSAR models. *Scientific Reports*, *10*(1). https://doi.org/10.1038/s41598-020-77327-0

Smer-Barreto, V., Quintanilla, A., Elliott, R. J. R., Dawson, J. C., Sun, J., Campa, V. M., et al. (2023). Discovery of senolytics using machine learning. *Nature Communications*, *14*(1), 1–15. https://doi.org/10.1038/s41467-023-39120-1

Stokes, J. M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N. M., et al. (2020). A Deep Learning Approach to Antibiotic Discovery. *Cell*, *180*(4), 688-702.e13. https://doi.org/10.1016/j.cell.2020.01.021

Su, X., Yan, X., & Tsai, C. (2012). Linear regression. *WIREs Computational Statistics*, *4*(3), 275–294. https://doi.org/10.1002/wics.1198

Tran-Nguyen, V. K., Junaid, M., Simeon, S., & Ballester, P. J. (2023). A practical guide to machine-learning scoring for structure-based virtual screening. *Nature Protocols*, *18*(11), 3460–3511. https://doi.org/10.1038/s41596-023-00885-w

Vincent, F., Nueda, A., Lee, J., Schenone, M., Prunotto, M., & Mercola, M. (2022). Phenotypic drug discovery: recent successes, lessons learned and new directions. *Nature Reviews Drug Discovery*, 1–16. https://doi.org/10.1038/s41573-022-00472-w

Wagner, B. K. (2016). The resurgence of phenotypic screening in drug discovery and development. *Expert Opinion on Drug Discovery*, *11*(2), 121–125. https://doi.org/10.1517/17460441.2016.1122589

Wong, F., Zheng, E. J., Valeri, J. A., Donghia, N. M., Anahtar, M. N., Omori, S., et al. (2023). Discovery of a structural class of antibiotics with explainable deep learning. *Nature*, 1–9. https://doi.org/10.1038/s41586-023-06887-8