NNT : 2024AIXM0221

THÈSE DE DOCTORAT

Soutenue à Aix-Marseille Université
le 18 juillet 2024 par

Saiveth HERNANDEZ HERNANDEZ

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

Prédiction conforme de la résistance aux médicaments à petites molécules dans les lignées cellulaires cancéreuses

Discipline
Biologie Santé

Spécialité
Bioinformatique et Génomique

École doctorale
ED 62 Sciences de la Vie et de la Santé

Laboratoire/Partenaires de recherche
Centre de Recherche en Cancérologie de Marseille (CRCM)

Financement
Consejo Nacional de Humanidades, Ciencia y Tecnologia (CONAHCYT), Mexico

Composition du jury

Delphine FRADIN CRCI2NA, Nantes, France	Rapporteure
Mohamed ELATI University of Lille, France	Rapporteur
Bianca HABERMANN IBDM, Marseille, France	Présidente du jury et Examinatrice
Pedro BALLESTER CRCM, Marseille, France, and Imperial College London, London, UK	Directeur de thèse

CRCM
Centre de Recherche en Cancérologie de Marseille

Affidavit

I, undersigned, Saiveth HERNANDEZ HERNANDEZ, hereby declare that the work presented in this manuscript is my own work, carried out under the scientific supervision of Dr. Pedro J. BALLESTER, in accordance with the principles of honesty, integrity and responsibility inherent to the research mission. The research work and the writing of this manuscript have been carried out in compliance with both the French national charter for Research Integrity and the Aix-Marseille University Charter on the fight against plagiarism.

This work has not been submitted previously either in this country or in another country in the same or in a similar version to any other examination body.

In Marseille, France, July 18, 2024



This work is licensed under Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International Public License

List of publications and conference participation

List of publications produced as part of the thesis project:

1. Saiveth Hernández-Hernández, Sachin Vishwakarma, and Pedro Ballester. "Conformal prediction of small-molecule drug resistance in cancer cell lines". In: *Conformal and Probabilistic Prediction with Applications*. PMLR. 2022, pp. 92–108,
2. Saiveth Hernández-Hernández and Pedro J Ballester. "On the Best Way to Cluster NCI-60 Molecules". In: *Biomolecules* 13.3 (2023), p. 498, https://doi.org/10.3390/biom13030498
3. Saiveth Hernández-Hernández, Qianrong Guo, and Pedro J Ballester. "Conformal prediction of molecule-induced cancer cell growth inhibition challenged by strong distribution shifts". In: *Preprint from bioRxiv* (2024), pp. 1–16, https://doi.org/10.1101/2024.03.15.585269
4. Saiveth Hernandez-Hernandez, and Pedro J. Ballester, "On the Best Way to use Conformal Prediction for Prospective Purposes". *In preparation*

Participation in conferences and training courses during the thesis period:

1. INSERM - Atelier "Introduction à l'Apprentissage automatique de la Biologie à la Santé" (11 octobre 2021 - 13 octobre 2021) Inserm - Bordeaux (Inserm workshop 266, online course).
2. 3rd edition of the PhD Day (May 2021). Oral Presentation
3. 11th Symposium on Conformal and Probabilistic Prediction with Applications (August 24-26, 2022). Brighton, UK. Oral Presentation
4. 4th edition of the PhD Day "Popularization of science" (September 12, 2022). Poster presentation
5. COLLOQUE Ecole Doctorale Sciences de la Vie et de la Santé (EDSVS) 2023 (30 mai 2023 - 31 mai 2023), Marseille, France.

Abstract

Phenotypic Drug Design (PDD) is a promising approach that can revolutionize the time-consuming and expensive drug discovery process, especially when combined with artificial intelligence methods. PDD has re-emerged as an effective approach to drug discovery by identifying compounds that can produce a desired therapeutic effect on the disease phenotype. The availability of a vast amount of preclinical data has made machine learning (ML)-based models highly effective in drug discovery. Notably, the NCI-60 panel, which has been used since 1990, exemplifies this synergy by identifying compounds with promising anticancer properties. Despite these advances, ML may encounter challenges in predicting the efficacy of novel compounds and assessing the reliability of individual molecules.

When employing ML for critical decision-making, accurately quantifying the confidence in predictions becomes imperative. Conformal Prediction (CP) emerges as a robust method to quantify uncertainty, applicable across various ML algorithms. CP operates on the principle of predicting that a new instance will exhibit a label similar to previous instances (calibration set) in a specific manner defined by a calibration function. The degree of similarity between the new instance and the calibration set is quantified through a calibration score, which in turn serves to estimate the uncertainty associated with the prediction.

This thesis project aims to develop advanced predictive models by harnessing the synergy between ML techniques and CP frameworks to identify optimal drug candidates for treating various types of cancer. By leveraging ML algorithms and incorporating CP to quantify prediction uncertainties, our approach aims to prioritize compounds exhibiting high potency and reliable prediction outcomes. This will help us to prioritize the most promising candidates for further preclinical evaluation.

This thesis project establishes a proof-of-concept framework, which provides a strong basis for evaluating the efficacy of new compounds in inhibiting or killing tumor cells. We began with a baseline CP model to measure its accuracy in predicting inhibitory activity of molecules in cancer cell lines. Through systematic optimization, we fine-tuned key parameters of these models to find an optimal balance between predictive performance and computational cost. Subsequently, we rigorously assessed the optimized CP model's efficacy under real-world conditions through retrospective and prospective evaluations.

The results demonstrate that the optimized CP models exhibit superior generalization than their underlying ML models alone. This is achieved for a range of test sets across

cancer cell lines, including those comprising molecules dissimilar from those in the training set. This thesis shows that CP can be employed to enhance ML predictions on cancer cell lines by adding a new dimension to PDD decision-making. Compounds can now be selected for in vitro tests not only because they are predicted to have potent activity, but also because those predictions are estimated to be reliable.

Keywords: Phenotypic Drug Design, artificial intelligence, conformal prediction, uncertainty quantification, small molecules

Résumé

La conception phénotypique de médicaments (PDD) représente une approche prometteuse susceptible de révolutionner le processus coûteux et chronophage de découverte de médicaments, notamment lorsqu'elle est associée à des méthodes d'intelligence artificielle. PDD est réapparue comme une approche efficace de découverte de médicaments en identifiant des composés capables de produire un effet thérapeutique désiré sur le phénotype de la maladie. L'abondance des données précliniques disponibles aujourd'hui rend les modèles basés sur l'apprentissage automatique particulièrement efficaces pour la découverte de médicaments. Par exemple, le panel NCI-60, utilisé depuis 1990, illustre cette synergie en identifiant des composés aux propriétés anticancéreuses prometteuses. Malgré ces progrès, l'apprentissage automatique peut rencontrer des difficultés pour prédire l'efficacité de nouveaux composés et évaluer la fiabilité de molécules individuelles

Dans le contexte de l'utilisation de l'apprentissage automatique (ML) pour des prises de décisions critiques, il est crucial de quantifier avec précision la confiance associée aux prédictions. La prédiction conforme (CP) se présente comme une méthode robuste pour quantifier l'incertitude, applicable à différents algorithmes de ML. Le principe de la prédiction conforme repose sur l'idée de prédire qu'une nouvelle instance aura une étiquette similaire à celle des instances précédentes (“ ensemble de calibration ”) de manière spécifique définie par une fonction de calibration. Le degré de similitude entre la nouvelle instance et l'ensemble de calibration est quantifié par un score de calibration, permettant ainsi d'estimer l'incertitude associée à la prédiction.

Ce projet de thèse vise à développer des modèles prédictifs avancés en exploitant la synergie entre les techniques de ML et les cadres de CP pour identifier les médicaments candidats optimaux pour le traitement de divers types de cancer. En intégrant les algorithmes de ML avec les CP pour quantifier les incertitudes de prédiction, notre approche vise à prioriser les composés présentant une forte efficacité et des prédictions fiables. Cette méthodologie nous permettra de sélectionner les candidats les plus prometteurs en vue d'une évaluation préclinique approfondie.

Ce projet de thèse établit un cadre de preuve de concept, fournissant une base solide pour évaluer l'efficacité des nouveaux composés dans l'inhibition ou la destruction des cellules tumorales. Nous avons débuté avec un modèle CP de base afin de mesurer sa précision dans la prédiction de l'activité inhibitrice des molécules dans les lignées cellulaires cancéreuses. Par le biais d'une optimisation systématique, nous avons

affiné les paramètres clés de ces modèles afin d'atteindre un équilibre optimal entre la performance prédictive et le coût computationnel. Par la suite, nous avons rigoureusement évalué l'efficacité du modèle CP optimisé dans des conditions réelles, en menant des évaluations rétrospectives et prospectives.

Les résultats démontrent que les modèles CP optimisés présentent une généralisation supérieure à celle de leurs modèles ML sous-jacents seuls. Cela est obtenu pour une gamme d'ensembles de tests à travers des lignées cellulaires cancéreuses, y compris ceux comprenant des molécules différentes de celles de l'ensemble d'apprentissage. Cette thèse montre que le CP peut être utilisé pour améliorer les prédictions de ML sur les lignées cellulaires cancéreuses en ajoutant une nouvelle dimension à la prise de décision en PDD. Les composés peuvent désormais être sélectionnés pour des tests in vitro non seulement parce qu'ils sont prédits comme ayant une activité puissante, mais aussi parce que ces prédictions sont estimées fiables

Mots clés: conception phénotypique de médicaments, intelligence artificielle, prédiction conforme, quantification de l'incertitude, petites molécules

Acknowledgements

The pursuit of a dream led me to join an exceptional research team, Machine Learning for Precision Oncology and Drug Design, at CRCM and the Aix-Marseille Universite four years ago. This thesis represents the culmination of years of effort and perseverance. For that, I am deeply grateful for the support of individuals and organizations who helped me along the way.

First, I would like to express my sincere gratitude to my thesis supervisor, Dr. Pedro J. Ballester, for their invaluable guidance and support. His expertise, insightful feedback, and patience have played a vital role in shaping this research.

To my family, I owe a debt of gratitude that words cannot fully express. Your endless love, encouragement, and understanding throughout this academic journey have been my rock. To Migue, Bris, Anita, Annet, Rodri and Pablo, thank you for being my cornerstones and for always being by my side. I especially want to thank my parents, Angel and Filomena, for their unwavering love, encouragement, and sacrifices. Thank you for your complete support in my decision to move away from Mexico to pursue this dream. Your confidence in my abilities has been my greatest motivation. This achievement is as much yours as it is mine.

To all our immediate family, in-laws, close friends, and the compassionate hearts whose generosity breathed life into my father's journey, your kindness was a symphony of hope amidst the silence of uncertainty. Your help has allowed me to share and celebrate this milestone with my entire family, making this accomplishment even more special and meaningful. Though I may not know all your names, I will forever cherish you and be profoundly grateful for your boundless generosity.

I want to express my heartfelt gratitude to Yves, Stephanie, and Isabelle who have become my second family during this academic endeavor. Yves, from the first moment we met, you have consistently looked out for me, offering support and guidance that has meant the world to me. Stephanie, I am deeply thankful for your exceptional kindness and generosity. You not only provided me with a place I can call home but also offered me a profound sense of belonging and stability that made all the difference in pursuing this thesis. Isabelle, thank you for all the wonderful yoga teachings and relaxation techniques you have shared with me. Your valuable advice has truly made a positive impact on me, and I will always hold it close to my heart. I will always be thankful to all of you for that and so much more.

I am also grateful to my friends, who have made this experience more enjoyable and memorable. Chucho, I am profoundly grateful for the care and support you extended to my family during those challenging times. Your watchful presence brought light into our darkest moments. Derly, from our childhood days to now, you have stood by my side, always believing in me. My lifelong friend, thank you for being there through every step of this journey. Ivette, thank you for your presence when I needed you most, your friendship has been invaluable in

keeping me moving forward. Thanks also to the ephemeral souls whose presence adorned my path with brief yet profound grace.

I am grateful for the support and trust extended to me by two esteemed people from Mexico, Mr. Erick García and Mr. Juan Manuel Carreño, former colleagues and mentors at the MRCI group. Their encouragement and belief in my decision to pursue studies in France were invaluable. I would also like to acknowledge that the support extended by Mr. Juan Manuel Carreño and Mr. Erick García is deeply rooted in the support of Mr. José Antonio Pérez and Mrs. Cecilia Pérez, the founding partners of the MRCI group. To them, I extend my sincere gratitude and deepest appreciation for welcoming me into the MRCI family from a distance.

I extend my sincere appreciation to the people at the CRCM and the Aix-Marseille Universite for providing the necessary resources and facilities for conducting my experiments and analysis. Thank you to CONAHCYT for funding this project. Special thanks to my colleagues in Marseille and London for their collaborative spirit and insightful discussions, which significantly contributed to the success of this project. I especially thank Qianrong Guo for her valuable and insightful comments. Thanks to Scarlett, Vivi, Eugine, Benoit, and Sachin for their company and friendship during my stay at the CRCM. Thank you to my jury members for taking the time to review this work.

I am truly grateful to everyone mentioned above and to those who have supported me in ways beyond words. Thank you.

To my family
Ángel, Filomena, Miguel Ángel, Bris,
Anita, Annet, Rodri, Pablo and the "Rockys"

Contents

Affidavit 2

List of publications and conference participation 3

Abstract 4

Résumé 6

Acknowledgements 8

Contents 11

List of Figures 14

List of Tables 21

Liste des acronymes 23

1. Introduction 25
- 1.1. An overview of key project areas . . . 26
- 1.2. Cancer Overview . . . 27
  - 1.2.1. Introduction to cancer . . . 27
  - 1.2.2. Diagnosis and treatment options . . . 31
- 1.3. Drug design and drug discovery . . . 33
  - 1.3.1. Introduction to Drug Discovery and Drug Development . . . 33
  - 1.3.2. Phases of Drug Discovery . . . 35
  - 1.3.3. Approaches to drug discovery . . . 37
  - 1.3.4. Ligand-Based Drug Design . . . 39
- 1.4. Uncertainty quantification in drug design . . . 40
  - 1.4.1. Definition and sources of uncertainty . . . 40
  - 1.4.2. Uncertainty types . . . 42
  - 1.4.3. Uncertainty Quantification methods . . . 43
- 1.5. Artificial intelligence in drug discovery and development . . . 45
  - 1.5.1. A comprehensive overview of Artificial Intelligence . . . 45
  - 1.5.2. Brief introduction to Machine Learning . . . 46
  - 1.5.3. Conformal Prediction . . . 47
- 1.6. Scope and Structure . . . 52

2. Research Framework **54**
2.1. Research framework overview . . . 54
2.2. Motivation and problem statement . . . 55
2.3. Literature review of the latest advances in Conformal Prediction . . . 56
2.3.1. Prominent studies using Conformal Prediction in Drug Design . 56
2.3.2. Findings from the Literature Review . . . 63
2.4. Research Questions . . . 64
2.5. Objectives . . . 65

3. Materials and Methods **66**
3.1. An overview of the development of this project . . . 66
3.2. NCI-60 panel . . . 67
3.2.1. Database description and preprocessing . . . 67
3.2.2. Molecular descriptors . . . 70
3.3. Machine Learning algorithms . . . 73
3.4. Model Evaluation . . . 74
3.4.1. Evaluation Metrics for Machine Learning Models . . . 74
3.4.2. Evaluation Metrics for Conformal Prediction Models . . . 75

4. Results and Discussion **77**
4.1. Applying CP to predict the inhibitory activity of molecules . . . 78
4.2. Baseline model investigation . . . 80
4.2.1. Validity and efficiency hold in CP models as calibration set size is reduced . . . 81
4.2.2. Accuracy assessment of CP-based models . . . 82
4.2.3. CP baseline model considerations . . . 84
4.3. Baseline model optimization . . . 84
4.3.1. Hyperparameter tuning . . . 84
4.3.2. Increasing MFP size leads to more accurate predictions . . . 86
4.3.3. Removing molecules with measured $pGI_{50} = 4$ decreases the model performance . . . 88
4.3.4. The importance of optimal data partitioning in cross-validation strategies . . . 90
4.4. Data preparation for a real-world scenario . . . 91
4.4.1. Leveraging molecular clustering to improve model validation . . 91
4.4.2. Relevance of outlier molecules definition . . . 94
4.4.3. Outlier detection to create an external validation set . . . 95
4.4.4. Understanding the impact of different clustering algorithms on molecular clustering . . . 96
4.4.5. Analyzing molecular clustering quality with Tanimoto similarity matrix . . . 98
4.5. Integration of the previous steps in a real-world scenario . . . 99
4.5.1. Retrospective evaluation of CP models on dissimilar test sets . . 99
4.5.2. Comparing CP-based models for Drug Design . . . 110

4.5.3. Prospective evaluation of CP models to guide decision-making . 111

5. Conclusions and future work **114**
5.1. Conclusions 114
5.1.1. General conclusions 114
5.1.2. On the accuracy of CP models to predict pGI_{50} values for different compounds in cancer cell lines 115
5.1.3. On the impact of data split and molecular representation 116
5.1.4. On the CP model generalization to predict pGI_{50} values for dissimilar compounds 117
5.1.5. On the integration of CP into decision-making processes to support prospective drug screening 117
5.2. Future work 118

Bibliography **119**

ANNEXES **130**
A. COPA paper 130
B. Clustering paper 148
C. LDMO-CCP paper 168

List of Figures

1.1. **Relationship of the four main topics of this thesis work: cancer, drug design, uncertainty quantification, and artificial intelligence (AI).** a) The Introduction chapter begins with an overview of cancer, its causes, and treatments like chemotherapy. b) Drug discovery, a crucial process in developing chemotherapy drugs, often involves virtual screening (VS). c)-d) Advancements in computational methodologies have led to the integration of AI into VS approaches. Methods like conformal prediction enhance VS by providing uncertainty quantification. This ensures reliable predictions in drug design, improving the selection of potential therapies. 26

1.2. **Comprehensive overview of the crucial elements involved in cancer development.** a)Disruptions in normal cell growth lead to uncontrolled multiplication and invasion of surrounding tissues, resulting in cancer. b) Cancer cells can break away from the primary tumor and spread to other areas of the body, which is referred to as metastasis. c) DNA changes that cause cancer usually occur in specific sections of DNA known as genes. d) Oncogenes and e) tumor suppressor genes, which play essential roles in normal cells, are responsible for driving cancer. f) Certain risk factors, including smoking, exposure to carcinogens, and heredity, among others, can elevate the chances of developing cancer. The images in this figure are from the National Cancer Institute's website (`https://www.cancer.gov`). They are in the public domain and not subject to copyright restrictions. 28

1.3. **Cancer types studied in this work, grouped by body systems.** These types of cancer develop in the digestive system (colon), urinary system (renal), breast, reproductive system (ovarian and prostate), lungs (non-small cell lung cancer), skin (melanoma), blood and lymph system (leukemia), and central nervous system (CNS). Image created with (`BioRender.com`). 31

1.4. **The 2022 global cancer statistics cover 185 countries and 36 types.** a) Approximately 20 million new cancer cases and 9.7 million deaths worldwide were reported in 2022. b) Among the newly reported cases, lung cancer was found to be the most common type in males, accounting for 15.3% of all cases. On the other hand, breast cancer was the most prevalent type in females, accounting for 23.8% of all cases. 32

1.5. **Drug discovery and development process**. This process involves several stages, from left to right: drug discovery, preclinical research, clinical trials (Phases I-III), and Regulatory review (Phase IV Post-marketing surveillance). This process typically spans 12-15 years, with costs ranging from $1.8 and $2.6 billion. Only a small fraction of compounds tested advance to approval, illustrating the rigorous nature of drug development in ensuring safety and efficacy for patient use. It is essential to mention that the number of molecules progressing between stages in drug discovery varies according to the work reported. The number reported in this figure serves as an illustrative example highlighting this process's challenging and rigorous nature [56, 81]. 34

1.6. **Overview of Conformal Prediction methodology and key components.** a) The CP framework predicts a new instance's label by assessing its (dis)similarity to old instances using a calibration function. In the so-called normalized CP, two models are built using the instances in the proper training set, an underlying model $h(x)$ and an error model $g(x)$. b) Using the underlying model, the error model, and the calibration score, CP generates a range of predictions with varying degrees of confidence to ensure the actual label falls within the prediction region. c) Predictions are evaluated by calculating efficiency (How informative are the predictions?) and validity (How reliable are the predictions?). . . . 49

2.1. **Overview of the Research Framework chapter**. This chapter lays the foundation for this thesis project. a) We aim to develop machine learning (ML) and conformal prediction (CP) models to predict potent molecules for various cancer types. b) Next, we conduct a state-of-the-art review of existing methods of Conformal Prediction (CP) applied to drug design. The literature review helped us identify research gaps. c) These research gaps led to the proposal of four research questions. and d) objectives. . 55

3.1. **Overview of the development of this thesis project**. This project focuses on the NCI-60 panel, a dataset for discovering compounds with potential anticancer activity. a) The first step is data extraction and data preprocessing. b) Compounds within the NCI-60 panel are represented using either Morgan fingerprints or physicochemical descriptors. c) These descriptors are inputs for machine learning and conformal prediction models. d) The model evaluation stage comprises regression, classification, and conformal prediction metrics. 67

3.2. **NCI-60 Growth Inhibition Bulk Data description.** a) The dataset utilized in this study is from the NCI-60 panel, which contains 60 distinct cell lines representing nine different types of cancer. b) Distribution of the pGI_{50} measurements in the molecules across the 60 cell lines for a total of 2,707,434 data points, where potent molecules ($pGI_{50} \geq 6$) are rather scarce. c) Distribution of the number of unique molecules tested per cell line. A color code is assigned to show the number of potent molecules ($pGI_{50} \geq 6$) per cell line. 68
3.3. **Graphical representation of CP evaluation.** a) Efficiency is the size of the predicted interval, where smaller intervals are more efficient. b) Validity refers to the number of errors obtained compared to the number requested. If these values coincide, it is said to be a calibrated model. Otherwise, the model can make more errors (red line) or fewer errors (blue line). c) Predictions are categorized into CP-valid if the predicted interval contains the true value and CP-invalid otherwise. 76

4.1. **Applying CP to predict the inhibitory activity of molecules in cancer cell lines of the NCI-60 panel**. a) CP baseline models were developed to set a benchmark for future investigations. This preliminary study focused on predicting the inhibitory activity of molecules using normalized CP models trained on randomly partitioned datasets. b) In the second stage, we focus on optimizing the performance of the baseline model, for which three primary aspects were explored. c) In the third stage, we prepare the data for a more realistic and rigorous evaluation, moving beyond random data separation. We investigated cluster analysis for model validation, also emphasizing outlier detection. d) The final stage aims to use the optimized CP models for practical use in a real-world drug design scenario, covering both retrospective and prospective evaluations. 79
4.2. **Initial research on the predictive accuracy of CP models for predicting inhibitory activity of molecules**. a) Molecules from the NCI-60 panel are represented using 256-bit MFPs, then used as input features for CP models. b) The dataset is randomly divided into training and test sets to build CP models. The training set is then subdivided into a proper training set and a calibration set. This step is repeated for each of the 60 data sets, one per cell line. A total of 1080 models, 60 cells × 6 CP modes × 3 data partitions, were built and evaluated in this initial study. c) The prediction performance is evaluated using metrics such as RMSE and Rp. We propose the terms "CP-valid" and "CP-invalid" predictions to complement the evaluation of the CP models. 81

4.3. **Grid search results** for a) RF and b) XGB. For both models, a color code represents the "n_estimator" hyperparameter. For the XGB, the markers represent the "cosample_bytree" value evaluated. The red arrow indicates the specific hyperparameter combination, resulting in the optimal model performance regarding the RMSE. 86
4.4. **Increasing the size of the Morgan fingerprints (MFP) leads to more accurate predictions across multiple evaluation metrics**. a) RMSE, b) Rp, and c) Rs values were computed between the observed and predicted pGI_{50} value, using the RF model as the underlying model. CP-valid predictions were computed at an 80% confidence level. Color code refers to the MFP size used. 87
4.5. **Removing molecules with measured $pGI_{50} = 4$ decreases the model performance, across multiple evaluation metrics.** a) RMSE, b) Rp, and c) Rs values were computed between the observed and predicted pGI_{50} value, using the RF model as the underlying model. CP-valid predictions were computed at an 80% confidence level. Color code refers to the set from which the compounds with measured $pGI_{50} = 4$ are eliminated. None: Non-removed, Train: Removed only from the training set, and Train-Test: Removed from both training and test sets. 89
4.6. **Data partition strategy is crucial as a CV does not significantly improve over a single-run approach.** a) RMSE, b) Rp, and c) Rs values were computed for the non-CP (left) or CP-valid (right) predictions. CP-valid predictions are obtained at a confidence level of 80%. The color code represents the type of model used, either ICP, ACP, 5-CCP, or 10-CCP. The results are from the RF model. 90
4.7. **The ACP and CCP approaches require more computational time to build the models**. The color code represents the type of model used, either ICP, ACP, 5-CCP, or 10-CCP. 91
4.8. **Leveraging molecular clustering to improve model validation.** a) We utilize MFPs as input for clustering molecules within the NCI-60 panel. b) We propose a novel method to identify outlier molecules and examine the impact of their removal on clustering quality. c) We compare the effectiveness of hierarchical, Taylor-Butina, and UMAP-based clustering techniques to determine the most suitable approach. d) The clustering results are evaluated using established metrics, and the quality of each cluster is assessed using an average similarity matrix. 93

4.9. **Description of the NCI-60 dataset after the outlier detection and clustering process.** a) pGI_{50} measurements from non-outlier and outlier molecules across the 60 cell lines for a total of 2,707,434 data points. Outlier molecules were retrieved at an outlier cutoff of 0.5. Potent molecules are rather scarce. b) UMAP-based clustering of the non-outlier molecules results in 7 clusters and provides the best set of clusters. These clusters, along with the outlier molecules, contain the 50,555 unique molecules in the NCI-60 panel. c) Number of unique molecules tested on each cell line broken down by cluster membership (same color code as in b). 98
4.10. **Summary of LDMO-CCP model development.** For each of the 60 cell lines: a) Generating the features for each molecule using MFPs and PCDs to encode their molecular characteristics. b) For each cell line and model, multiple Leave-dissimilar-molecules-out (LDMO)-CV validation runs are performed to evaluate how well the considered model can predict the cell activity of molecules with different chemistry. c) For each cell line, the underlying model is trained and tested on each fold of the non-outlier molecules. d) The standard machine learning metrics and the conformal prediction methodology metrics are used to evaluate CP model performance. Note that b) and c) correspond to the first stage of this study. e) In the second stage, the top two CP models identified during the first stage are subsequently re-trained using the non-outlier molecules. Following re-training, these models were evaluated using the outlier molecules to assess their performance on completely new chemistry. 101
4.11. **Trade-off between prediction uncertainty and measured error of CP models.** For each CP model, median error and uncertainty are computed using MFPs+PCDs as input descriptors at different confidence levels. The cell lines' performance is summarized as 7-fold-median uncertainty (y-axis). Thus, boxplots contain 60 points, representing the median uncertainty per cell line. The median of 7-fold-median error values is calculated across 60 cell lines to summarize the error performance. A single point is obtained that represents the median error of the entire group, and each boxplot is positioned along the x-axis based on the corresponding median error obtained. The vertical dotted lines represent the expected error at each confidence level. Color code refers to the four CP models ($h(x) - g(x)$) employed. Markers refer to the requested confidence level. 103

4.12. **Dissimilarity measure between the top 3 CP-valid molecules and those observed during training.** a) We initially pinpointed the top 100 CP-valid molecules with the highest predicted pGI_{50} values for the selected cell lines (see Table 1C for reference). These molecules are highlighted within a red oval region for easy identification. Subsequently, we narrowed our focus to the three molecules with the highest measured pGI_{50} values, which are denoted within a blue oval region in the figure. b) We find the closest training molecule for each of the three best CP-valid molecules by computing the Tanimoto similarity. We select the most similar molecule from six candidates (one per cluster), as the other five are more dissimilar. The most similar molecule from the training set to the test molecule is highlighted in bold. 105

4.13. **CP models can effectively generalize to outlier test sets and make low-uncertainty predictions.** Two LDMO-CCP models were compared using outlier and non-outlier datasets at different expected errors. The model's performance within the non-outlier dataset was evaluated through 7-fold-median scores and combined to find their median across all 60 cell lines. In the outlier dataset, a straightforward calculation of the median across the 60 cell lines leads to the single point encompassing the median uncertainty or error of the entire group of cell lines. a) A calibration plot illustrates the expected and actual error relationship. b) Error-Uncertainty plot to illustrate the relationship between the expected error and the associated uncertainty attributed to achieving the requested error. Color code refers to LDMO-CCP models, either RF-RF or XGB-RF. Markers refer to datasets that are either non-outliers or outliers. 107

4.14. **CP-valid molecules derived from models assessed on the outlier test set yielded higher hit rates.** A comparative analysis evaluated the performance of the two best LDMO-CCP models at an 80% confidence level. The a) hit rate and b) MCC distribution across 60 cell lines were analyzed. Statistical test results for the underlying models were displayed at the top of the boxplots. The color code stands for a key assigned to identify the model used and the type of prediction (nonCP or CP-valid). 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; ****:p ≤1.00e-04. 107

4.15. pGI_{50} **predictions of molecule-cell line pairs improve when using LDMO-CCP.** The pGI_{50} value observed and predicted in the best-predicted cell line, obtained by the RF-RF (top) and the XGB-RF (bottom) CP models. The color code indicates the test set molecules, either CP-valid (left) or CP-invalid (right). The vertical and horizontal dotted lines show the threshold for inactive and active molecules ($pGI_{50} \geq 6$). The number of FP, TP, TN, and FN is indicated for CP-valid and CP-invalid predictions. These values are used to compute the hit rate and MCC for each cell line. 108

4.16. Boxplots for the three criteria selected. From left to right, each boxplot represents the distribution of the top N molecules with the highest predicted pGI_{50}, the top N molecules with the highest predicted pGI_{50} (top M), and with the smallest interval; the top N molecules with the highest pGI_{50}-low (best lower endpoint of the predicted interval). The horizontal line on measured pGI_{50}=6 represents the threshold between active and inactive molecules: every above is true positive, below false positive. a) Top N=10, b) Top N=20, c) Top N=50, d) Top N=100. All metrics are calculated using only the top N molecules. Predictions are from the RF-RF model. 112
4.17. CP-enhanced predictions, defined as the 50 non-outlier molecules with a combination of high predicted pGI_{50} and small interval (criterion 2), provide better hit rates than the other two criteria. Correlation plot between a) criterion 1, 50 molecules with the highest predicted pGI_{50}, and criterion 2, the 50 molecules with a combination of high predicted pGI_{50} and small interval, b) between criterion 2, the 50 molecules with a combination of high predicted pGI_{50} and small interval, criterion 3, 50 molecules with the highest predicted pGI_{50}-low, c) criterion 2, the 50 molecules with a combination of high predicted pGI_{50} and small interval, and criterion 3, 50 molecules with the highest predicted pGI_{50}-low. 113

List of Tables

1.1. Description of the three main sources of uncertainty that can be encountered during the drug design and discovery process using AI-based methods. 41

2.1. Prominent studies using Conformal Prediction (CP) for QSAR/QSPR. A chronological overview of major studies utilizing conformal prediction for drug design, highlighting the number of compounds examined in each study, the model validation strategy, and the types of conformal prediction employed. The subscript notation in the validation type column indicates the subsequent partitioning of the training set into the proper training and calibration set. 58

3.1. The NCI-60 Growth Inhibition Data, which includes the following fields, is available for download in comma-delimited file format. This data can be downloaded from `https://wiki.nci.nih.gov/display/NCIDTPdata/NCI-60+Data+Download+-+Previous+Releases` 69

3.2. Brief description of physicochemical descriptors (PCDs) used in this study. These descriptors cover the pharmacokinetic and pharmacodynamics aspects of drug action. Lipinski's rule of five considers the first four descriptors, which are directly related to the pharmacokinetic properties. The total polar surface area of molecules is related to their pharmacodynamic aspect. 72

3.3. Summary of the machine learning (ML) models used in this thesis project. The "Used as" column describes whether the algorithm has been used as an underlying ($h(x)$) or error model ($g(x)$) while using conformal prediction. The "Comments" column includes additional information about using the corresponding algorithm. 73

4.1. The time to build the model grows exponentially as we increase the size of the MFPs. The table contains the computational times, in minutes, needed for generating MFP (three values: 256, 1024, and 2048 bits) and building models (two models: RF and XGB). MFPs were calculated for 50,555 unique molecules, while model building only included the approximately 41K molecules screened in the SR cell (Figure 3.2-c). . . 87

4.2. Models trained on datasets containing molecules with measured pGI_{50} = 4 are more confident in their predictions. The expected error is set at 0.2, corresponding to an 80% confidence level. The results are shown for both models, RF and XGB, considering the three scenarios analyzed (None: Non-removed, Train: Removed only from the training set, and Train-Test: Removed from both training and test sets). 89
4.3. Summary of the LDMO-CCP models built and evaluated. Each CP model is specified by its $h(x) - g(x)$ combination. The algorithms selected are LR, RF, and XGB, where LR serves as a benchmark for comparing the performance of non-linear algorithms. Therefore, some $h(x) - g(x)$ combinations are excluded from this study. RF and XGB models underwent a hyperparameter tuning step to optimize their performance. The LR algorithm utilizes default parameters without additional tuning. 102
4.4. Summary of the advantages and disadvantages of the CP-based models compared in this study. By systematically comparing and evaluating four distinct CP models, this research contributes valuable insights to the field of drug design. 110

Liste des acronymes

ACP

Aggregate Conformal Prediction. 51, 52

AI

Artificial Intelligence. 40, 45, 46, 55, 78

CCP

Cross Conformal Prediction. 51, 52

CP

Conformal Prediction. 15, 47–49, 52, 54–56

iid

independent and identically distributed. 57

LDMO

Leave-dissimilar-molecules-out. 18, 100, 101

MFP

Morgan fingerprints. 69, 78, 102

ML

Machine Learning. 46, 54, 55, 73, 78

PCD

physicochemical descriptors. 69, 71, 102

PD

pharmacodynamic. 71

PDD

Phenotypic Drug Discovery. 55, 78

PK

pharmacokinetic. 71

QSAR

Quantitative Structure-Activity Relationship. 40, 54

RF

Random Forest. 57

UQ

Uncertainty quantification. 40

1. Introduction

Contents

1.1. An overview of key project areas . . . 26
1.2. Cancer Overview . . . 27
1.2.1. Introduction to cancer . . . 27
1.2.2. Diagnosis and treatment options . . . 31
1.3. Drug design and drug discovery . . . 33
1.3.1. Introduction to Drug Discovery and Drug Development . . . 33
1.3.2. Phases of Drug Discovery . . . 35
1.3.2.1. Target Identification and validation . . . 35
1.3.2.2. Hit identification . . . 36
1.3.2.3. Hit to lead phase . . . 36
1.3.2.4. Lead optimization . . . 37
1.3.3. Approaches to drug discovery . . . 37
1.3.4. Ligand-Based Drug Design . . . 39
1.3.4.1. Quantitative Structure-Activity Relationship . . . 40
1.4. Uncertainty quantification in drug design . . . 40
1.4.1. Definition and sources of uncertainty . . . 40
1.4.2. Uncertainty types . . . 42
1.4.2.1. Aleatoric uncertainty . . . 42
1.4.2.2. Epistemic uncertainty . . . 42
1.4.2.3. Other uncertainty types . . . 43
1.4.3. Uncertainty Quantification methods . . . 43
1.4.3.1. Frequentist approaches . . . 43
1.4.3.2. Bayesian approaches . . . 44
1.5. Artificial intelligence in drug discovery and development . . . 45
1.5.1. A comprehensive overview of Artificial Intelligence . . . 45
1.5.2. Brief introduction to Machine Learning . . . 46
1.5.3. Conformal Prediction . . . 47
1.5.3.1. Basic concepts on Conformal Prediction . . . 47
1.5.3.2. Approaches to Conformal Prediction . . . 49
1.5.3.3. Advanced Conformal Prediction methods . . . 51
1.6. Scope and Structure . . . 52

1.1. An overview of key project areas

This chapter introduces the four key themes central to this thesis project. It aims to comprehensively understand the interplay between cancer, drug design and discovery, uncertainty quantification, and artificial intelligence (AI).

Section 1.2 provides an overview of the broad topic of cancer. This section presents the definition of cancer in detail, along with a brief explanation of how uncontrolled cell growth and division lead to cancer development. Furthermore, a comprehensive categorization of cancer is discussed. Additionally, the section covers the available diagnosis and treatment options, including chemotherapy (Figure 1.1-a).

The drug discovery process, presented in Section 1.3, is a pivotal step in developing chemotherapy drugs. It involves identifying novel compounds that can potentially treat diseases, followed by their optimization for potency, selectivity, and safety. This process is highly complex and involves the use of various techniques such as high-throughput screening, computational modeling, and virtual screening, among others. The final goal is to identify a drug candidate that can be further tested in preclinical and clinical trials for its efficacy and safety before it can be approved for patient use (Figure 1.1-b).

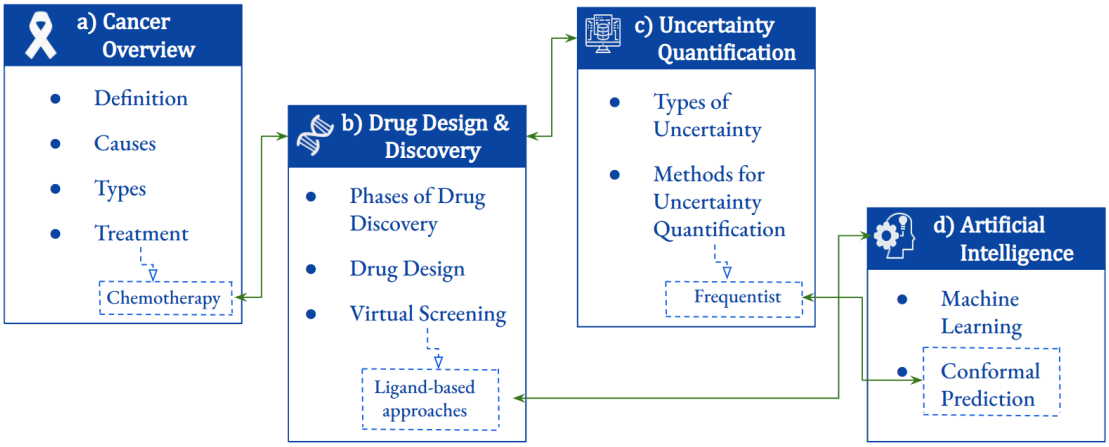


Figure 1.1.: **Relationship of the four main topics of this thesis work: cancer, drug design, uncertainty quantification, and artificial intelligence (AI).** a) The Introduction chapter begins with an overview of cancer, its causes, and treatments like chemotherapy. b) Drug discovery, a crucial process in developing chemotherapy drugs, often involves virtual screening (VS). c)-d) Advancements in computational methodologies have led to the integration of AI into VS approaches. Methods like conformal prediction enhance VS by providing uncertainty quantification. This ensures reliable predictions in drug design, improving the selection of potential therapies.

Section 1.4 discusses the topic of uncertainty quantification, with a particular focus on its relevance in drug design. This section highlights the crucial role of managing uncertainty in making informed decisions and the need for robust techniques to quantify it. We present an overview of the various types of uncertainty and the methods

that have been developed to address them, providing a comprehensive understanding of the field (Figure 1.1-c).

Section 1.5 provides a comprehensive overview of the impact of artificial intelligence on cancer research, with a particular focus on drug design. We delve into the specific applications of machine learning and conformal prediction, two subfields of artificial intelligence that have been extensively employed in drug design (Figure 1.1-d).

1.2. Cancer Overview

1.2.1. Introduction to cancer

The human body contains more than 10^{14} cells that, under normal conditions, grow and multiply to form new cells as the body needs them [5]. Cells consist predominantly of water, proteins, nucleic acids, metabolites, and a lipid membrane encapsulating them. Standard cellular processes, like metabolism, cell division, protein synthesis, signal transduction, and DNA replication/repair, are vital for cell survival, growth, and functioning [24]. In healthy cells, numerous genes intricately regulate cell division. Healthy growth depends on a delicate equilibrium between genes promoting cell proliferation and those inhibiting it. Additionally, genes signaling the initiation of apoptosis in damaged cells play a crucial role [105].

Cancer occurs when the regular growth of cells is disrupted, causing the abnormal or damaged cells to multiply uncontrollably and ultimately invade surrounding tissues and metastasize to distant sites [98]. Therefore, cancer development is a multistep process involving mutation and selection for cells with progressively increasing capacity for proliferation, survival, invasion, and metastasis. Tumor initiation occurs due to a genetic alteration in a single cell that triggers abnormal cell proliferation. Cell proliferation then leads to the outgrowth of clonally derived tumor cells. The next step is tumor progression, where additional mutations occur within cells of the tumor population that confer selective advantages, such as more rapid growth. The mutated cells and their descendants dominate the tumor population in a process known as clonal selection [34]. While cancer cells typically grow and form tumors (Figure 1.2), certain types of cancer, such as leukemia, do not follow this pattern.

Metastasis is a multistep process that involves cells breaking away from the primary tumor, invading local tissues and vessels, and setting up new cellular colonies in distant sites, causing secondary tumors to form [49]. Each step is quite complex, involving intricate molecular mechanisms that remain poorly understood. Cancer cells that can metastasize are often more resistant to apoptosis than normal cells, which allows them to survive and grow after leaving their original location. The presence of metastasis usually indicates advanced-stage cancer and is associated with a poorer prognosis and may have a higher likelihood of recurrence following treatment. Metastatic cancers are generally more aggressive and become extremely

difficult to remove through surgery or localized radiation treatments [5].

Cancer is a group of diseases resulting from alterations in the genes responsible for regulating cell growth and division. Among these genetic alterations are changes in oncogenes and tumor suppressor genes or TSGs (Figure 1.2) through mechanisms such as DNA mutation, chromosomal rearrangements, changes in the sections of DNA that act like "on" or "off" switches for nearby genes, and epigenetic changes [98]. Oncogenes stem from mutated versions of normal cellular genes called proto-oncogenes. In cancer, activating mutations in proto-oncogenes lead to uncontrolled cell division, increased cell survival, and the dissemination of cancer cells [62]. In contrast to the activating mutations, TSGs and the proteins they encode are functionally inactivated by mutations [23]. A TSG is a type of cancer gene created by loss-of-function mutations, and the idea is that these cellular genes function at least in part to suppress malignant transformation in normal cells [49].

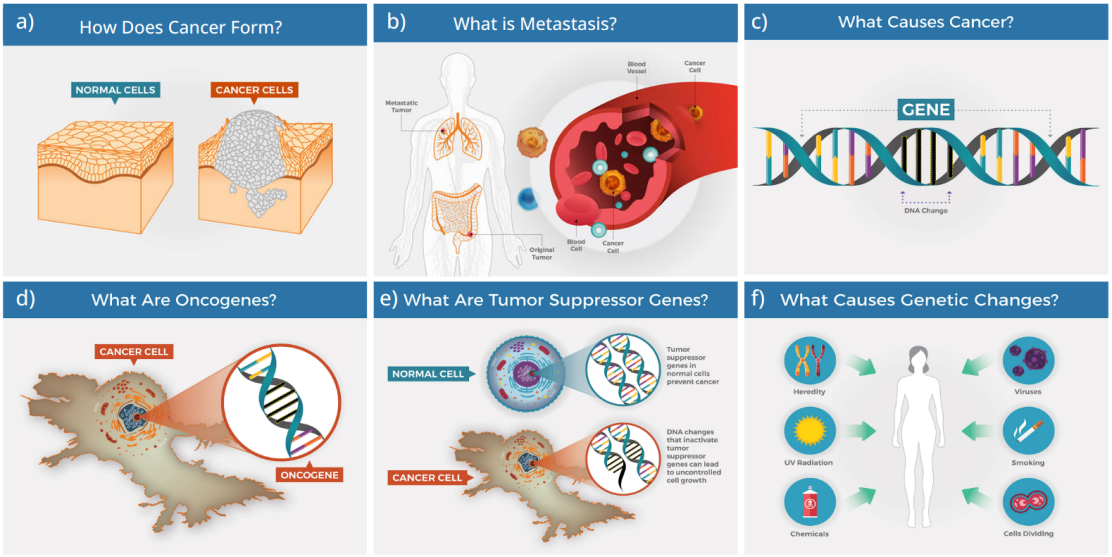


Figure 1.2.: **Comprehensive overview of the crucial elements involved in cancer development.** a)Disruptions in normal cell growth lead to uncontrolled multiplication and invasion of surrounding tissues, resulting in cancer. b) Cancer cells can break away from the primary tumor and spread to other areas of the body, which is referred to as metastasis. c) DNA changes that cause cancer usually occur in specific sections of DNA known as genes. d) Oncogenes and e) tumor suppressor genes, which play essential roles in normal cells, are responsible for driving cancer. f) Certain risk factors, including smoking, exposure to carcinogens, and heredity, among others, can elevate the chances of developing cancer. The images in this figure are from the National Cancer Institute's website (`https://www.cancer.gov`). They are in the public domain and not subject to copyright restrictions.

While the underlying reasons for specific individuals being more prone to cancer than others can be challenging to discern, extensive research in recent years has focused on understanding the risk factors associated with cancer development [32, 76, 77, 125, 127, 159]. Scientists have conducted epidemiology studies to identify possible risks and protective factors. The latter, according to the National Cancer

Institute (NCI), are factors that are linked to a lower risk of cancer. It is essential to mention that although most cancer risk and protective factors are initially identified in epidemiology studies, such studies on their own cannot prove that a behavior or substance causes cancer. This association can be established when many studies all point to a similar association between a potential risk factor and an increased risk of cancer and when there is a possible mechanism to explain how the risk factor might cause cancer. Some of the most-studied known or suspected risk factors for cancer are tobacco, alcohol, cancer-causing substances[1], hormones, obesity, age, diet, radiation, and sunlight, among others [98].

The World Health Organization (WHO) and the Union for International Cancer Control (UICC) jointly established a classification system for malignant tumors. This system categorizes tumors in four ways: firstly, by the affected tissue, organ, and system; the second and third classifications adhere to WHO guidelines, detailing the specific type and grade of the tumor. The fourth classification is based on the extent of spread using the Tumor Node Metastasis system. In this section, we describe only the initial classification, while further details about subsequent classifications can be found in [27].

Carcinoma, sarcoma, myeloma, leukemia, and lymphoma are tumors in the broad tumor classification [5]. The most common type of cancer is carcinoma, which results from altered epithelial cells. It begins in the skin or tissues lining or covering internal organs. Different subtypes include adenocarcinoma, basal cell carcinoma, squamous cell carcinoma, and transitional cell carcinoma [1]. Sarcomas starts in connective tissues, including muscle, fat, blood vessels, lymph vessels, and fibrous tissue. Sarcomas are much less common than carcinomas and can be broadly categorized as bone sarcomas and soft tissue sarcomas [26, 98]. Myeloma is a cancer that starts in plasma cells, a type of white blood cell made in the bone marrow [1]. Leukemia is a cancer of the white blood cells. As mentioned before, these cancers do not form solid tumors. Instead, many abnormal white blood cells, known as leukemia cells and leukemic blast cells, accumulate in the blood and bone marrow. This accumulation of abnormal cells displaces and crowds out normal blood cells [1, 98]. Lymphoma is a type of cancer that originates in the lymph glands or the cells of the lymphatic system, which runs throughout the body. As a result, lymphoma can occur in many different parts of the body [1, 26].

Below is the list of the cancer types studied in this thesis project, organized by system. A comprehensive list of several cancer types and their relevant information can be found in [9].

- Digestive system: Colorectal cancer is a malignant growth in the colon or rectum resulting from abnormal cell growth in the large intestine. Colon and rectal cancers share many features and are often grouped together. This type of cancer

[1] The National Toxicology Program has published the 15th Report on Carcinogens, a list of chemical substances known to be human carcinogens. The complete list can be consulted at https://ntp.niehs.nih.gov/whatwestudy/assessments/cancer/roc (accessed 25.01.2024).

is the third most frequent cancer in both men and women, accounting for 9.6% of cases worldwide [53].

- Urinary system: Renal or kidney cancer often develops in the lining of the small tubes within the kidney. Renal cancer is among the top 15 most frequent cancers, accounting for 4.4% of cases worldwide [53].
- Breast: Breast cancer is a malignant tumor that forms in the cells of the breast. It can occur in both men and women but is more common in women. Breast cancer accounts for 11.6% of cancer cases in women worldwide, making it the second most common cancer [53].
- Reproductive system
  - Ovarian cancer has three main types: epithelial, germ cell, and sex-cord-stromal. It is often called the "silent killer" and as of 2018, around 240,000 new cases were reported globally in women [144].
  - Prostate cancer develops from the gland cells in the prostate. This cancer accounts for 14.2% of cancer cases in males worldwide, making it the second most common cancer [53].
- Skin: Melanoma is a skin cancer caused by the uncontrolled growth of melanocytes. It is less common but more dangerous than other skin cancers because it can spread to other parts of the body if not detected early [4].
- Brain and nervous system: Central Nervous System (CNS) cancer refers to tumors that develop in the brain or spinal cord. Approximately 80,000 adults and 3,600 children were diagnosed with primary Central Nervous System tumors in 2018 [44].
- Blood and lymph system: Leukemia (acute or chronic) is a cancer of the white blood cells. However, some leukemias start in other blood cell types. Leukemia is the most common cancer in children, accounting for about one-third of childhood cancers [160].
- Lung: Lung cancer is divided into non-small cell lung cancer (NSCL) and small cell lung cancer (SCL). NSCL accounts for 80-85% of lung cancers and includes adenocarcinoma, squamous cell carcinoma, and large cell carcinoma [98].

Figure 1.3 displays the origin site of the nine cancer types examined in this thesis project.

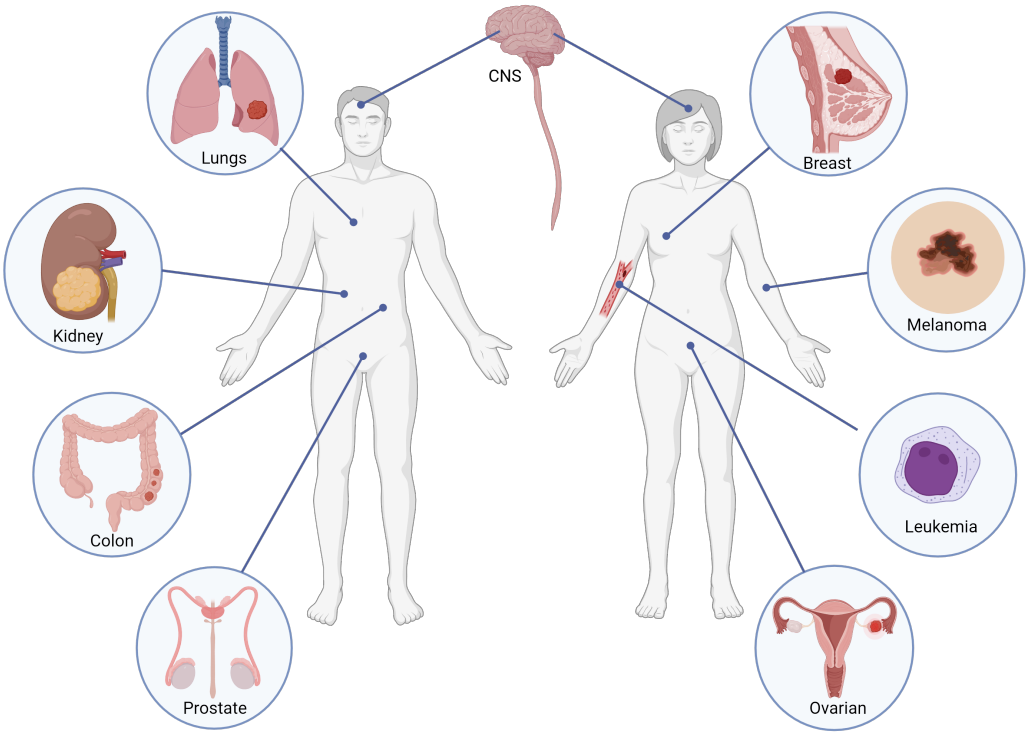


Figure 1.3.: **Cancer types studied in this work, grouped by body systems.** These types of cancer develop in the digestive system (colon), urinary system (renal), breast, reproductive system (ovarian and prostate), lungs (non-small cell lung cancer), skin (melanoma), blood and lymph system (leukemia), and central nervous system (CNS). Image created with (BioRender.com).

1.2.2. Diagnosis and treatment options

The International Agency for Research on Cancer (IARC) [2] has recently released the latest global cancer burden estimates [53]. CANCER TODAY offers interactive data visualization tools to explore estimates of incidence, mortality, and prevalence in 185 countries or territories for 36 cancer types by sex and age group. The report states that in 2022, there were approximately 20 million new cancer cases and 9.7 million deaths worldwide. Europe accounted for roughly 22.4% of these new cases, while Latin America and the Caribbean accounted for 7.77%.

A ranking of the top five most frequent cancers affecting men and women shows that among males, lung cancer dominated with 15.3% of all cases, followed closely by prostate cancer (14.2%), colorectal cancer (10.4%), stomach cancer (6.1%), and liver cancer (5.8%). The combined incidence of these top five cancers accounted for 52.8% of cases in men. Conversely, breast cancer held the highest frequency among females, comprising 23.8% of all cases. Lung cancer (9.4%), colorectal cancer (8.9%), cervix uteri cancer (6.8%), and thyroid cancer (6.4%) rounded out the top five. Together,

[2]The IARC is the cancer agency of the World Health Organization (WHO)

these five cancers represented 55.3% of all cases in women. The report also shows the 5-year prevalence of cases, which in 2022 reached about 53.5 million cases worldwide (Figure 1.4).

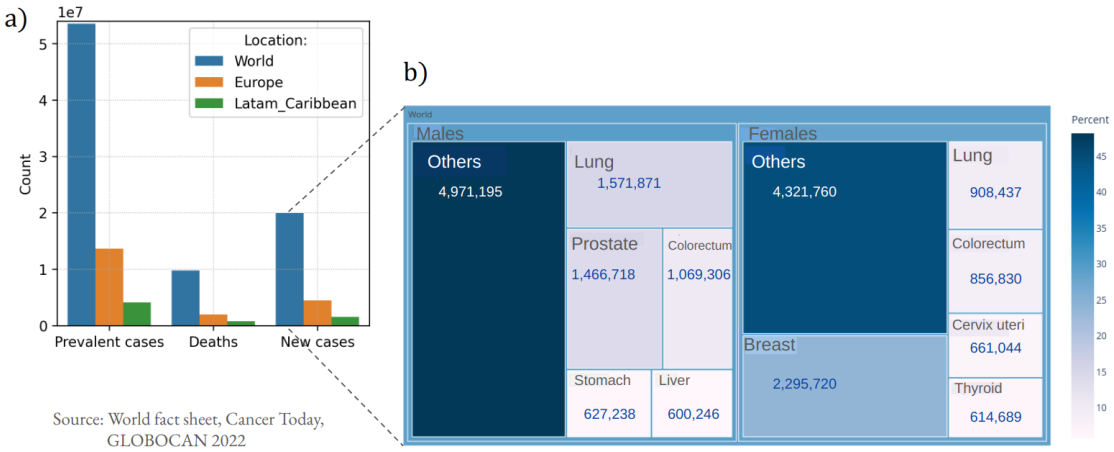


Figure 1.4.: **The 2022 global cancer statistics cover 185 countries and 36 types.** a) Approximately 20 million new cancer cases and 9.7 million deaths worldwide were reported in 2022. b) Among the newly reported cases, lung cancer was found to be the most common type in males, accounting for 15.3% of all cases. On the other hand, breast cancer was the most prevalent type in females, accounting for 23.8% of all cases.

These statistics indicate a high incidence of cancer and highlight the urgent need for advancements in treatment methods. Cancer treatments can be broadly classified into two categories, local treatments and systemic treatments. Local treatments, such as surgery and radiation therapy, target a specific tumor or area of the body. On the other hand, systemic treatments refer to drug treatments such as chemotherapy, immunotherapy, hormone, or targeted therapy, which can impact the entire body [9].

The choice of cancer treatment and its course of action is dependent on various factors, such as the type of cancer, its location, and the stage of its progression. Surgery is one of the most popular and traditional methods used to prevent, diagnose, stage, and treat cancer. This treatment is considered to be very promising for both benign and malignant tumors as it causes minimal damage to the surrounding tissues, unlike chemotherapy and radiotherapy [1]. Radiation therapy aims to destroy or damage cancer cells in a specific body area by using high-energy particles or waves, such as x-rays, gamma rays, electron beams, or protons [9].

Chemotherapy is a systematic treatment utilizing chemicals to treat disease. It involves drugs that travel throughout the body to kill cancer cells, including those distant from the primary tumor [1]. Despite its effectiveness against cancer, challenges include undesirable side effects on normal cells and the development of drug resistance, hindering optimal outcomes [109]. Immunotherapy treatment aims to boost the body's natural immune response to cancer. It involves using substances, such as immune checkpoint inhibitors, monoclonal antibodies, and adoptive cell transfer, to enhance

or stimulate the immune system's ability to recognize and destroy cancer cells [50, 106]. Hormonal therapies operate on the premise that hormones play a crucial role in regulating the behavior of specific cells. Some types of cancer are hormone-dependent for their growth. Thus, treatments that either block or modify hormones can sometimes help slow or stop the development of these cancers [9]. Targeted therapy aims to deliver drugs to cancer-specific genes or proteins, minimizing off-target side effects on healthy tissues. Targeted therapy can be used by itself or combined with other treatments, such as traditional or standard chemotherapy, surgery, or radiation therapy [109].

1.3. Drug design and drug discovery

1.3.1. Introduction to Drug Discovery and Drug Development

According to [99], a drug is a synthesized pharmaceutical molecule or compound intended to diagnose, cure, mitigate, treat, or prevent disease. The Food and Drug Administration (FDA) glossary, provides a more extensive definition, indicating that a drug is also a substance, other than food, intended to affect the structure or function of the body.

Developing a new drug is an expensive and time-consuming process that typically takes 12 to 15 years and costs between $1.8 and $2.8 billion from discovery to market [45, 67, 135]. The purpose of developing new drugs is to treat diseases that have no cure or to improve upon existing treatments by offering better efficacy, fewer side effects, greater compliance, fewer drug interactions, and an overall better quality of life for patients.

The drug discovery and development process involves various disciplines, such as medicinal chemistry, pharmacology, bioinformatics, structural biology, toxicology, and clinical research [135, 138]. These disciplines work together to identify and develop new therapeutic agents. It also implies the generation and use of vast data, usually obtained via different high-throughput technologies. Automation through computer-assisted methods is frequently employed to expedite specific stages of the process, enhance knowledge acquisition, and minimize errors. Therefore, drug discovery identifies new treatments using a combination of computational, experimental, translational, and clinical models. This process generally includes four main stages: drug discovery, preclinical development, clinical trials, and post-marketing surveillance [18]. Figure 1.5 provides an overview of these stages.

In this section, we will provide a brief overview of the stages of drug discovery, with further details to follow in the next section. The initial phase, known as target selection, is an early step wherein researchers identify and prioritize specific molecules associated with a disease like cancer and could be potential targets for therapeutic intervention. Subsequently, the process involves screening compounds to identify initial

hits and optimizing lead compounds to enhance efficacy, safety, and pharmacokinetic properties [96].

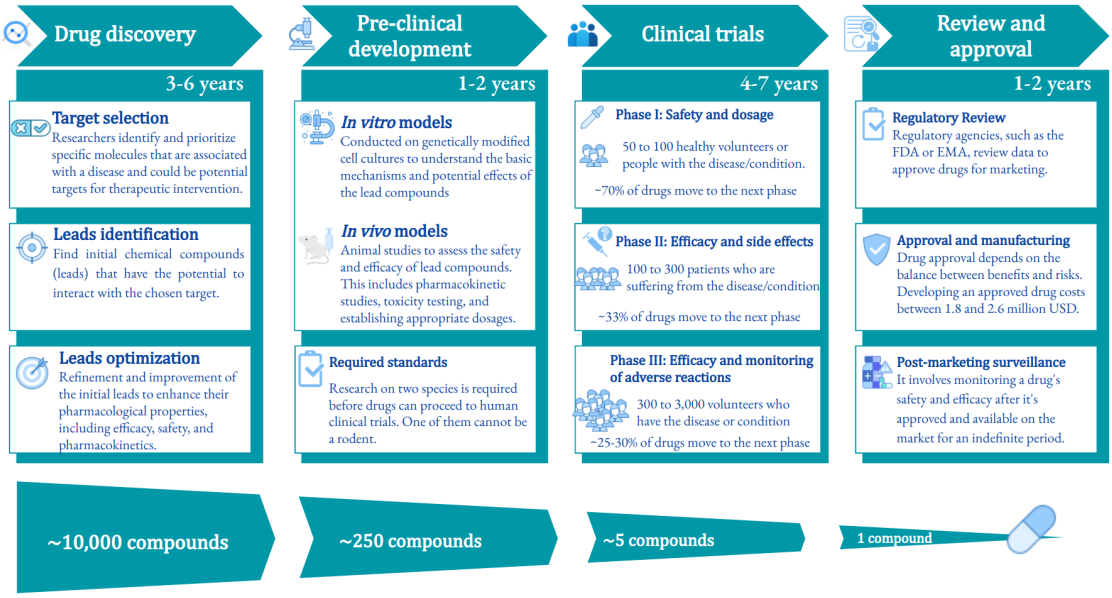


Figure 1.5.: **Drug discovery and development process**. This process involves several stages, from left to right: drug discovery, preclinical research, clinical trials (Phases I-III), and Regulatory review (Phase IV Post-marketing surveillance). This process typically spans 12-15 years, with costs ranging from $1.8 and $2.6 billion. Only a small fraction of compounds tested advance to approval, illustrating the rigorous nature of drug development in ensuring safety and efficacy for patient use. It is essential to mention that the number of molecules progressing between stages in drug discovery varies according to the work reported. The number reported in this figure serves as an illustrative example highlighting this process's challenging and rigorous nature [56, 81].

In the preclinical phase, researchers focus on addressing fundamental questions regarding the safety of potential drugs. These preclinical or laboratory studies involve in vitro or in vivo models. The first step is usually in vitro testing on genetically modified cell cultures. In vitro tests offer valuable insights into intracellular processes, facilitate preliminary toxicity testing, and aid in comprehending the mechanisms of action of new drugs. On the other hand, in vivo models are employed to account for the complexity of the entire organism. Therefore, animal experiments are required for in vivo models, and this type of research must be performed on two different species, one of which is not a rodent [18]. Researchers usually use two kinds of models; for example, drugs that have been successful in cell line models can be further evaluated in animal models to estimate their safety in a living organism.

The preclinical stage of drug development estimates the drug's safety in living organisms. However, evaluating the drug's safety in the human body is necessary, which is beyond the scope of animal testing. This is where the clinical research stage comes in. There are four phases in the clinical trial. However, phase IV, review and control, is usually considered an independent stage [96]. Phase I is carried out on either healthy

or sick volunteers. This part of the study primarily assesses the drug's safety and determines the expected dosage, for which between 50 and 100 volunteers take part. Around 70% of the drugs initially tested remain at the end of this stage. Next, phase II concerns people diagnosed with the disease. Its primary purposes are to confirm the drug's safety and obtain data on its effectiveness. This phase usually involves a group of between 100 and 300 diagnosed patients. After phase II, around 33% of the drug passes to the next phase. Studies in Phase III are carried out on a larger population, between 300 and 3000 diagnosed patients, and can last several years. This research aims to determine the optimal dose and confirm the effectiveness of the study drug. A successful output of Phase III enables the registration of a new drug [18].

1.3.2. Phases of Drug Discovery

As we saw in the previous section, the drug discovery and development process is very time-consuming and costly. In this section, we describe the stages of drug discovery.

1.3.2.1. Target Identification and validation

There are two main reasons why drugs fail in clinical trials, either they do not produce the intended results, or they pose a risk to human health [67]. To prevent these failures, it is crucial to identify and validate the target of a new drug in the early stages of drug discovery. Therefore, identifying and validating the target is one of the most crucial stages in developing a new drug.

In the realm of drug discovery, biologically active compounds can take many forms, ranging from small synthetic molecules to large, complex antibodies. Regardless of their size and composition, these compounds achieve their therapeutic effects by interacting with specific molecular structures within the human body [138]. This interaction is commonly referred to as targeting, and the molecular structure that is being targeted is known as the "target." Researchers must look at various molecular structures to identify these targets, including enzymes, receptors, metabolites, substrates, ion channels, transport proteins, DNA, RNA, and ribosomes [70]. When a drug compound interacts with its designated target, it can trigger a cascade of chemical reactions that ultimately produce the desired therapeutic effect.

Target identification can be done through phenotypic screening or genetic association studies. Phenotypic screening involves evaluating compounds in cell-based assays or animal models to identify active compounds [95]. The molecular target is then determined through genetic or chemical proteomic-based approaches. Unlike target-based screening, which focuses on a specific molecular target, phenotypic screening is more concerned with observing a compound's functional outcome or effect on the biological system. Genetic association studies identify genetic variants associated with disease risk or progression [138]. Target identification has been addressed in the field of artificial intelligence (AI), particularly with machine learning (ML) models.

These ML models can analyze large datasets generated from genomics, proteomics, and other omics studies to identify patterns and potential targets associated with diseases [113].

After identifying a potential drug target, it needs to undergo the process of validation wherein its function in a disease state is ascertained. Various techniques can be employed to validate the accuracy of the identified target. These techniques include using in vitro tools, whole animal models, or modulating the target in disease patients [67]. While each approach is valid and useful in its own right, it is recommended to use a multi-validation approach to increase confidence in the observed outcome. This will help ensure the identified target is accurate and reliable, reducing its chances of failure in clinical trials.

1.3.2.2. Hit identification

Once a potential therapeutic target has been validated, the subsequent step in the drug discovery process involves identifying "hits". Essentially, "hits" refer to chemical compounds demonstrating the intended biological activity in a screening assay [138]. These compounds are a starting point for further optimizing and developing a promising drug candidate. This phase involves developing compound screening assays that help identify potential drug candidates. These assays are designed to test the effectiveness of the compounds in inhibiting or activating the target.

Different hit identification methods exist, such as high throughput screening (HTS), where libraries of up to a million compounds are quickly evaluated against a protein target in a biochemical or cell-based assay [138]. Other methods for hit identification are virtual screening, fragment-based screening, and Structural aided drug design, among others [67].

Virtual Screening (VS) is a widely used technique in the field of drug discovery and development. It involves using computational methods to identify potential drug candidates from a vast library of chemical compounds. The primary objective of VS is to prioritize the compounds that are most likely effective against a particular disease or condition [30]. This approach has gained popularity as a complementary method to HTS, often used as the primary method for identifying drug leads [67]. VS techniques can be used to filter large chemical libraries down to a more manageable number of compounds, making the drug discovery process more efficient and cost-effective.

1.3.2.3. Hit to lead phase

During the hit-to-lead phase of drug discovery, the goal is to optimize the initial hits identified in the screening phase and transform them into lead compounds [67]. A chemical lead can be described as a synthetically feasible, stable, and drug-like molecule. In addition, it should be active in primary and secondary assays and exhibit acceptable specificity and selectivity for the target [145].

At this stage, medicinal chemists select three to five different chemical series for lead generation and evaluate similar compounds to establish a quantitative structure-activity relationship towards desired activity, target selectivity, physicochemical properties, ADME (Absorption, Distribution, Metabolism, and Excretion) properties, pharmacokinetics, and toxicity potential [138, 145]. To achieve this, chemists use various approaches, such as conventional organic chemistry and combinatorial chemistry, to synthesize compounds.

An important step in this phase is to assess the drug-likeness of a compound based on its physicochemical properties. To do this, the Lipinski Rule of Five is considered to assess the drug-likeness [145]:

- Its molecular mass is less than 500 daltons
- Its logP, which is a measure of lipophilicity, is less than 5
- The number of hydrogen bond donors is less than 5
- The number of hydrogen bond acceptors is less than 10

These guidelines are used to evaluate whether a compound has properties that are associated with oral bioavailability and favorable pharmacokinetic characteristics.

1.3.2.4. Lead optimization

Lead optimization is the final stage in drug discovery and involves making chemical modifications to the lead structure to overcome its limitations and create potential preclinical development candidates [145]. The primary objective of lead optimization is to enhance the physicochemical and ADME properties while minimizing the toxicity liabilities of the lead structure [138]. This process is undertaken to identify a potentially safe compound with favorable pharmacokinetics, which can be further developed for clinical trials.

1.3.3. Approaches to drug discovery

There are four main approaches to drug discovery: serendipity, chemical modifications, screening of databases, and drug design [45]. In modern drug design, these approaches are often combined to get the best out of each of them and speed up the overall process. In the following, we will explain the most important points of each approach, with particular attention to drug design.

Serendipity is the oldest approach in drug discovery. Although the word "serendipity" has its roots in a Persian fairy tale, it is now defined as "the faculty of making happy and unexpected discoveries by accident" or "the faculty of finding valuable or agreeable things not sought for" [15, 90]. In the medical field, it refers to an unintentional discovery [142]. As of 2012, it was estimated that around 6% of all drugs on the market

involved serendipity, with penicillin being the best known. In the case of anticancer drugs, around 35% can be attributed to a serendipitous event [61].

The second approach is the chemical modification of known drugs or natural products. This approach is also known as molecular or structural modification [42] and refers to the intentional alteration or manipulation of the chemical structure of a compound to improve its properties, such as efficacy, selectivity, pharmacokinetics, or safety [47]. Aspirin was discovered using this method [45].

Screening of databases is another way to discover new drugs. This screening can be done virtually or by HTS assays. HTS is employed to detect initial compounds for small-molecule drug design campaigns. The use of robots in HTS enables the screening of huge compound libraries at a rate that can surpass a few thousand compounds per day or per week [19]. A list of HTS-based drugs can be found in [87]. On the other hand, VS is a computational approach used to identify hit molecules from libraries of chemical compounds for developing novel lead molecules [43]. It involves using filters to evaluate hits against a specific drug target. VS and HTS are complementary techniques used in lead identification.

Drug design, a leading and cost-effective approach in drug discovery, has seen rapid progress fueled by artificial intelligence (AI) integration. This fast progress is primarily due to the increasing availability of relevant experimental data and the easy accessibility of high-quality, well-documented implementations of various machine learning (ML) algorithms [13]. In drug design, biological targets are first identified, followed by discovering molecules that interact with these targets. Compounds with optimal properties such as affinity, selectivity, and non-toxicity are derived through iterative structure optimization [1]. Ligand-based and structure-based approaches are the two main approaches in drug design. The ligand-based approach will be discussed later in this section.

Developing *in silico* modeling technologies in modern drug design represents a significant advancement in this field. These *in silico* technologies can be applied for virtual screening [59, 149], structure-activity relationships (SAR) and quantitative SAR (QSAR) analysis [65], ADME (Absorption, Distribution, Metabolism, and Excretion) modeling [16, 73], and modeling of drug-target interactions [112]. To facilitate the analysis and manipulation of molecules, they are often represented numerically using methods such as InChI and SMILES, which are based on linear notations. Other methods include molecular fingerprints like structural keys and circular fingerprints [10]. The molecular properties can also be encoded, for example, by using the molecular weight, the number of hydrogen bond acceptors and donors, and the polar surface area, among others. These descriptors will be discussed later in this work. However, the numerical encoding can lead to a loss of information, as different descriptors capture different properties of the molecules. This information loss has led to an emerging research area for molecular representation learning [51, 83, 161], where models are developed to represent the intricate nature of a molecule.

Dose-response quantification is also needed in drug design to perform different analyses. Among the indicators of toxicity and pharmacological activity are the Half-Maximal Inhibitory Concentration (IC_{50}), the Lethal Dose-50 (LD_{50}), and the half-maximal effective concentration of a compound (EC_{50}) [75]. IC_{50} is a measure of the concentration of the inhibitor that causes 50% inhibition of enzyme or receptor activity at the target level. LD_{50} is the amount of a substance that kills 50% of a group of animals. Though it is criticized for low reproducibility, it is still used for assigning substances to a toxicity class. The EC_{50} measures the potency of a compound in producing 50% of its maximum effect, whether it be pharmacological efficacy or acute toxicity.

The quantitative characterization of molecules and the quantification of their dose response serve as the foundation for various analyses. The relationships between molecular descriptors and the activities of a series of molecules or compounds are identified as QSAR analyses. In contemporary settings, QSAR analyses are predominantly executed using ML techniques, including decision trees, random forests, and neural networks, among others. Despite their efficacy in delivering accurate predictions, these methods are commonly called "black box" approaches due to their inherent complexity [1]. While these advanced techniques contribute significantly to predictive modeling, enhancing interpretability for a more comprehensive understanding of the underlying molecular interactions remains challenging.

1.3.4. Ligand-Based Drug Design

Virtual Screening, or VS, is a widely used computational technique to identify potential drug candidates from a vast library of chemical compounds [30]. It aims to prioritize the most likely effective compounds against a particular disease or condition, making the drug discovery process more efficient and cost-effective. Two standard methods used in virtual screening are ligand-based and structure-based approaches [92].

Structure-based virtual screening (SBVS) relies on docking large libraries of compounds into 3D structures of the target protein and evaluating the protein-ligand contacts that drive binding [58]. With recent advances in sampling, scoring functions [149], and strategies for considering protein and ligand flexibility, SBVS has become more effective and accurate in identifying potential ligands.

However, in some cases, no information is available about the structure of the target protein, making it difficult to design drugs that can effectively bind to it. Therefore, without any structure information available for the therapeutic target, the alternative approach is Ligand-based virtual screening (LBVS) [92]. This approach does not require prior knowledge about how the target protein functions but instead relies on the known properties of small molecules that have previously been shown to interact with similar targets. By identifying molecules with similar structures and bioactivity to these known compounds, scientists can develop new drug candidates that are more likely to be effective [131].

The most common LBVS techniques include molecular similarity-based search, quantitative structure-activity relationship (QSAR), and pharmacophore modeling [131]. In this thesis project, we focus on QSAR modeling.

1.3.4.1. Quantitative Structure-Activity Relationship

Quantitative Structure-Activity Relationship (QSAR) is a widely used approach to evaluate the relationship between the chemical structures of a series of compounds and their corresponding chemical or biological activity [3]. QSAR method is based on the hypothesis that compounds with similar chemical or physicochemical properties exhibit similar activity. In other words, QSAR can predict the activity of new compounds based on their structural features, which makes it a useful tool in drug discovery and other areas of chemical research. According to Acharya et al., the general methodology of QSAR is built upon a series of the following consecutive steps [3]:

- Identify ligands with experimentally measured values of the desired biological activity.
- Determine the molecular descriptors that are linked to the different structural and physicochemical characteristics of the molecules being analyzed.
- Identify the links between the biological activity and the molecular descriptors that can clarify the differences in activity found in the dataset.
- Test the statistical stability and predictive power of the QSAR model.

QSAR tools have evolved into AI-based approaches such as linear discriminant analysis, support vector machines, random forests, and decision trees [111]. The main difference between traditional QSAR modeling and AI-based QSAR modeling is the methods used to predict these relationships. Traditional QSAR models are based on mathematical equations derived from statistical analyses of chemical and biological data. These equations typically involve descriptors derived from the chemical structure of molecules.

1.4. Uncertainty quantification in drug design

1.4.1. Definition and sources of uncertainty

The Oxford English Dictionary defines uncertainty as "a state of doubt; want of assurance or confidence". In other words, it refers to the lack of complete knowledge or precision about a system or outcome.

Nowadays, where numerous applications heavily depend on simulations and Artificial Intelligence (AI)-based algorithms, it is essential to address and consider the uncertainties associated with their outputs. Uncertainty quantification (UQ) is a

broad interdisciplinary field that involves methods and techniques to assess, characterize, and manage uncertainty in various scientific and engineering applications [153]. AI techniques can be applied within the context of uncertainty quantification, especially as AI-based models play an increasingly significant role in supporting decision-making processes. The primary aim of UQ is to assess the reliability of predictions, account for model variations, randomness, and misspecification, and ultimately assist in decision-making.

Various factors can contribute to uncertainty in mathematical models, including human factors, inaccuracies in the input data, and assumptions made in the model's formulation. These uncertainties can significantly impact the accuracy and reliability of the model's predictions and should be carefully considered and addressed [126]. Table 1.1 provides a summary of the categories of these three primary sources of uncertainty, as found in the literature [2, 79, 91, 126, 153].

Table 1.1.: Description of the three main sources of uncertainty that can be encountered during the drug design and discovery process using AI-based methods.

Source	Categories	Description
Human factors	Judgment and subjectivity	Uncertainty introduced by subjective interpretations biases, or assumptions made by individuals involved in data collection, analysis, or decision-making.
	Interpretation of results	Uncertainty associated with the interpretation and communication of analysis results, leading to different conclusions or actions based on the same data.
Input data	Error measurement	Errors introduced during data collection or observation processes.
	Instrumentation precision	Variability and limitations in measurement instruments.
	Sampling variability	Variability due to selecting a subset of a larger population for measurement.
	External interactions	Uncertainty arising from interactions with other systems or external influences.
Model formulation	Parameter estimation	Uncertainty in determining model parameters from data.
	Model uncertainty	Uncertainty stemming from the choice and complexity of mathematical or statistical models used to represent real-world phenomena.
	Simplifying assumptions	Uncertainty introduced by simplifying complex systems or processes into more manageable models, potentially neglecting important factors.

Understanding these sources of uncertainty is crucial for developing appropriate strategies to quantify, mitigate, and manage uncertainty in different domains.

1.4.2. Uncertainty types

Uncertainty is an ever-present element in our understanding of the world and plays an essential role in decision-making and predictive modeling. Commonly, uncertainty can be broadly classified into two categories: aleatoric uncertainty and epistemic uncertainty. In addition to the usual sources of uncertainty, the complexity of the models used to analyze these data and the variability inherent in the data itself can introduce uncertainty that falls outside of the traditional aleatoric and epistemic categories. This section briefly explains these uncertainty types.

1.4.2.1. Aleatoric uncertainty

Aleatoric uncertainty, often called statistical uncertainty, encompasses the inherent noise present in observations, stemming from the intrinsic complexity of the underlying data. Because of this, aleatoric uncertainty is also referred to as data uncertainty. This type of uncertainty can arise in both the training data used to build a model and the new data that the model is being used to predict.

The primary sources of data uncertainty include measurement errors, misclassification errors, binning, censoring and truncation, and missing values, introducing challenges in accurately modeling and predicting outcomes [79]. In domains like drug design, data uncertainty can be further categorized into natural and technical variability [91]. Natural variability may arise from factors such as diverse organisms and biological samples, contributing to the rich tapestry of data complexity.

On the other hand, technical variability can be attributed to measurement procedures, instrumentation nuances, sample preparation methods, and other technical issues. Aleatoric uncertainty is considered irreducible, meaning it is inherently unremovable from the system. This implies that regardless of the advancements in modeling or data collection techniques, certain uncertainties persist due to the intrinsic randomness or variability within the observed phenomena [164].

1.4.2.2. Epistemic uncertainty

Epistemic uncertainty describes a lack of knowledge of the trained model, and it usually arises when the model encounters situations or scenarios for which it lacks sufficient training data [79]. An example of epistemic uncertainty in drug discovery is the incomplete matrix of compounds tested across various assays [91]. This limited knowledge introduces uncertainty about the model's predictions in scenarios where empirical data is sparse or absent. However, unlike aleatoric uncertainty, epistemic uncertainty can be omitted if sufficient data are available.

Although these approaches cover different aspects of uncertainty, as AI-based models advance in sophistication, the complex relationships and representations they learn can introduce new types of uncertainty. Moreover, with the rise of data-driven approaches, AI-based models are trained on vast and diverse datasets. The inherent variability in these data sources and developing more sophisticated models introduce distinct forms of uncertainty beyond the conventional categories. This includes the emergence of approximation uncertainty and distributional uncertainty.

1.4.2.3. Other uncertainty types

Approximation uncertainty arises from the inherent limitations of simplistic models to accurately represent complex data structures [164]. A simple example is a linear model attempting to fit a non-linear dataset. This same situation applies to drug discovery, in which approximation uncertainty can be exemplified by the challenges faced in predicting molecular interactions or biological responses using linear models that may oversimplify the underlying complexities of biological systems. According to Yu et al., this type of uncertainty, like epistemic uncertainty, can be omitted using deep learning methods [164].

Distributional uncertainty arises when a model encounters test data that significantly differ from the data it was trained on, leading to a lack of confidence in its predictions [79, 88]. Unlike aleatoric uncertainty, this type extends beyond mere label or data noise as it stems from a mismatch between the distributions of the training and test datasets, a phenomenon commonly referred to as dataset shift [156]. In drug discovery, a QSAR model trained on a dataset of compounds with specific chemical properties may face challenges when predicting the properties of novel compounds with substantially different structures or characteristics. This distributional shift could arise due to variations in molecular structures not adequately represented in the training set, introducing uncertainties in the model's ability to generalize to diverse chemical spaces.

1.4.3. Uncertainty Quantification methods

In drug design, decision-making is inherently conducted in the presence of uncertainty, leading to the development of many methods to quantify uncertainty. In this subsection, we review the Frequentist and Bayesian approaches.

1.4.3.1. Frequentist approaches

Frequentist methods are a fundamental approach to statistical inference and uncertainty quantification, primarily focusing on repeated sampling and statistical procedures' long-run behavior [128]. Frequentist methods use the likelihood function to

draw inferences. Unlike Bayesian methods, which incorporate prior beliefs and update probabilities based on observed data, frequentist methods emphasize properties of estimators and tests based on hypothetical repeated sampling [91].

Frequentist approaches to statistical analysis can be broadly categorized into parametric and non-parametric (or distribution-free) methods [120]. In parametric methods, the form of the potential error distribution is specified and trusted in the analysis. On the other hand, in non-parametric methods, the error distribution is not specified.

In frequentist inference, uncertainty is quantified through point estimation, interval estimation, hypothesis testing, and resampling methods, among others [54]. Point estimation involves estimating unknown parameters based on observed data using estimators such as the sample mean or maximum likelihood estimator. The latter is the most common point estimator, which seeks to find parameter values maximizing the likelihood function given the observed data.

Confidence intervals measure the uncertainty associated with parameter estimates. A confidence interval is a type of interval estimate that, when the estimation process is repeated many times, contains the true parameter value with a certain probability [12].

Hypothesis testing is another key aspect of the frequentist approach. It involves setting up null and alternative hypotheses and using statistical tests (e.g., t-test) to evaluate whether there is sufficient evidence in the data to reject the null hypothesis in favor of the alternative. Frequentist resampling methods such as the bootstrap and permutation tests are utilized to estimate the sampling distribution of statistics and assess uncertainty without explicit parametric assumptions [91].

Conformal prediction is a method that provides accurate predictions and calibrated measures of uncertainty based on observed data [12, 155]. It follows frequentist principles but is more flexible and adaptive than traditional frequentist approaches [91]. This makes it particularly useful for complex and non-standard data settings. Unlike traditional frequentist methods that rely on assumptions about data distributions and parameter estimators, conformal prediction focuses on constructing prediction regions that guarantee valid statistical properties without imposing strict distributional assumptions [155]. These regions are designed to provide accurate predictions with high confidence levels, thus ensuring that the model can make reliable predictions in a wide range of complex scenarios. The details regarding this method will be elaborated further in Section 1.5.3.

1.4.3.2. Bayesian approaches

The Bayesian approach to statistical estimation involves a two-stage experiment and is based on Thomas Bayes's famous theorem. Bayes' theorem provides a formula for calculating the probability of a particular first-stage outcome if the outcome of the second stage is known [128]. Bayes' theorem forms the cornerstone of Bayesian inference,

which updates prior beliefs (expressed as a prior distribution) about parameters based on observed data, yielding a posterior distribution that combines prior knowledge with the likelihood of the data [54].

In contrast to the frequentist approach, Bayesian inference treats probability as a measure of uncertainty representing subjective beliefs about the likelihood of events or parameters. Instead of point estimates, Bayesian inference provides a complete distribution of plausible parameter values, allowing for richer uncertainty characterization [166].

Bayesian methods offer flexibility in incorporating prior information into the analysis, enabling researchers to explicitly incorporate expert knowledge or historical data into the modeling process [128]. This feature is particularly advantageous when dealing with small sample sizes or complex models where prior information can help regularize parameter estimates and improve predictive performance. However, Bayesian inference can be computationally intensive, especially for high-dimensional or complex models, due to the need to calculate posterior distributions.

Gaussian processes are one of the most famous methods in Bayesian machine learning [40, 107]. Neural networks and deep learning are also top-rated, including recent research on Bayesian versions of these networks that are especially interesting in quantifying uncertainty [2, 72, 88]. Imprecise probabilities are closely related to the concept of credal sets, and they can be thought of as a generalization of Bayesian inference where a set of priors is used instead of a single prior distribution [68].

In summary, the frequentist approach emphasizes objective data analysis by relying on long-run frequencies and classical statistical techniques such as confidence intervals and hypothesis testing. On the other hand, the Bayesian approach incorporates subjective beliefs through prior distributions, enabling richer uncertainty quantification via posterior distributions. Both methods have their strengths and limitations, and the choice between them depends on the nature of the problem, available data, and underlying modeling assumptions.

1.5. Artificial intelligence in drug discovery and development

1.5.1. A comprehensive overview of Artificial Intelligence

AI refers to developing computer systems that can perform tasks that typically require human intelligence [80]. Although the term human intelligence is too broad, intelligence in AI refers to the ability of computer systems to perform tasks that include learning, reasoning, problem-solving, perception, speech recognition, and language understanding.

A British mathematician and computer scientist, Alan Turing, developed a theoretical

model called the Turing machine, which laid the groundwork for artificial intelligence. His concept of the Turing Test, introduced in 1950, aimed to evaluate a machine's ability to exhibit intelligent behavior. However, it was not until 1956 that John McCarthy introduced the term Artificial Intelligence during the Darmount Conference [17].

From this point on, many expectations arose about this new area. However, due to the limitations in terms of computational resources, the progress of artificial intelligence was limited. This lack of fulfillment in the expectations of artificial intelligence led to a period between the late 1970s and mid-1980s known as the "AI winter." This winter would be broken in the 1980s thanks to the interest in neural networks and expert systems. These expert systems would gain popularity due to their applications in medical diagnostics [167].

AI can be classified based on its scope into weak AI and strong AI. Philosopher and cognitive scientist John Searle popularized this categorization in the early 1980s in a paper titled "Minds, Brains, and Programs" [130]. The first category is weak or narrow AI, encompassing AI systems specifically designed and assessed to excel at particular tasks. Currently, most AI applications fall into this category as they are tailored for specific tasks, such as predicting ADMET properties and QSAR, among others. The second category is general or strong AI, which refers to a hypothetical form of AI capable of understanding, learning, and applying knowledge to a wide variety of tasks, similar to human intelligence. In other words, unlike weak AI, general AI is not task-specific. General AI is often featured in science fiction literature. Nowadays, a third category called super AI refers to a hypothetical AI level surpassing any human mind. However, the term super AI is less formally established or recognized than weak AI and strong AI in the academic and philosophical literature.

1.5.2. Brief introduction to Machine Learning

In the previous section, we briefly introduced the relationship between AI and Machine Learning (ML) and their applications in drug design. In this section, we will define machine learning formally and examine how it is used in phenotypic virtual screening.

Given a data set Z, called the example space, defined as $Z := X \times Y$, where X and Y are called the instance space and the label space, respectively. Therefore, each example $z_i \in Z$ consist of an instance x_i and its label y_i, i.e, $z_i = (x_i, y_i)$. We always assume that Z contains more than one element [12]. In a basic ML problem, the example space Z is usually split into a training and a test set. The examples in the training set will usually denoted $z_1, \ldots, z_n$, and the example to be predicted z_{n+1}. This way to represent z_i covers two fundamental machine learning problems: classification, in which Y is a finite set, and regression, in which $Y = \mathbb{R}$ is the set of real numbers [155].

ML is a vast field involving a wide range of techniques, each designed to attend to specific tasks. One such technique is supervised learning, where models learn to identify

patterns and map input data to labeled outputs, making it ideal for classification and regression tasks [97]. In other words, supervised learning involves training models with labeled data to make predictions based on new, unseen data.

Unsupervised learning is a type of ML where the model is trained on unlabeled data. In unsupervised learning, the model is left alone to find patterns and relationships in the data without any guidance or supervision [6]. This technique is often used for clustering [63], where the model groups similar data points together, and dimensionality reduction [114], where the model reduces the number of features in the data while retaining the most important information. Unlike supervised learning, unsupervised learning does not require any labeled data, making it particularly useful when obtaining labeled data is difficult or expensive.

Semi-supervised learning is a machine learning technique that combines the characteristics of both supervised and unsupervised learning [97]. This method is particularly useful when there is a shortage of labeled data. By utilizing both labeled and unlabeled data, semi-supervised learning can improve the accuracy of the models and reduce the need for manual labeling [41].

Reinforcement learning is another ML approach that enables agents to learn and make decisions by interacting with an environment to maximize rewards [94]. It involves trial-and-error learning, where the agent learns from past decisions and makes better decisions in the future. This learning technique is widely applicable in various fields, including drug discovery and development [108, 168].

Despite its advantages, AI and ML face significant data challenges, such as the data's scale, growth, diversity, and uncertainty.

1.5.3. Conformal Prediction

1.5.3.1. Basic concepts on Conformal Prediction

Using AI, especially black-box ML models, poses challenges, with uncertainty quantification being a significant concern. This is crucial when ML output influences decision-making, requiring quantification of prediction confidence. Uncertainty quantification is approached from various perspectives, e.g., using Bayesian models, which explicitly incorporate probability distributions to represent uncertainty. On the other hand, the chosen metric for assessing model performance should reflect its capacity to handle uncertainty. This section focuses on the CP framework, an alternative for uncertainty quantification.

Conformal prediction is built on top of any ML algorithm. Therefore, some steps are similar in both methods. As in ML, we start with a dataset Z with the same characteristics described in the previous section. ICP initially divides the dataset into training and test sets (step 1). Then, the training set is further split into a proper training set, Z_t, and a calibration set Z_c where $|Z_c| = q$ (step 2). In step 3, we fit a model

$h(x)$, called the *underlying model*, using those instances in the proper training set Z_t. In the so-called normalized ICP [103], a second model, $g(x)$, is built using the proper training set.

The model $g(x)$ is called the error model, and it aims to predict the error made when predicting y [12]. This works as follows. Given the prediction $h(x_i) = \tilde{y}_i$, we can calculate the error for each prediction as $|x_i - h(x_i)| = |x_i - \tilde{y}| = e$. Since the error e is calculated for each sample in the proper training set, we have a second example space defined as

$$Z_e := X \times E,$$

where E are the errors made by the underlying model when predicting y, and X is the same instance space used for training the underlying model.

The underlying and error models can be used to define a calibration function (step 4), which can be either a conformity or nonconformity function [155]. The calibration function assesses how similar (conformity) or dissimilar (nonconformity) a new instance is compared to the data the model has been built upon. Further details of the calibration functions for the special cases of classification and regression will be given later in this section. Regardless of the specific calibration function employed, switching between conformity or nonconformity measures can be done as follows:

- *conformity* := - *nonconformity*
- *conformity* := 1 - *nonconformity*

While any function can be used as a nonconformity function, choosing a function is the key to obtaining a predictor with high efficiency [8]. The nonconformity function is then applied to all elements in the calibration set Z_c to get a set of calibration scores (step 4). As a notation stated in [155], nonconformity scores are denoted by α_i while conformity scores are denoted by β_i.

Next, we fix an error or significance level $\epsilon \in (0,1)$, which is user-defined. Instead of the error, the user can specify a confidence level expressed in percentage. The error or confidence level is then used to obtain the calibration score (conformity or nonconformity score). In step 5, we will obtain the underlying model prediction $\tilde{y}$, which, together with the error prediction and the calibration score, will be used to obtain the prediction region.

Finally, two metrics are proposed in CP to evaluate the quality of the model by reflecting its ability to capture prediction uncertainty [12, 155]. The first is called *validity* and measures the coherence between ϵ and the actual error rate. The second one is known as *efficiency* and refers to the informativeness of the prediction region. i.e., the size of the output prediction sets. The average number of output labels for classification is used to measure the efficiency. The average length of predicted intervals is used to measure the efficiency in regression tasks.

To sum up, CP is based on the idea of predicting that a new instance will have a

label that makes it similar to the old instances (calibration set) in some specific way (calibration function) and using the degree to which the specific type of similarity holds (calibration score) within the old instances to estimate the confidence in the prediction [155]. The aim is to give subsets of Y large enough such that the true label y_n will fall in the subset. Therefore, instead of having point predictions (as is the case with a standard ML model), CP generates a range of more or less precise predictions, each labeled with a degree of confidence [12]. Each prediction is evaluated by calculating the efficiency and validity, which measure, respectively, how informative and reliable the predictions are. Figure 1.6 summarizes the basis of CP.

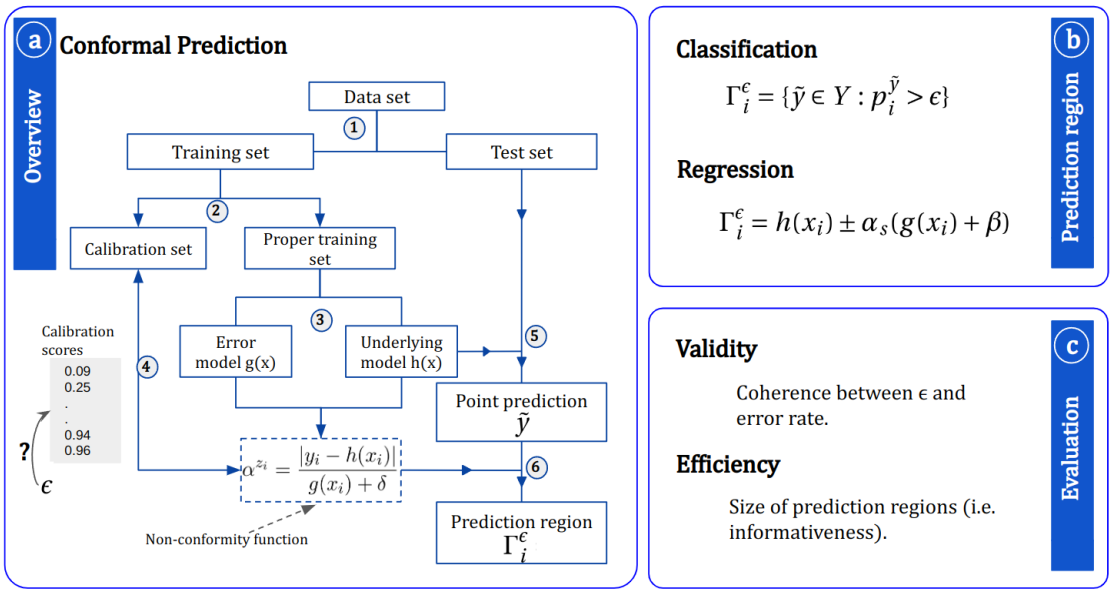


Figure 1.6.: **Overview of Conformal Prediction methodology and key components.** a) The CP framework predicts a new instance's label by assessing its (dis)similarity to old instances using a calibration function. In the so-called normalized CP, two models are built using the instances in the proper training set, an underlying model $h(x)$ and an error model $g(x)$. b) Using the underlying model, the error model, and the calibration score, CP generates a range of predictions with varying degrees of confidence to ensure the actual label falls within the prediction region. c) Predictions are evaluated by calculating efficiency (How informative are the predictions?) and validity (How reliable are the predictions?).

1.5.3.2. Approaches to Conformal Prediction

CP was initially developed as an online transductive framework, where predictions are made sequentially based on all previous instances rather than using a fixed batch. This differs from most previous work on machine learning, which uses an inductive and offline framework [12, 155]. In the offline inductive framework, a prediction rule is derived from a batch of old instances and then applied to new ones.

Although the online transductive CP (TCP) has some advantages over the offline inductive CP (ICP), such as adaptability to new information without requiring complete retraining, when it comes to handling large datasets, the ICP arises as an alternative,

whose computational efficiency is often much better [36]. However, this computational efficiency comes at a cost concerning the loss in predictive efficiency, leading to a weakened guarantee of validity [155]. Since the drug discovery process involves large data sets, we focus on the ICP rather than the TCP.

ICP for classification

Up to step 5, ICP for classification and regression shared the same outline. However, they differ in constructing the prediction region. In classification, for each possible label $\tilde{y} \in Y$ we compute their calibration score, i.e, $\alpha_i^{\tilde{y}} = f(x_i, \tilde{y})$. However, the numerical value of α_i does not, by itself, tell us how unusual f finds $(x_i, \tilde{y})$ to be [155]. For what, we need a comparison of the current α_i to other α_j. These "other α_j" are those calibration scores computed in step 4 and saved in the calibration table. Such comparison is made by computing the fraction

$$\frac{|\{j = 1, \ldots, q : \alpha_j \geq \alpha_i\}|}{q} \tag{1.1}$$

which lies between $\frac{1}{q}$ and 1; if it is small (close to its lower bound $\frac{1}{q}$ for a large q), then $(x_i, \tilde{y})$ is very nonconforming (an outlier). If it is large (close to its upper bound 1), then $(x_i, \tilde{y})$ is very conforming [103]. This fraction is called the *p-value* for $(x_i, \tilde{y})$. However, as pointed out by authors such as [8, 36, 103], this *p-value* is not to be confused with traditional p-values from statistics, since these p-values are the ranking of a test instance compared to known instances of each class.

Finally, in step 6, with a fixed error rate $\epsilon \in (0, 1)$, we compute the prediction set as follows:

$$\Gamma_i^{\epsilon} = \{\tilde{y} \in Y : p_i^{\tilde{y}} > \epsilon\}. \tag{1.2}$$

When performing binary classification, four possible outcomes can result from conformal prediction [12, 103]:

- Class Positive: {Active}, for instance.
- Class Negative: {Inactive}, for instance.
- Both classes: {Active, Inactive}.
- No class assignment: { }, (empty class).

The first two outputs, which correspond to the positive or negative class in this example, are the standard outputs of machine learning methods. If the output represents both classes, it means that additional information is necessary to differentiate between them. On the other hand, if the output represents an empty class, it indicates that the predicted instance falls outside the model's scope and should be determined experimentally. After experimentally determining its value, the instance can be incorporated into the updated model [8].

ICP for regression

In regression problems, the construction of the prediction region varies slightly from classification problems. In regression problems, we are not able to compute all possible labels for x_i, and therefore, we cannot compute the *p-value*. Instead, we fix the *p-value* and then find an appropriate $\alpha_i^{\bar{y}}$ [103, 155].

The way that we fix the *p-value* is by setting $s = \lfloor \epsilon(q+1) \rfloor$, which is the index of the $(1-\epsilon)-$percentile nonconformity score, α_s. Then, the prediction for a new example is (step 6):

$$\Gamma_i^\epsilon = h(x) \pm \alpha_s(g(x_i) + \delta) \tag{1.3}$$

where $g(x)$ is the error model and δ is a sensitivity parameter that determines the impact of normalization.

Equation 1.3 represents the version that includes the error model to control the predicted interval size. The calibration score and the error determine this size ($\alpha_s(g(x_i) + \delta)$). On the other hand, when the error model is not taken into consideration, the predicted interval is calculated based only on the calibration score α_s [12]:

$$\Gamma_i^\epsilon = h(x) \pm \alpha_s \tag{1.4}$$

In this case, the size of the interval remains fixed for all predictions since the calibration score is fixed for a given confidence level. However, having the same uncertainty level for all predictions is not always desirable, as some predictions may be easier or more challenging than others.

1.5.3.3. Advanced Conformal Prediction methods

Traditional conformal prediction methods are known to provide accurate measures of prediction uncertainty [12, 155]. However, they may face limitations if the test data distribution differs significantly from the training data. This can lead to reduced prediction accuracy and reliability. As a result, researchers are now exploring more advanced conformal prediction techniques, such as Aggregate Conformal Prediction (ACP) and Cross Conformal Prediction (CCP).

The ACP is an advanced version of the traditional conformal prediction that provides robust predictions. The key idea behind ACP is to use multiple inductive conformal predictors, each trained on a different data subset or using a different randomization strategy [28, 84]. By aggregating these diverse predictions, ACP not only improves prediction accuracy but also extends the range of applicability to heterogeneous datasets. This approach significantly reduces the risk of overfitting to specific data patterns, thereby boosting the reliability of prediction intervals. Such adaptability makes ACP a reassuring choice for scenarios where data variability is a primary concern. However, it is essential to note that ACP introduces increased computational complexity and

resource requirements due to the aggregation process, which can pose challenges, particularly with large datasets or complex models.

Similarly, CCP addresses distributional shifts between training and test datasets [154]. In CCP, the training data set is split into k non-overlapping sets. This is done in a manner similar to that of the k-fold cross-validation technique. Subsequently, k CP models are trained, each time using a different set as the calibration set [36, 146]. This approach is advantageous in real-world circumstances where data distributions may vary over time or across different environments, as it fosters confidence in the prediction performance.

The primary difference between ACP and CCP lies in their focus. The first emphasizes improving prediction reliability by combining outputs from multiple conformal predictors. In contrast, the second focuses on handling distributional differences between training and testing data to ensure prediction validity under varying conditions. Traditional CP methods remain valuable for their simplicity and straightforward implementation but may need more robustness and adaptability, such as that offered by ACP and CCP. The choice between these methods depends on the specific characteristics of the prediction task and the available resources and data.

1.6. Scope and Structure

This section outlines the fundamental topics involved in the thesis project at hand to provide a more comprehensive understanding of the project. As we saw, Chapter 1 of this thesis introduced four key themes - cancer, drug design, uncertainty quantification, and artificial intelligence (AI) - and aimed to understand their interplay. Section 1.2 provided an overview of cancer, including its definition, categorization, diagnosis, and treatment options. Section 1.3 defined drug discovery and presents the entire drug discovery and development process, focusing on ligand-based methods. Section 1.4 introduced uncertainty quantification, focusing on its impact on drug design, and presented different approaches to quantify and manage uncertainty. Finally, Section 1.5 discussed the role of AI in cancer research, particularly in drug design. The remaining work is organized into distinct chapters.

Chapter 2, Research Framework, provides a foundation for this thesis project. Section 2.2 explains the reasons for this project, while Section 2.3 presents a literature review of CP methods for drug design. We identify gaps and propose four research questions in Section 2.4. Finally, Section 2.5 outlines the project's objectives, aiming to develop reliable models for predicting potent molecules.

Chapter 3 of this thesis project presents the methods used to answer research questions. Section 3.2 covers data extraction and preprocessing of the NCI-60 Growth Inhibition Bulk Data. Section 3.3 describes the various machine learning models considered, and Section 3.4 presents the evaluation metrics used.

Chapter 4 presents the findings of this thesis project, which focuses on predicting the inhibitory activity of molecules on cancer cell lines of the NCI-60 panel. The chapter is derived from four studies conducted to address research questions. The studies involve developing a CP baseline model to establish a benchmark (Section 4.2) and optimizing its performance by exploring hyperparameter tuning, molecular representation, and assessing different CP methodologies (Section 4.3). The next stage prepared the data for real-world scenarios by exploring different molecular clustering methods (Section 4.4). The results of the previous steps are combined to train and evaluate CP-based models in both retrospective and prospective scenarios (Section 4.5).

Chapter 5 concludes this project and suggests potential future work.

2. Research Framework

Contents

2.1. Research framework overview 54
2.2. Motivation and problem statement 55
2.3. Literature review of the latest advances in Conformal Prediction 56
2.3.1. Prominent studies using Conformal Prediction in Drug Design . 56
2.3.2. Findings from the Literature Review 63
2.4. Research Questions 64
2.5. Objectives 65

2.1. Research framework overview

This chapter provides a comprehensive and organized foundation for the main body of this thesis project. In Section 2.2, we present the reasons that motivate this thesis project, which include the development of ML and CP models to predict potent molecules for different cancer types.

In Section 2.3, we conduct a state-of-the-art review of the latest advances in CP methods employed in drug design, particularly in QSAR modeling. This literature review aims to identify trends and critical gaps that exist in CP. For instance, the need for ML-based models that can generalize effectively to unseen data and the reliability of such predictions are some of the gaps we identify.

With the research gaps established, we propose four research questions (RQ) in Section 2.4 that will guide the project and provide a framework for our exploration. These research questions focus on the prediction accuracy of CP-based models, the importance of the features used as model input, the ability of the model to generalize to unseen compounds, and the possibility of adapting CP-based predictions to be integrated into existing pipelines to support prospective evaluations.

Finally, this chapter concludes with Section 2.5, where we present the objectives of this thesis project. These objectives are designed to align with the research questions and support our overall goal of developing robust and reliable models for predicting potent molecules. Figure 2.1 summarizes the contents of this chapter.

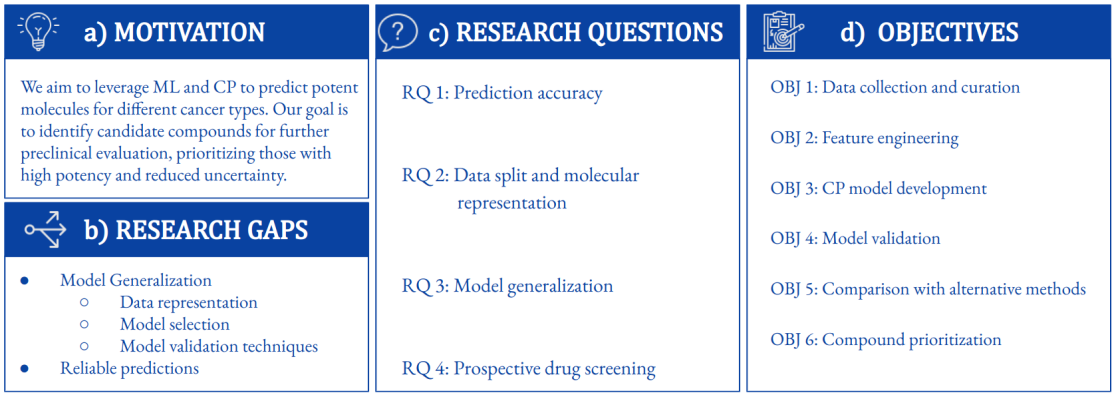


Figure 2.1.: **Overview of the Research Framework chapter**. This chapter lays the foundation for this thesis project. a) We aim to develop machine learning (ML) and conformal prediction (CP) models to predict potent molecules for various cancer types. b) Next, we conduct a state-of-the-art review of existing methods of CP applied to drug design. The literature review helped us identify research gaps. c) These research gaps led to the proposal of four research questions. and d) objectives.

2.2. Motivation and problem statement

As discussed in the first chapter, the drug discovery and development process is both time-consuming and expensive [67]. The initial stage of drug discovery involves the identification of a small molecule that can potentially become a drug candidate. One strategy for drug discovery that has regained popularity in recent years is Phenotypic Drug Discovery (PDD) [95]. This approach identifies compounds that can produce a desired therapeutic effect on the disease phenotype. Recent studies have shown that PDD can be a promising approach for discovering new drugs and accelerating the drug discovery process.

In parallel, AI has been widely used in the past few decades to develop predictive models for various applications in biomedicine and healthcare [14]. ML, a subfield of AI, has become an essential component in early drug discovery by creating quantitative structure-activity relationship (QSAR) models [13]. The preclinical data available today has made using ML-based models in drug discovery possible. One example is the National Cancer Institute (NCI-60) human tumor cell lines screen, which has been used by the cancer research community since 1990 to identify compounds with potential anticancer activity [100].

Although ML models have been widely used in drug design, several challenges remain, particularly in rigorous model validation for new compounds. Indeed, many studies overlook the reduction in predictive performance when a model, effective in one dataset, fails to generalize effectively to another, as discussed in [71]. Another challenge is the reliability of the predictive models at the instance level. While many supervised learning algorithms are employed to predict the phenotypic activities of molecules in cancer cell lines, they usually assign reliability to the entire model

(e.g., by calculating the RMSE between predicted and observed activities of test set molecules) rather than to individual molecules (e.g., a predicted activity interval where the observed activity of a given test set molecule is most likely to be) [65]. This can be problematic in drug design since it is essential to estimate the reliability of a prediction for each molecule.

This information is crucial for decision-making, such as selecting the most potent molecules with the most reliable predictions to reduce time and financial costs. Therefore, there is a need to develop models that can estimate the reliability of predictions for each molecule to improve the drug design pipeline. Hence, this thesis project aims to build predictive models using ML techniques and combine them with CP frameworks to predict the most potent molecules against different cancer types by screening compound libraries on cancer cell lines. Including a CP framework will allow us to obtain candidate compounds that could be further evaluated preclinically that are likely to be potent and have a lower uncertainty.

2.3. Literature review of the latest advances in Conformal Prediction

This section presents a comprehensive literature review of the latest advances in CP methods employed in QSAR modeling. Our objective is to provide an overview of current trends and identify the potential shortcomings that may exist in the intersection of QSAR models and CP methods. Ultimately, our goal is to contextualize the findings within the broader context of the drug design process to provide a more nuanced understanding of how these two fields intersect. We describe and discuss 12 prominent studies using CP for drug design to achieve this. These studies are arranged chronologically and summarized in Table 2.1.

2.3.1. Prominent studies using Conformal Prediction in Drug Design

In their work, Norinder et al. [103] proposed CP as an alternative to applicability domain determination challenges. They applied CP to ten publicly available datasets, consisting of four balanced binary classification problems and six regression problems. The sizes of the datasets used for classification and regression vary. The number of compounds in classification problems ranges from 132 to 397, while in regression problems, it ranges from 339 to 1,230. It is important to note that the datasets are relatively small, with six of them having less than 400 compounds. This can lead to challenges in capturing the complexity of diverse chemical environments encountered in drug discovery. Also, models trained on small data sets are more susceptible to overfitting and poor generalization [31]. The structures used in this study were described using signature descriptors, representing the environment of an atom up to a certain

height. In other words, the signature of a molecule is a vector of the occurrence numbers of atomic signatures [52]. The authors use Random Forest (RF) as the underlying model for both normalized ICP (regression) and Mondrian CP (classification). The study utilizes the standard data partitioning method from CP. Firstly, the data sets were randomly divided into a training set (80%) and a test set (20%). The training set was then further divided into two subsets: a "proper" training set (70%) and a calibration set (30%), selected randomly. The conclusions highlight the advantages of CP, emphasizing the mathematical framework and straightforward understanding of confidence levels compared to traditional applicability domain concepts.

Eklund et al. [48] conducted a study to evaluate the use of ICP on four large in-house datasets from AstraZeneca. The datasets included two binary classification problems and two regression problems, but the specific endpoints were not disclosed for legal reasons. Instead, the authors presented summary statistics on each dataset, comprising data from four assays conducted over five years at AstraZeneca. The size of the initial training set ranges from 1,780 to 10,264 structures, while the test set sizes range from 6,054 to 19,318. The structures used in this study were described using signature descriptors [52]. The authors use support vector machines (SVM) as the underlying model for offline ICP (classification and regression problems) and semi-offline ICP (regression). Additionally, the study analyzed the impact of different calibration set sizes on the reliability of confidence estimations. The findings revealed that inadequate calibration set sizes resulted in a higher variance of confidence, while insufficient sizes of the proper training set led to a downward bias in confidence. Despite drug discovery data often violating the independent and identically distributed (iid) assumption, the study concluded that CP significantly improved predictions compared to traditional QSAR applications.

In their study, Cortés-Ciriano et al. [39] analyze large-scale omics data to develop computational approaches for more efficient and targeted use of anticancer therapeutics. They use RF and SVM models, along with CP, to model pGI_{50} values for 17,142 compounds on 59 cancer cell lines from the NCI60 panel [100]. The models were trained using chemical and cell line profiling data as input features to predict the pGI_{50} values. However, compound-only (QSAR) models were exclusively trained using compound fingerprints as input features. The size of the fingerprints was set to 256 bits, and the maximum radius of the substructures considered was set to 2. For the CP evaluation, the dataset was split into an external set (20%) and a training set (80%), which was further divided into a calibration set (30%) and a proper training set (70%). The study demonstrates the potential of their approach to accurately predict drug activity on new cell lines and tissues, as well as the possibility of extrapolating to dissimilar compounds. However, the study needs more discussion on the efficiency (uncertainty) of the pGI_{50} predictions.

Table 2.1.: Prominent studies using Conformal Prediction (CP) for QSAR/QSPR. A chronological overview of major studies utilizing conformal prediction for drug design, highlighting the number of compounds examined in each study, the model validation strategy, and the types of conformal prediction employed. The subscript notation in the validation type column indicates the subsequent partitioning of the training set into the proper training and calibration set.

Dependent variable	**Learning task**	**Validation type**	**Input features**	**CP type**	**Dataset description**	**Refs.**
10 endpoints	Regression, Classification	Random splitting: $80_{70/30}/20$	Signature descriptors	ICP, MCP	10 datasets containing between 132 and 1,230 compounds	Norinder et al. [103]
4 endpoints	Regression, Classification	Random splitting	Signature descriptors	ICP, MCP	4 Astra Zeneca in-house datasets, starting from 1,780 molecules, progressing to 10,264	Eklund et al. [48]
pGI_{50}	Regression	Random splitting: $80_{70/30}/20$	MFP (r=2, bits=256)	ICP	17,142 compounds screened against 59 cancer cell lines from the NCI60 panel	Cortés-Ciriano et al. [39]
Active, Inactive	Classification	Random splitting: $70_{5CV}/30$	Signature descriptors	MCCP, CCP	18 datasets, from ExCAPE-DB, containing between 863 and 390,191 compounds	Sun et al. [146]
29 endpoints including pGI_{50}	Regression	Random splitting: 60/20/20	97 physiochemical descriptors (PCD)	ACP	29 datasets, from ChEMBL v19, containing between 155 and 4,325 compounds	Svensson et al. [147]

Continued on next page

Continuation of Table 2.1

Dependent variable	**Learning task**	**Validation type**	**Input features**	**CP type**	**Dataset description**	**Refs.**
pIC_{50}	Regression	Random splitting: 70/15/15	MFP (r=2, bits=2048)	DCP	24 datasets, from ChEMBL v23, containing between 479 and 5,207 compounds	Cortés-Ciriano et al. [35]
Active, Inactive	Classification	Stratified random splitting $80_{70/30}/20$	MFP (r=2, bits=2048) + 6 PCD	MCP	550 human protein targets, from ChEMBL v23, containing between 76 and 7,707 compounds. ChEMBL v24	Bosc et al. [20]
pGI_{50}	Regression	Generic Bemis-Murcko scaffold split $80_{80/20}/20$	ECFP6 FP (4,096 bits) + 9 PCD	ICP, DCP	Retrospective setting: ChEMBL25 dataset Semi-prospective setting: GC4 dataset	Parks et al. [110]
Inhibitors, non-inhibitors	Classification	Stratified random splitting: $80_{70/30}/20$	97 PCD	MCP	3 datasets containing between 667 and 1258 compounds	Alvarsson et al. [8]
Toxic, non-toxic	Classification	Stratified 10-fold splitting 70/20/10	MFP (r=4, bits=1024) + 97 PCD	MCP	Tox21 dataset containing between 7,081 and 9,219 compounds	Zhang et al. [165]
pGI_{50}	Regression	Random splitting: $90_{90/10}/10$	MFP (r=2, bits=256)	ICP	50,555 compounds screened against 60 cancer cell lines from the NCI-60 panel	Hernández et al. [65]
pIC_{50}	Regression, classification	Random splitting: $70_{5CV}/15/15$	Gene expression values	ICP	GDSC release 8.3	Lenhof et al. [82]

Sun et al. [146] proposed the Mondrian cross-conformal prediction (MCCP) method, which combines Mondrian inductive CP with cross-fold calibration. The study applied this method to 18 publicly available datasets extracted from the ExCAPE-DB database, which display varying levels of imbalance, ranging from 1:10 to 1:1000, regarding the ratio of active to inactive compounds. Signature descriptors [52] were generated for all the compounds. The study compared the performance of MCCP with cross-conformal prediction (CCP) methods, using SVM as the underlying model. The dataset was split into a 70% training set and a 30% external test set. The training set was partitioned randomly into five folds to estimate prediction regions of the test set compounds. The authors found that while CCP models cannot ensure label-conditional validity for all labels on imbalanced datasets, MCCPs can achieve both global and label-conditional validity, even on highly imbalanced datasets. The study also examined a compound similarity-based nonconformity measure, which was found to produce valid predictions but with notably inferior efficiency compared to nonconformity measures obtained from supervised learning. The authors concluded that MCCP can be a promising approach to estimating prediction confidence on large imbalanced bioactivity datasets.

Svensson et al. [147] conducted a study to investigate the use of aggregated CP (ACP) in modeling molecular properties, biological activity values (pIC_{50}), and toxicity. They explored various strategies to scale prediction intervals to optimize their efficiency. The study evaluated nine regression algorithms, including RF, Gradient Boosting, and Bayesian Ridge, to estimate prediction uncertainty on 29 publicly available datasets from ChEMBL version 19. The number of compounds in each dataset set ranges from 118 to 4,325, with 75% of the datasets having less than 1,112 compounds. Each compound was encoded using 97 RDKit descriptors. Using the ACP (200 CP models were aggregated), the authors found that the best results were obtained using the Random Forest (RF) as the underlying model. This model was used to explore the impact of six distinct nonconformity measures based on different methods to estimate the uncertainty of predictions for individual instances. Although the validity of the different nonconformity functions varied, all models exhibited an error rate close to the defined significance level. This highlighted that the CP models are calibrated regardless of the nonconformity metric. However, the authors found that the choice of nonconformity function significantly impacted the average prediction range (efficiency).

Cortés-Ciriano et al. [35] combined Test-Time Dropout methods and CP to introduce a framework called Dropout Conformal Predictors (DCP) for computing reliable errors. The study utilized 24 pIC_{50} datasets from ChEMBL version 23. The size of the data sets ranges from 479 to 5,207 compounds, each represented using Circular Morgan fingerprints, with a radius of 2 and a fingerprint size of 2,048 bits. The authors compared two algorithms, Deep Neural Networks (DNN) and RF. The datasets were randomly split into training (70%), validation (15%), and test (15%) sets to build the DNN and RF models. The results demonstrate the validity and efficiency of DCP, correlating true values within predicted intervals with confidence levels. The authors also found that these predictors outperform RF-based CP, yielding narrower predicted

confidence intervals. Additionally, they discovered that retrospective virtual screening experiments show comparable retrieval rates of active compounds between DCP and RF-based CP.

A case study conducted by Bosc et al. [20] compared the performance of CP to that of traditional QSAR methods for predicting target-ligand binding in drug discovery. They used 550 human protein targets from ChEMBL (version 23) and calculated molecular descriptors, which included Morgan fingerprints (radius of 2 and a length of 2048 bits) and six physicochemical descriptors (molecular weight, number of hydrogen bond donors, number of hydrogen bond acceptors, number of rotatable bonds, lipophilicity and the topological polar surface area). The authors use the Mondrian CP approach since they are working with classification problems. Two models were created for each target, one QSAR model and one MCP model, with the RF algorithm set as the underlying model for the MCP. The dataset was divided into training (80%) and test (20%) sets by applying a stratification sampling on the activity class. For MCP, the training set is randomly divided into a proper training set (70%) and a calibration set (30%), using a splitting procedure repeated 100 times. Both models were evaluated on new data (ChEMBL version 24) to simulate a real-world application. The results showed that both QSAR and MCP approaches offer good predictive performance, with MCP generally outperforming QSAR for most of the targets analyzed.

In the research conducted by Parks et al. [110], the authors compared the RF and Feedforward Neural Network (FFN) proteochemometric (PCM) models for predicting pIC_{50} measurements from the ChEMBL version 25 database. The dataset was preprocessed to represent each compound with 4,096-bit length ECFP6 fingerprints, molecular weight, topological surface area, number of hydrogen donors, number of hydrogen acceptors, LogP, heavy atom count, number of rotatable bonds, and ring count. The final dataset comprised 302,325 data points of 213,502 unique SMILES strings across 940 unique UniProt IDs. The study assessed CP for calibrating prediction intervals in a retrospective and semi-prospective test using the Grand Challenge 4 dataset as an external test set. The results from comparing the RF and FNN showed that while RF prediction intervals are less efficient, they exhibited better validities than FFN. Moreover, both models demonstrated good overall quality validities. The authors conclude that both RF and FFN models showed good retrospective performance but faced challenges in the semi-prospective setting. However, CP prediction intervals proved well-calibrated both retrospectively and semi-prospectively, indicating their utility in guiding hit discovery and lead optimization campaigns.

Alvarsson et al. [8] conducted a study exploring the use of conformal prediction as a framework to address the challenge of quantifying the reliability of predictions in drug discovery. They used three datasets (Bcrp, BSEP, and Pgp inhibition) containing between 667 and 1,258 compounds. Each compound was described using 97 RDKit descriptors. The researchers conducted classification modeling for the three transporters, combining training and test sets from the original publication. They then selected an external test set using stratified random sampling. Mondrian Conformal

Prediction (MCP) models were trained based on underlying RF models. The study found that lower error rates led to more multiclass predictions, while higher error rates resulted in more empty prediction sets.

Zhang et al. [165] conducted a study combining deep learning-based (DL) predictors with the CP framework to develop models with well-defined uncertainties for toxicity predictions. To achieve this, they employed several deep feedforward neural networks and graph neural networks within a CP setting. The authors also compared DL with decision tree-based algorithms like RF and LightGBM models. The models were evaluated using data from the Tox21 challenge, which includes toxicities of compounds on 12 nuclear receptors and stress response-related targets measured in *in vitro* test systems. The study categorized the datasets into active and inactive compounds, treating the problem as a binary classification issue. Each structure was represented with 97 structural and physicochemical properties [8] and extended-connectivity Morgan fingerprints calculated with radius=4 and hashed to 1024 bits. To train the models, 10-fold cross-validation was used, with 10% for validation, 20% for calibration, and 70% for proper training. Overall, the results found by the authors demonstrate the effectiveness of CP in providing toxicity predictions with a high level of confidence. CP offers statistical guarantees on model performance and improved predictions of the minority class compared to underlying models.

In our previous work [65], we investigated the impact of CP on predicting the inhibitory activity of molecules on cancer cell lines within the NCI-60 panel [100] while analyzing variability across cancer types. We used a 256-bit Morgan fingerprint with a radius of 2 to represent molecules. After preprocessing, we obtained 50,846 unique NSC IDs and 2,707,434 pGI_{50} measurements corresponding to 50,555 unique molecules. For each cell line, we built regression models using Python implementations of RF and extreme gradient boosting (XGB) algorithms. We applied a 90-10 random partition, with 10% of molecules reserved as a test set and the remaining 90% used for training. We further subdivided the training set into proper training (90%) and calibration sets (10%). Our results showed that underlying models can predict pGI_{50} of test set molecules with reduced accuracy for potent molecules. However, relaxing the confidence level improves CP-valid predictions and enhances the prediction accuracy of pGI_{50} errors. We found that molecules deemed reliable by conformal prediction are substantially better predicted than others. Our study suggests the promise of computational tools in guiding phenotypic drug design.

In a recent preprint by Lenhof et al. [82], the authors discussed two significant challenges in optimizing anticancer drug treatments through machine learning systems. These challenges are making reliable predictions of anticancer drug responses and prioritizing drugs for specific cancer samples. To address these challenges, the authors proposed a flexible approach to predicting and prioritizing drug sensitivity using conformal prediction and presented a Python framework applicable to various methods, including classification, regression, and joint classification and regression methods, the latter is called the SAURON-RF which stands for SimultAneoUs Regres-

sion and classificatiON Random Forests. Their proposed approach combines the SAURON-RF and CP and is applied to the GDSC dataset, one of the largest publicly available pharmacogenomic databases. This dataset includes processed gene expression data, pre-computed logarithmized IC_{50} drug responses, and raw drug responses. The dataset was divided into a training (70%), calibration (15%), and test (15%) set. The training set was further divided into five subsets for 5-fold cross-validation (CV). The authors found that CP significantly improved the predictions by reducing false predictions while retaining correct ones, which helped in drug prioritization.

2.3.2. Findings from the Literature Review

When applied to the drug discovery process, two significant challenges with AI-based models, whether ML or CP, are their performance and their ability to generalize to new unseen compounds. For example, models trained on small data sets are more susceptible to overfitting and poor generalization [31]. Therefore, the choice of the dataset plays an essential role in developing AI-based models.

The first analysis point in the literature review focused on data representation, another factor impacting model performance and generalizability. How the data is represented affects the model's ability to capture relevant patterns and generalize to new data. Signature descriptors, MFP, and PCD emerged as common choices, with each approach having advantages and limitations. Signature descriptors provide a comprehensive structural representation, MFP captures molecular features, and PCD offers physicochemical insights. However, achieving a balance between informativeness and model complexity is a challenge.

The second aspect analyzed in this literature review is the choice of algorithms, which impacts the model's ability to learn from the data and make accurate predictions. RF and SVM methods are commonly applied to QSAR modeling. RF offers versatility and robustness, particularly in handling complex relationships within the data. SVM, on the other hand, excels in high-dimensional spaces. However, both algorithms have limitations. RF may be susceptible to overfitting, while kernel and parameter choices can influence SVM's performance. Exploring a broader range of ML algorithms, including DL approaches, may lead to better predictive performance and uncover novel insights.

Lastly, an essential factor to consider is the model validation technique. Most studies analyzed in this literature review use random partitioning, with or without stratification, which may affect the model's generalization capabilities. Random partitioning can lead to imbalances in the distribution of critical features, potentially hindering the model's ability to extrapolate to new data. Stratified partitioning, conversely, ensures a more representative split but may introduce biases depending on the stratification criteria. Therefore, more complex model validation techniques reflecting real-world conditions should be analyzed.

In summary, the performance and generalization of CP models can be significantly improved by accurately representing data, selecting the appropriate model, and employing effective model validation techniques. These areas will guide the research questions proposed in the next section, ensuring that the CP models are reliable and efficient.

2.4. Research Questions

The research questions in this thesis are tailored to the application of CP in drug discovery. These research questions are intended to guide the creation of ML and CP models for predicting potent molecules for different cancer types. The build models are designed to identify candidate compounds, prioritizing those with high potency and lower uncertainty.

RQ 1. How accurately can CP-based models predict pGI_{50} values for different compounds in cancer cell lines?

Understanding the accuracy of CP models in predicting pGI_{50} values is crucial for assessing the reliability of these predictions. One of the main advantages of leveraging CP models is that, unlike ML models, an uncertainty is assigned at the instance level. That is, for each molecule, we can estimate the reliability of the delivered prediction and focus on the most reliable ones. With reliable predictions, researchers could identify compounds with potential therapeutic efficacy, which could significantly impact the efficiency of the drug discovery process.

RQ 2. How does the data split and molecular representation choice impact pGI_{50} predictions?

The representation of a compound directly impacts the accuracy of pGI_{50} predictions. A precise representation ensures that the model captures the essential features and characteristics of the molecule. If the representation captures the fundamental properties of molecules across different datasets, the model is more likely to perform well on new molecules or datasets that it has not encountered during training. To answer this question, we will focus on MFPs and PCDs.

RQ 3. Can the CP model generalize well to predict pGI_{50} values for dissimilar compounds in cancer cell lines?

The ability of CP models to predict pGI_{50} values for new compounds is essential for their practical use in decision-making problems. If the model can accurately predict outcomes for unseen data, it could facilitate the identification of potential drug candidates for further evaluation in experimental studies. This generalization ensures that the CP-driven approach remains robust and applicable beyond the training dataset.

RQ 4. How can CP be integrated into existing workflows and decision-making processes to support prospective drug screening?

While individual prospective validations of CP models are yet to be reported, some studies have shown a prospective benefit of CP in combination with other virtual screening methods [110, 150]. However, it is essential to conduct prospective evaluations to obtain a more precise assessment of the model's performance in practical applications. Integrating CP into existing decision-making processes to support prospective screenings can enhance decision-making by providing reliable confidence estimates for predictions. This helps prioritize and focus resources on compounds that are most reliable and likely to succeed, leading to more informed decisions in drug discovery.

2.5. Objectives

We have outlined a set of specific objectives that should be achieved to provide answers to the research questions. These objectives will guide our research process and serve as measurable outcomes to assess our progress.

- Data collection and curation: Download the NCI-60 growth inhibition data. The NCI-60 data must undergo a preprocessing stage (data curation) to ensure high-quality data. It involves cleaning, filtering, and validating the data to ensure reliability and accuracy.
- Feature engineering: Explore the different ways of representing the molecules and the impact that such representation has on the developed models.
- CP Model Development: Develop a robust CP model tailored for predicting pGI_{50} values of compounds based on phenotypic screening data.
- Model Validation: Rigorously validate the CP model's performance through cross-validation and benchmarking against existing methods for predicting pGI_{50}.
- Comparison with traditional methods: Compare the efficiency and effectiveness of CP-driven models with traditional techniques to highlight the advantages of the CP approach.
- Compound Prioritization: Utilize the CP model to prioritize compounds from large chemical libraries based on their predicted pGI_{50} values, identifying potential candidates for further investigation.

3. Materials and Methods

Contents

3.1. An overview of the development of this project 66
3.2. NCI-60 panel 67
3.2.1. Database description and preprocessing 67
3.2.2. Molecular descriptors 70
3.2.2.1. Morgan Fingerprints 70
3.2.2.2. Physicochemical descriptors 71
3.3. Machine Learning algorithms 73
3.4. Model Evaluation 74
3.4.1. Evaluation Metrics for Machine Learning Models 74
3.4.1.1. Regression metrics 74
3.4.1.2. Classification metrics 75
3.4.2. Evaluation Metrics for Conformal Prediction Models 75

3.1. An overview of the development of this project

This chapter presents a comprehensive guide to the procedures and techniques employed in this thesis project, composed of four studies designed to answer the research questions posed in Section 2.4. It will cover the various stages of the project's development and provide insights into the key concepts and methodologies used.

For our thesis project, we started by preparing the NCI-60 Growth Inhibition Bulk Data, which has been used for over 30 years to discover compounds with potential anticancer activity. Section 3.2 presents some insights from this dataset and explains data extraction and preprocessing as the initial steps (Figure 3.1-a). These initial steps are essential for ensuring data quality. We utilized this dataset for all the studies carried out in this project, therefore, data extraction and preprocessing are common steps.

Once the data is preprocessed, we represent it using Morgan fingerprints or a combination of Morgan fingerprints and physicochemical descriptors (Figure 3.1-b). The selection of feature sets relies on the particular goals of each study involved in this project and plays a critical role in delivering accurate and reliable results. Chapter 4 contains a detailed description of the specific configuration of these feature sets. These molecular descriptors are introduced in Subsection 3.2.2.

In this project, we employ conformal prediction, a framework that can be integrated with machine learning models. Section 3.3 describes the various models we considered for this project. However, since each study uses different models, we have provided the specific configurations of the models in Chapter 4. The molecular descriptors introduced in 3.2.2 are the input features for ML and CP models (Figure 3.1-c). The trained models are evaluated by computing various metrics for both regression and classification problems. This evaluation is complemented by the evaluation metrics used in conformal prediction (Figure 3.1-d). Section 3.4 presents these evaluation metrics.

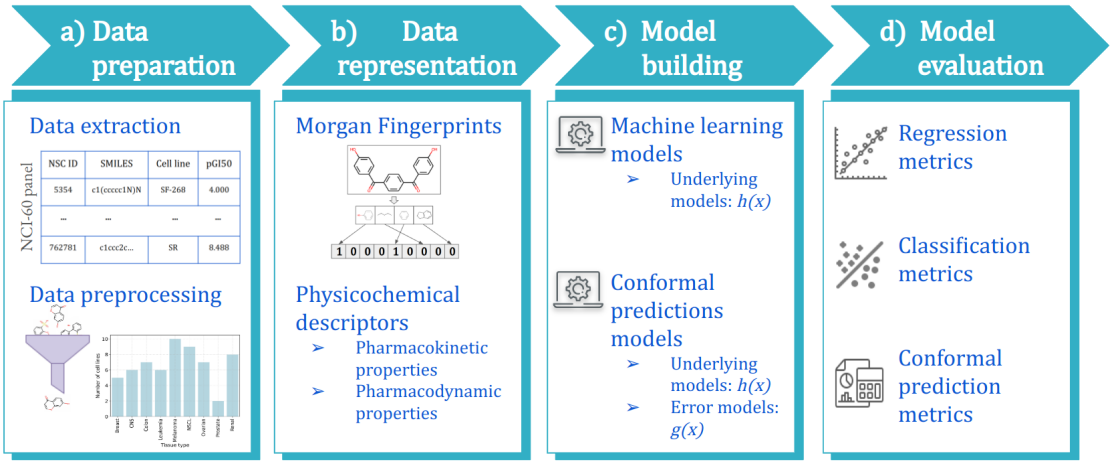


Figure 3.1.: **Overview of the development of this thesis project**. This project focuses on the NCI-60 panel, a dataset for discovering compounds with potential anticancer activity. a) The first step is data extraction and data preprocessing. b) Compounds within the NCI-60 panel are represented using either Morgan fingerprints or physicochemical descriptors. c) These descriptors are inputs for machine learning and conformal prediction models. d) The model evaluation stage comprises regression, classification, and conformal prediction metrics.

3.2. NCI-60 panel

3.2.1. Database description and preprocessing

The NCI Development Therapeutics Program (DTP) offers services and resources to facilitate discovering and developing new cancer therapeutic agents for research communities worldwide. Indeed, since 1990, the cancer research community has used the NCI-60 Human Tumor Cell Lines Screen to find compounds with potential anticancer activity [100]. The NCI-60 panel is designed to screen up to 7,000 small molecules (synthetic or purified natural products) per year, utilizing 60 different human tumor cell lines to identify and characterize those compounds with growth inhibition or killing of tumor cell lines [133]. These 60 different human tumor cell lines comprise 9 cancer types: leukemia (LE), melanoma (MEL), cancers of the lung (NSCL), colon (CO), brain (CNS), ovary (OV), breast (BR), prostate (PR), and kidney

cancers (RE) (Figure 3.2-a).

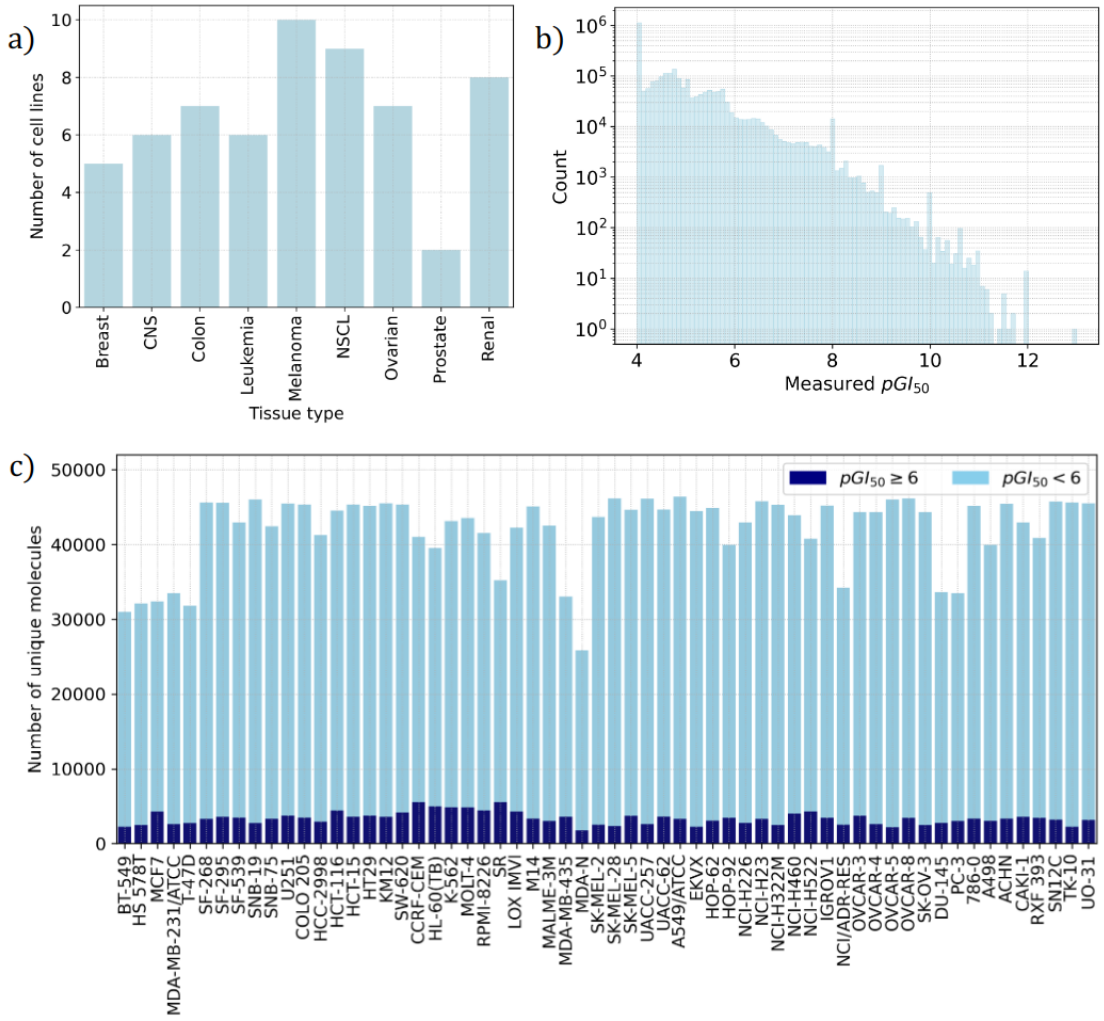


Figure 3.2.: **NCI-60 Growth Inhibition Bulk Data description.** a) The dataset utilized in this study is from the NCI-60 panel, which contains 60 distinct cell lines representing nine different types of cancer. b) Distribution of the pGI_{50} measurements in the molecules across the 60 cell lines for a total of 2,707,434 data points, where potent molecules ($pGI_{50} \geq 6$) are rather scarce. c) Distribution of the number of unique molecules tested per cell line. A color code is assigned to show the number of potent molecules ($pGI_{50} \geq 6$) per cell line.

Each compound submitted to the NCI-60 panel for testing and evaluation is identified with a unique registration number called the National Service Center (NSC) ID. According to the NCI-60 Screening Methodology, all compounds screened are initially tested in a one-dose assay on the full NCI-60 panel. Compounds that fulfill a predetermined threshold inhibition criteria within a minimum number of cell lines are further evaluated against the NCI-60 panel in a 5-dose assay [133]. Historical DTP screening data was carefully analyzed to determine the threshold criteria for efficiently capturing compounds with anti-proliferative activity. Thus, the threshold may be

updated as additional data becomes available. The results of these experiments are available for download from the DTP Bulk Data for Download website.

We used the NCI-60 Growth Inhibition Bulk Data for this thesis project, from where we modeled the pGI_{50} of molecule-cell line pairs. The pGI_{50} is defined as the negative logarithm of the half-maximal inhibitory concentration of the molecule on the cell line. A summary of the additional information for the NCI-60 Growth Inhibition Data is provided in Table 3.1.

Table 3.1.: The NCI-60 Growth Inhibition Data, which includes the following fields, is available for download in comma-delimited file format. This data can be downloaded from https://wiki.nci.nih.gov/display/NCIDTPdata/NCI-60+Data+Download+-+Previous+Releases

File column headers	Description
NSC	The NCI's internal ID number
CONCUNIT	Concentration unit either M for molar or u for μ g/ml
LCONC	Log of the highest concentration tested
PANEL	Panel name for the cell line
CELL	Cell line name
PANELNBR	Panel number of the cell line
CELLNBR	Cell number of the cell line
GI_{50}	Endpoint calculated from concentration curves
INDN	number of tests for this NSC and cell line
TOTN	Maximum number of tests for this NSC
StdDev	Standard Deviation of the Log_{10} of the results averaged across all tests for this NSC and cell line

High-quality data is essential for developing accurate ML/CP models in drug design. Therefore, data cleansing is necessary to ensure reliable data when generating these models. The data preprocessing steps are similar in all the projects presented in the Results section, and they can be found in Appendices A and C. Here, we summarize the main steps of this preprocessing. As part of the data-cleaning process, we remove low-activity molecules ($pGI_{50} < 4$). In a therapeutic context, drugs with low potency may not achieve sufficient efficacy at practical doses, making them less likely to be effective in treating the targeted disease or condition.

Additionally, removing compounds with low activities helps enhance the quality and reliability of the dataset used for analysis and modeling. Next, we calculated the mean pGI_{50} if multiple 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. Then, SMILES (Simplified Molecular Input Line Entry System) were standardized using the Molvs package. These steps are common to both projects, [65] and [64].

Depending on the specific requirements and goals of the projects, each molecule is represented by a Morgan fingerprints (MFP) [123] alone or a combination of MFP and physicochemical descriptors (PCD). We will discuss these molecular descriptors further in the next section. After the preprocessing stage, 2,707,434 data points remain

(50,555 small molecules screened against 60 cancer cell lines), which represents a matrix completeness of 89.25% (Figure 3.2-b).

The 60 cell lines in the NCI-60 panel have been extensively screened against an average of 45,315 compounds, with a standard deviation of 5,138. This large dataset size per cell line ensures that CP models trained on this data are reliable and robust for predictive analysis. However, the proportion of compounds exhibiting high potency is very low. Indeed, each cell line shows high potency ($pGI_{50} \geq 6$) in only 4 to 15% of the compounds screened (Figure 3.2-c). This sparsity of potent molecules is a major obstacle in identifying and characterizing potent compounds, which can significantly impact the performance of CP models trained on this dataset.

3.2.2. Molecular descriptors

A primary objective in computer-aided drug design research is to differentiate between drug-like and non-drug-like molecules. [66, 129]. Drug-likeness depends on several molecular properties that influence pharmacodynamics and pharmacokinetics, such as absorption, distribution, metabolism, and excretion in the human body. To perform computational studies, molecular properties must be accurately and quantitatively represented as molecular descriptors while preserving their chemical information. The selection of appropriate molecular descriptors is crucial for predicting the drug-likeness of a molecule during screening processes [93].

In our analysis, we utilized two types of molecular descriptors: Morgan fingerprints (MFPs), which encode structural information based on atom neighborhoods and bond connectivity patterns within a specified radius around each atom [123], and physico-chemical descriptors (PCDs), which quantify the physical and chemical properties of molecules such as molecular weight, lipophilicity, and hydrogen bonding capacity [46]. By leveraging these diverse molecular descriptors, we aim to comprehensively characterize the chemical space and structural diversity of the NCI-60 dataset. In this section, we describe the generalities of both descriptors, while the particular configurations for each are presented in Appendices A, B, and C, and revisited in Chapter 4.

Finally, the RDKit cheminformatics library [118] was employed to compute both molecular descriptors, MFPs and PCDs. RDKit is an open-source toolkit for cheminformatics, widely used in computational chemistry and drug discovery, offering extensive capabilities for molecular representation, substructure analysis, and descriptor calculation.

3.2.2.1. Morgan Fingerprints

Molecules from the NCI-60 panel can be encoded using SMILES notation. Different algorithms can process SMILES strings to generate fingerprints [86]. These algorithms analyze structural information within the SMILES and assign binary values to specific

features or substructures. Chemical fingerprints, represented as high-dimensional vectors, are frequently employed in similarity-based virtual screening [22, 60]. Some feature selection [69, 143] is usually used to identify subsets of bits from a fingerprint, leading to optimal performance in predictive models.

Traditionally, fingerprint descriptors are represented as bit strings, in which each bit indicates the presence (1) or absence (0) of a particular molecular feature. Binary fingerprints can be hashed or keyed based on substructural features [60]. Circular fingerprints, proposed in 1965, capture novel fragments and generate single, separate circular structural fragments [163]. The algorithm centers on each compound's nonhydrogen atom or molecular fragment and extends the fragment to its neighbors based on predefined rules. Common types include ECFPs, FCFPs, Molprint2D, and Molprint3D.

Fingerprints have both advantages and drawbacks in their application. One of the main advantages is that they rely on a 2D representation of molecules, which allows for rapid computation, even for extensive compound collections like combinatorial libraries. While fingerprints do not consider 3D information, a study comparing 2D and 3D descriptors found that 2D parameters outperformed their 3D counterparts, contrary to initial expectations [85].

3.2.2.2. Physicochemical descriptors

PCDs describe a molecule's physical and chemical properties in cheminformatics. They are also important in drug design to predict and optimize drug behavior in the body. Sometimes, PCDs can be derived from SMILES strings and fingerprints, or vice versa.

Designing effective drugs that produce minimal side effects and do not cause toxic reactions requires consideration of pharmacokinetic (PK) and pharmacodynamic (PD) parameters. These parameters are crucial for ensuring that drugs remain effective even at minimum concentrations [136]. PK involves the study of how drugs are absorbed, distributed, metabolized, and excreted (ADME) within the body, while PD is the study of the magnitude of drug response [124, 136].

In this thesis project, we calculated 7 PCDs for each molecule to cover both the PK and PD aspects of drug action. PCDs such as molecular weight, lipophilicity (logP), hydrogen bond characteristics (e.g., number of hydrogen bond donors/acceptors), and structural features like the number of aliphatic and aromatic rings are often associated with pharmacokinetic properties [124, 136]. These descriptors are summarized in Table 3.2 and are briefly discussed below.

Molecular weight is the sum of atomic weights of all elemental atoms in a substance, usually measured in g/mol [46]. LogP is related to the drug lipophilicity and is crucial for drug solubility, absorption, and tissue penetration [46, 129]. Hydrogen bond donors give away hydrogen atoms, while hydrogen bond acceptors are ready to re-

ceive hydrogen atoms to form hydrogen bonds [151]. These first four descriptors are considered in Lipinski's rule of five [145].

Table 3.2.: Brief description of physicochemical descriptors (PCDs) used in this study. These descriptors cover the pharmacokinetic and pharmacodynamics aspects of drug action. Lipinski's rule of five considers the first four descriptors, which are directly related to the pharmacokinetic properties. The total polar surface area of molecules is related to their pharmacodynamic aspect.

PCD	Abbr.	Description	Refs.
Molecular weight	Mw	It is the sum of the atomic weights of all atoms in a molecule. Molecular weight can influence factors such as absorption, distribution, and metabolism of a drug.	[46]
Lipophilicity	LogP	Lipophilicity refers to the ability of a molecule to dissolve in lipids or non-polar solvents. It is often quantified using the octanol-water partition coefficient (LogP). Lipophilicity can affect a drug's ability to cross cell membranes and penetrate tissues.	[129], [46]
Number of hydrogen bond donors	HBD	The count of functional groups or atoms in a molecule that can donate a hydrogen atom to form a hydrogen bond with another molecule.	[151], [46]
Number of hydrogen bond acceptors	HBA	The count of functional groups or atoms in a molecule that can accept a hydrogen atom to form a hydrogen bond with another molecule.	[151], [46]
Total polar surface area	TPSA	It is based on the polarized atoms of the molecule, that allows estimation of transport properties and of the passive molecular transport through membranes of the drug.	[29], [46]
Number of aromatic rings	AroR	An AroR is a special type of ring-shaped structure in a molecule with alternating double and single bonds between carbon atoms. The number of AroR refers to how many of these stable ring structures are present.	[122]
Number of aliphatic rings	AlkR	An AlkR are made up of carbon atoms bonded together, in a cyclic arrangement with only single bonds. The number AlkR in a molecule indicates how many of these simple ring structures are present.	[122]

The number of aromatic and aliphatic rings are descriptors of the presence and type of ring structures within a molecule [121, 122]. Aromatic rings are characterized by alternating double bonds and stability, while aliphatic rings consist of carbon atoms bonded with single bonds in a cyclic form. In [121], it was determined that oral drug candidates with fewer aromatic rings are likely to be more suitable for development. Additionally, molecules containing more than three aromatic rings tend to have lower developability, increasing the risk of failure during drug development. The molecules' total polar surface area (TPSA), which quantifies the surface area

available for polar interactions, is an example of a PD descriptor. TPSA can affect the ability of a molecule to form hydrogen bonds with biological targets, potentially impacting binding affinity and potency. Overall, the combination of MFPs and PCDs can provide a comprehensive set of information about a molecule and can be used to predict its properties and behavior in various applications.

3.3. Machine Learning algorithms

In this thesis project, we utilized the Python implementations of four different ML algorithms, namely k-Nearest Neighbors (kNN), Linear Regression (LR), Random Forest (RF), and Extreme Gradient Boosting (XGB), from the widely used scikit-learn library. These ML algorithms are widely used and have proven effective in various applications (see Table 2.1). These algorithms have been described in our previous work [64, 65] and can be found in Appendices A and C, respectively. For convenience, we summarize the main features of these algorithms in Table 3.3.

Table 3.3.: Summary of the machine learning (ML) models used in this thesis project. The "Used as" column describes whether the algorithm has been used as an underlying ($h(x)$) or error model ($g(x)$) while using conformal prediction. The "Comments" column includes additional information about using the corresponding algorithm.

Algorithm	Description	Used as	Comments
kNN	It predicts the class or value of a data point by averaging the values of its k nearest neighbors	$g(x)$	Default error model in [102]
LR	It aims to minimize the sum of squared differences between observed and predicted values	$h(x)$ & $g(x)$	Comparing linear vs non-linear models
RF	It builds multiple decision trees during training and outputs the average prediction of the individual trees	$h(x)$ & $g(x)$	Promising ML model
XGB	It is an implementation of gradient-boosted decision trees designed for speed and performance	$h(x)$ & $g(x)$	Promising ML model

These four algorithms were chosen for this project for various reasons. For instance, the kNN algorithm was selected since it is the default error model in the NonConformist Python package [102]. By using this algorithm, we can measure the impact of using Python packages without modifications. The LR model was selected to compare the performance of linear and nonlinear models. As discussed in Section 2.3, the RF model is commonly used in QSAR problems, which is why it was chosen for analysis in this project. Lastly, the XGB model was selected for its performance in similar problems.

3.4. Model Evaluation

Assessing the performance of an ML-based model is essential to determining its efficacy in solving practical problems. When dealing with regression or classification tasks, various metrics measure the model's predictive ability and generalization capacity. This section explores various commonly used evaluation metrics for regression and classification scenarios. We also complement the previous explanation of the metrics used in CP.

3.4.1. Evaluation Metrics for Machine Learning Models

3.4.1.1. Regression metrics

We employ three well-known metrics briefly described below to evaluate the performance of regression problems. The first metric is the Root Mean Squared Error (RMSE), which measures the average deviation of predicted pGI_{50} values ($\hat{y}_i$) from the actual pGI_{50} values (y_i). It can be mathematically represented as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3.1}$$

where n is the number of observations. RMSE values range from 0 to infinity, with 0 representing a perfect fit where predicted values perfectly match the actual values. Therefore, the RMSE tells us how far off our predictions are from the actual values, on average.

The second regression metric calculated is the Pearson's correlation coefficient (R_p). It measures the linear relationship between the predicted and actual values as follows:

$$R_p = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{3.2}$$

where x_i and y_i represent the actual and predicted values, respectively, and $\bar{x}$ and $\bar{y}$ denote their respective means. The R_p coefficient ranges from -1 to 1. A coefficient close to 1 signifies a strong positive linear relationship, -1 indicates a strong negative linear relationship, and 0 suggests no linear relationship.

The next metric is Spearman's correlation coefficient (R_s), which assesses the monotonic relationship between two variables. Unlike Pearson's correlation, Spearman's correlation is based on the ranks of the values rather than their actual values. R_s can be mathematically represented as:

$$R_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \tag{3.3}$$

where d_i represents the difference between the ranks of corresponding pairs of values and n is the number of samples. R_s ranges from -1 to 1. A value close to 1 indicates a strong positive monotonic relationship, -1 signifies a strong negative monotonic relationship, and 0 implies no monotonic relationship.

3.4.1.2. Classification metrics

In this thesis project, we selected a specific threshold for defining active compounds based on the sensitivity and specificity of ML and CP models. This process allows us to categorize the predictions and calculate the classification metrics accordingly. A threshold of 6 pGI_{50} units is used to determine whether predictions are positive (active) or negative (inactive). A prediction is positive if it is higher than or equal to 6 ($\hat{y}_i \geq 6$); otherwise, it is negative ($\hat{y}_i < 6$). In addition, Norinder et al., [103] also identified this threshold value as suitable for distinguishing active compounds. After this step, predictions are summarized as True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

The first metric considered in this case is the hit rate (HR), which measures the proportion of TP in compounds identified as positive. The hit rate is defined as:

$$HR = \frac{TP}{TP + FP} \times 100\% \tag{3.4}$$

The hit rate allows us to prioritize compounds, aligning with the objective of virtual screening to identify potential leads.

The last metric is the Matthews Correlation Coefficient (MCC), which provides a balanced measure by considering TP, TN, FP, and FN as follows:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{3.5}$$

MCC values range from -1 to 1. A score of 1 indicates perfect classification, 0 denotes random classification, and -1 represents total disagreement between prediction and observation.

3.4.2. Evaluation Metrics for Conformal Prediction Models

Section 1.5.3 briefly explains how the CP predictions are evaluated. In this section, we explore the evaluation of regression problems in more detail. The two metrics defined in CP are efficiency and validity [12, 155]. The efficiency refers to the informativeness of the prediction region. i.e., the average size of predicted intervals. The validity measures the coherence between ϵ and the actual error rate.

In CP, a single molecule can be predicted by considering different confidence levels, such as 80% or 90% confidence. The interval for each of these predictions is centered

on the point prediction $\hat{y}$ and extends up to the value $\alpha_s g(x)$, where α_s is the non-conformity score and $g(x)$ is the predicted error. The most effective predictions are those where the interval is smaller since the model is more "confident" of the prediction (Figure 3.3-a).

Note that efficiency does not consider whether the predicted interval contains the true value, which is why the validity of the predictions is calculated. Validity refers to whether the model has met the requested error, and if so, the model is said to be well-calibrated. If not, the observed error may be greater than expected, so the model needs more information to derive reliable predictions. In the case of a lower error than expected, although this is not considered a calibrated model since it is not strictly complying with the requested error, this scenario can be understood as an advantage since the model is "wrong" fewer times than the maximum allowed (Figure 3.3-b).

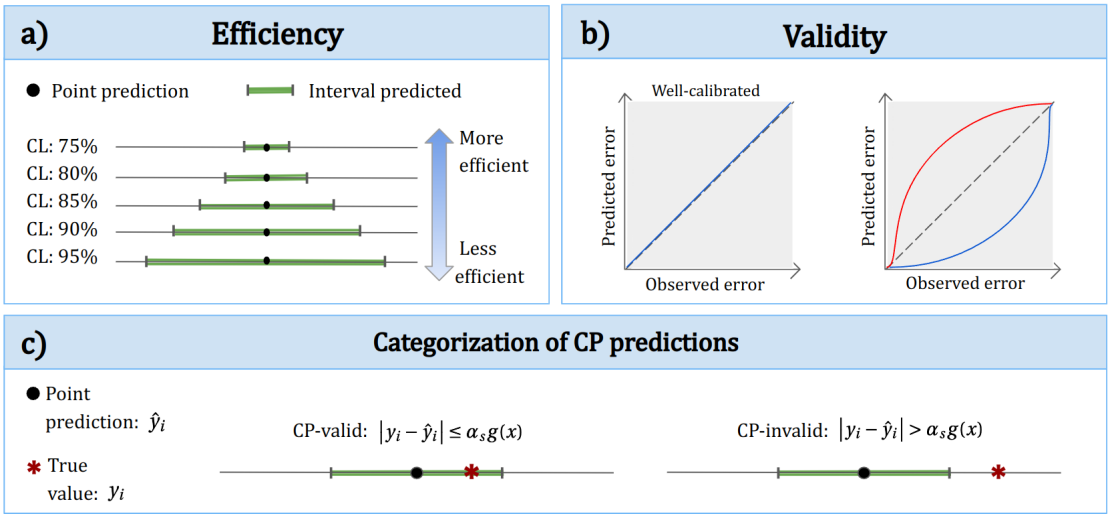


Figure 3.3.: **Graphical representation of CP evaluation.** a) Efficiency is the size of the predicted interval, where smaller intervals are more efficient. b) Validity refers to the number of errors obtained compared to the number requested. If these values coincide, it is said to be a calibrated model. Otherwise, the model can make more errors (red line) or fewer errors (blue line). c) Predictions are categorized into CP-valid if the predicted interval contains the true value and CP-invalid otherwise.

Since calculating validity involves counting how many times the model has been wrong, we categorized this type of prediction in a previous work [65]. If the predicted interval contains the true value, it is called a CP-valid prediction; otherwise, it is called a CP-invalid prediction (Figure 3.3-c). Mathematically, this can be described as follows:

$$\text{CP-valid} := |y_i - \hat{y}_i| \leq \alpha_s g(x), \text{ and CP-invalid} := |y_i - \hat{y}_i| > \alpha_s g(x) \tag{3.6}$$

where $\alpha_s g(x)$ is the prediction range.

4. Results and Discussion

Contents

4.1. Applying CP to predict the inhibitory activity of molecules 78
4.2. Baseline model investigation 80
4.2.1. Validity and efficiency hold in CP models as calibration set size is reduced 81
4.2.2. Accuracy assessment of CP-based models 82
4.2.3. CP baseline model considerations 84
4.3. Baseline model optimization 84
4.3.1. Hyperparameter tuning 84
4.3.2. Increasing MFP size leads to more accurate predictions 86
4.3.3. Removing molecules with measured $pGI_{50} = 4$ decreases the model performance 88
4.3.4. The importance of optimal data partitioning in cross-validation strategies 90
4.4. Data preparation for a real-world scenario 91
4.4.1. Leveraging molecular clustering to improve model validation . . 91
4.4.2. Relevance of outlier molecules definition 94
4.4.3. Outlier detection to create an external validation set 95
4.4.4. Understanding the impact of different clustering algorithms on molecular clustering 96
4.4.5. Analyzing molecular clustering quality with Tanimoto similarity matrix 98
4.5. Integration of the previous steps in a real-world scenario 99
4.5.1. Retrospective evaluation of CP models on dissimilar test sets . . 99
4.5.1.1. Retrospective evaluation results in the non-outlier data set 102
4.5.1.2. Outlier set results 106
4.5.2. Comparing CP-based models for Drug Design 110
4.5.3. Prospective evaluation of CP models to guide decision-making . 111

4.1. Applying CP to predict the inhibitory activity of molecules

This chapter presents the principal findings of this thesis project, which focuses on predicting the inhibitory activity of molecules on cancer cell lines of the NCI-60 panel using CP-based models. The information presented in this chapter is derived from four distinct studies conducted to address the research questions posed in Section 2.4. Three of these four studies can be found in Appendices A, B, and C.

Drug discovery, a time-consuming and expensive process, can be revolutionized by approaches like PDD [33, 67]. PDD, which identifies compounds that can produce a desired therapeutic effect on the disease phenotype, has shown great promise in recent studies as a potential accelerator for drug discovery [95].

AI and ML are promising tools in biomedicine and healthcare, particularly in early drug discovery. The abundance of preclinical data available today has made using ML-based models in drug discovery possible and highly effective [13, 14]. A prime example is the NCI-60 panel that has been used since 1990 to identify compounds with potential anticancer activity. However, ML may face challenges in the performance of new compounds and in assigning reliability to individual molecules.

This thesis project aims to create predictive models using ML techniques and CP frameworks to identify the most effective molecules for treating various types of cancer. This will help us identify candidate compounds that are likely to be potent and have a lower uncertainty level, making them good candidates for preclinical evaluation. One reason for using CP models for this type of task is their ability to assign a level of uncertainty to the predictions. Focusing on reliable predictions increases the likelihood of successful outcomes in subsequent stages of drug development, thereby enhancing decision-making confidence.

The first stage was to develop a CP baseline model to establish a benchmark for subsequent investigations (Figure 4.1-a). The preliminary investigation focused on predicting molecules' inhibitory activity, employing models (underlying and error) trained on randomly partitioned datasets. This study presents an initial evaluation of the performance of CP models in drug design. The findings, discussed in Section 4.2, demonstrate that this method has promising capabilities in identifying potent molecules. Furthermore, this preliminary investigation has identified areas where further improvements can be made to enhance the accuracy and efficiency of CP models. This initial research resulted in the first published paper of the thesis (see Appendix A).

In the second stage, we focus on optimizing the performance of the baseline model. To this end, three primary aspects were explored: hyperparameter tuning, molecular representation employing varying sizes of MFPs, and assessing different CP methodologies such as Cross CP and Aggregated CP (Figure 4.1-b). In Section 4.3, we comprehensively explain the optimization process. Our research reveals that larger MFPs

improve model performance. Moreover, we found that optimizing data partitioning is critical for advanced CP versions. This stage results in improved CP models with optimized parameters, which are expected to perform better than the baseline model established in the previous stage.

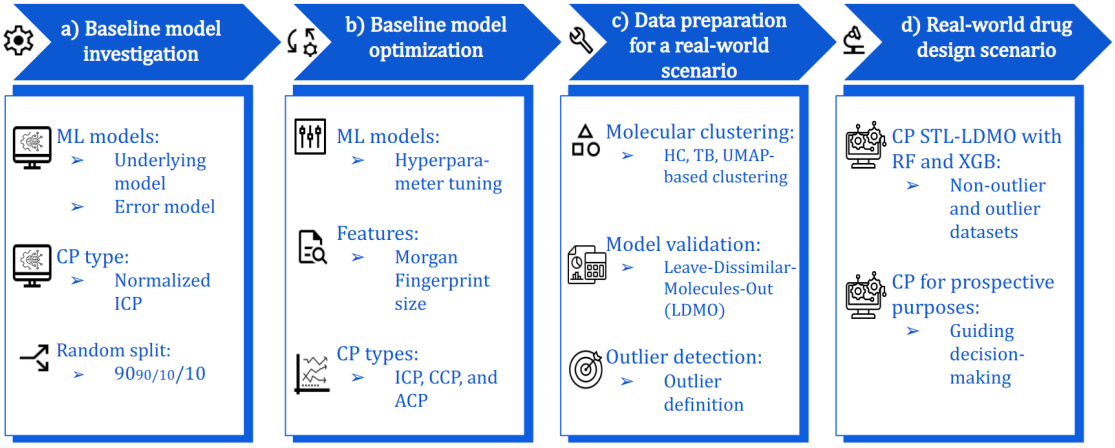


Figure 4.1.: **Applying CP to predict the inhibitory activity of molecules in cancer cell lines of the NCI-60 panel**. a) CP baseline models were developed to set a benchmark for future investigations. This preliminary study focused on predicting the inhibitory activity of molecules using normalized CP models trained on randomly partitioned datasets. b) In the second stage, we focus on optimizing the performance of the baseline model, for which three primary aspects were explored. c) In the third stage, we prepare the data for a more realistic and rigorous evaluation, moving beyond random data separation. We investigated cluster analysis for model validation, also emphasizing outlier detection. d) The final stage aims to use the optimized CP models for practical use in a real-world drug design scenario, covering both retrospective and prospective evaluations.

Optimizing a model is crucial for improving its performance, and an essential part of this process is to have a robust model validation strategy. Using cross-validations with random partitions of data is not adequate for this purpose. Instead, molecule clustering is emerging as a promising alternative for generating clusters that can be used for model validation (Figure 4.1-c). We evaluated three known clustering methods as part of this third stage in preparing the dataset for training CP models on real-world scenarios. One of these methods, which is based on the UMAP technique, demonstrated superior performance compared to the others. We have published the details of these experiments in [63], which can be found in Appendix B. Additionally, in Section 4.4, we discuss the main findings of this study.

The concluding step of this research project involves performing experiments to evaluate the optimized CP models in scenarios that simulate real-world situations (Figure 4.1-d). To achieve this, the final stage is divided into two parts. The first part is a retrospective assessment, where we evaluate the performance of CP models on a test set that contains molecules that are different from those used for model development. The results of this retrospective evaluation can be found in [64], as well as in Appendix C. The second part involves a prospective evaluation, in which we propose an optimal

strategy for using CP predictions when class labels are not available. The findings of both stages are discussed in Section 4.5.

4.2. Baseline model investigation

A key question of this thesis is to understand the accuracy of CP models in predicting pGI_{50} values, which is crucial for assessing the reliability of these predictions. Therefore, one of the first stages in this thesis project involved building and evaluating six CP models. The aim was to lay the groundwork for assessing the performance of these models in drug design applications.

Detailed findings from these experiments have been made publicly available through an open-access, peer-reviewed article, "Conformal prediction of small-molecule drug resistance in cancer cell lines", orally presented at the 11th *Symposium on Conformal and Probabilistic Prediction with Applications* (or COPA conference). This work was published in the Proceedings of Machine Learning Research (PMLR):

> Saiveth Hernández-Hernández, Sachin Vishwakarma, and Pedro Ballester. "Conformal prediction of small-molecule drug resistance in cancer cell lines". In: *Conformal and Probabilistic Prediction with Applications*. PMLR. 2022, pp. 92–108.

In this thesis project, we refer to this paper as the "COPA paper", which can be found in Appendix A. The COPA paper is an initial research study on the predictive accuracy of CP models in predicting the inhibitory activity of molecules. CP-based models can estimate uncertainty at the level of individual instances. This means that the predicted interval length represents an associated uncertainty level for each pGI_{50} prediction made for a molecule. The error model controls the predicted interval size in the normalized CP (see Section 1.5.3 for reference). A CP model is represented by the duple $h(x) - g(x)$, where $h(x)$ is the underlying model and $g(x)$ is the error model.

For this initial work, the molecules in the NCI-60 panel were represented using a 256-bit MFP size with a radius of 2 (Figure 4.2-a). The data preprocessing stage can be consulted in Section 3.2. We adopt a random partitioning strategy to divide the dataset into training and test sets. After the initial partition, the training set was randomly split into a proper training set and a calibration set (Figure 4.2-b). We tested three different 'proper training-calibration' configurations to measure the impact of decreasing the number of molecules in the calibration set.

Six CP models were built and evaluated for each cell line to predict the pGI_{50} value of a given test set molecule from its molecular features. Table 1A shows the combinations of the considered CP models, each evaluated at four confidence levels: 80%, 85%, 90%, and 95%. Thus, a total of 1080 models (60 cells × 6 CP modes × 3 data partitions) were built and evaluated in this initial study. Then, we computed the RMSE and Rp metrics for each 60 sets, one per cell line. This evaluation was complemented using

CP's validity and efficiency metrics (Figure 4.2-c).

In the final step of our analysis, we implement a threshold that helps us distinguish between the active and inactive molecules. This process allows us to calculate the hit rate of our predictions accurately. The main findings of this work are discussed in the following subsections.

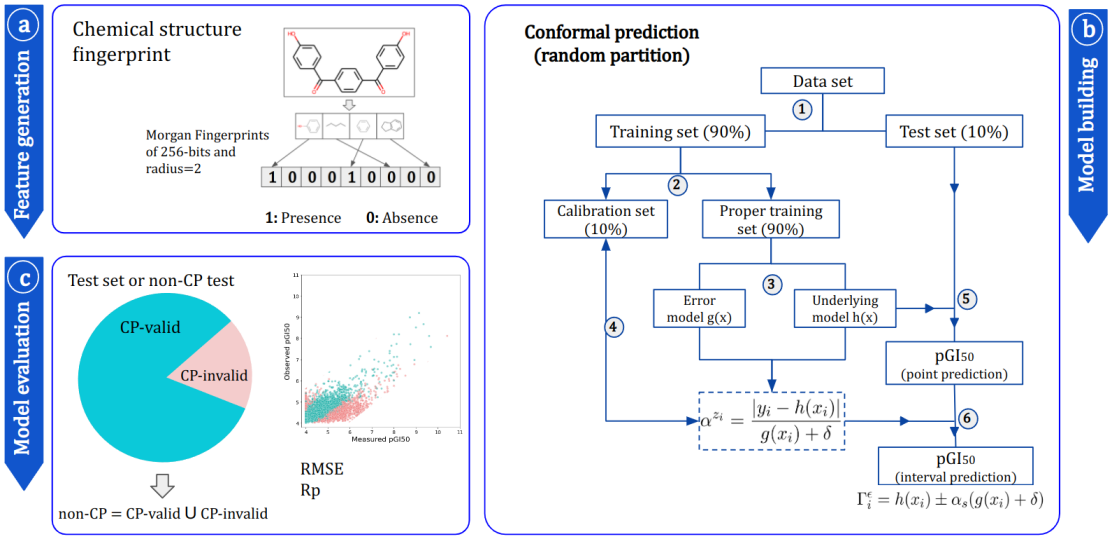


Figure 4.2.: **Initial research on the predictive accuracy of CP models for predicting inhibitory activity of molecules**. a) Molecules from the NCI-60 panel are represented using 256-bit MFPs, then used as input features for CP models. b) The dataset is randomly divided into training and test sets to build CP models. The training set is then subdivided into a proper training set and a calibration set. This step is repeated for each of the 60 data sets, one per cell line. A total of 1080 models, 60 cells × 6 CP modes × 3 data partitions, were built and evaluated in this initial study. c) The prediction performance is evaluated using metrics such as RMSE and Rp. We propose the terms "CP-valid" and "CP-invalid" predictions to complement the evaluation of the CP models.

4.2.1. Validity and efficiency hold in CP models as calibration set size is reduced

The calibration set size in CP is crucial for calibrating prediction intervals and estimating uncertainties in model predictions. However, the calibration and training set sizes' impact on CP-based models has been relatively unexplored [48], especially compared to the extensive research on the training set and test set sizes for ML-based models.

An experiment conducted in this initial study focused on setting the test set size and examining various training and calibration set sizes. We explored three different sizes for the training and calibration sets to discern any variations in model performance. The combinations evaluated are 90-10%, 80-20%, and 70-30%. We assessed the impact of this variation in CP-based models by considering validity (expected error) and efficiency (uncertainty).

Our results reveal that modifying the proportion of training and calibration set sizes does not significantly affect the validity or the efficiency evaluation metrics (Figure 3A). This suggests that, within the scope of the NCI-60 dataset, the diversity and richness of information remain robust even when the number of molecules in the calibration set is reduced. Moreover, this observed consistency across different dataset sizes underscores the reliability of the CP-based model in providing accurate predictions and uncertainty estimates. Therefore, we keep the 90-10% data partition for all the experiments in the COPA paper.

Since decreasing the number of calibration molecules for CP models shows no impact on their validity or efficiency, it implies that resources can be more efficiently allocated during model training by reducing the calibration set size. This adjustment allows more data to be assigned to training, enhancing the overall model's effectiveness.

Although our study has demonstrated the stability of our results in terms of validity and efficiency, it is important to note that the scope of our dataset limits these findings and may not be generalizable to other datasets. Indeed, Eklund et al. [48] analyzed the impact of different calibration set sizes (in-house datasets) and found that inadequate calibration set sizes resulted in a higher variance of confidence.

4.2.2. Accuracy assessment of CP-based models

Experiments in the early stages of our research focused on assessing the accuracy of CP models in predicting the inhibitory activity of molecules on cancer cell lines. We compared the outcomes of these experiments with those derived from standard ML models to determine whether CP models improve on standard ML techniques. Our analysis used the RMSE and Rp metrics to ensure a balanced comparison.

To complement this comparison, we introduce a novel approach by defining "CP-valid" predictions as the outcomes where, given a confidence level, the true label falls within the predicted interval (Figure 3.3-c). Our definition of CP-valid and CP-invalid predictions contributes to the existing literature, where we observed very few studies (Table 2.1) focusing on assessing only correctly predicted molecules by CP models. Indeed, CP models are traditionally evaluated based on validity (expected error) and efficiency (interval size) metrics calculated for the entire test set. In Norinder et al. [103], the authors define an erroneous prediction as one where the difference between experimental and predicted activity is larger than the prediction range. Their definition is similar to the one we provide for CP-invalid predictions (Equation 3.6), but they do not give a formal definition or capitalize on the potential for correct predictions.

To the best of our knowledge, our work represents the first instance in the literature of defining and leveraging CP-valid predictions, filling a gap in existing evaluation methodologies. Unlike previous studies, which predominantly rely on traditional evaluation metrics, our approach leverages the predictive power of CP-valid predic-

tions to prioritize molecules with a high probability of success in future stages of drug development. While our approach enhances the reliability of prediction outcomes, it may inadvertently exclude certain molecules that fall outside the defined CP-valid criteria, potentially overlooking valuable candidates deserving of further investigation. Therefore, continuously refining and optimizing the CP-valid definition may be essential to balance reliability and inclusivity during the selection process. This could ensure that valuable molecules are not missed during the screening process.

By comparing the performance of CP models with standard ML models, we observed a consistent trend across all analyzed experiments: the CP models consistently outperform the standard ML models in terms of RMSE and Rp (Figure 5A). We also found that relaxing the requested confidence level leads to more accurate predictions (Figures 5A and 6A). This finding highlights the trade-off between confidence level and prediction accuracy and demonstrates the adaptability of CP to different confidence levels. This flexibility is particularly valuable in scenarios where strict adherence to a high confidence level may lead to overly conservative predictions. By allowing for greater uncertainty in predictions, CP models can provide intervals that better capture the true label, resulting in higher accuracy for CP-valid predictions. However, it is essential to balance confidence level and prediction accuracy. While relaxing the confidence level may improve accuracy for CP-valid predictions, extensive intervals may limit the practical utility of CP models by reducing the precision of predictions.

Regarding the size of the intervals (uncertainty), our results show that even though all 6 CP models are either valid or near-valid models (the median validity is close to the required error in all CP models with $\epsilon \pm 0.002$), the choice of the kNN algorithm as the error model leads to intervals that do not have practical utility as they are very long intervals (Figure 4A). For instance, a median efficiency of around 9 pGI_{50} units is obtained by the RF-kNN. For the remaining CP models, better efficiency is obtained when RF is used as the underlying model, with median efficiency ranging from 1.43 to 2.74 pGI_{50} units in RF-RF and RF-XGB CP models. When the XGB model is used as the underlying model, the median efficiency ranges from 1.64 to 5.92 pGI_{50} units in XGB-RF and XGB-XGB CP models. Another significant finding is that CP-valid predictions demonstrate a better prediction of the pGI_{50} error (Figure 6A). These findings suggest that CP-valid predictions not only accurately capture the true range of inhibitory activity but also provide more reliable estimates of the prediction error, thereby enhancing the overall reliability of predictions. This finding underscores the value of CP in drug discovery, where accurate estimation of prediction errors is crucial for decision-making processes.

This study also established a criterion for assessing cell line performance based on efficiency metrics derived from CP models. The best-performing cell line is defined as the one exhibiting the lowest uncertainty, while the highest uncertainty characterizes the worst-performing cell line. The best-performing cell lines identified are the OVCAR-5 for the RF and the SNB-19 for the XGB models. Conversely, according to the RF and XGB models, SR emerged as the worst-performing cell line with the highest

uncertainty (see Table 2A). Moreover, the hit rates for CP-valid predictions for these cell lines were between 75.8% and 85% (see Figure 8A).

4.2.3. CP baseline model considerations

The COPA paper presents a promising approach for using CP-based models to predict molecules that may have inhibitory effects in cell lines. This highlights CP's potential as a robust tool for early-stage drug discovery and therapeutic development in cancer treatment.

While the results show promising performance in predicting potent molecules, it is essential to acknowledge that the results presented here were achieved with relatively minimal model tuning and using a random partitioning of the NCI-60 dataset, which may lead to model overestimation. To address potential biases introduced by these factors, the subsequent phases of this thesis focused on a comprehensive optimization of four promising CP models identified in this preliminary analysis.

Therefore, in the following sections, we explore various advanced techniques and methodologies to enhance the reliability and generalizability of CP-based models. This includes investigating data representations and feature engineering approaches to capture more nuanced patterns and relationships within the data. In addition, we explore alternative data partitioning techniques to assess model performance better.

4.3. Baseline model optimization

AI-based models, such as ML and CP, are powerful tools for solving complex problems and making predictions based on data. However, it is crucial to understand that their performance can be highly sensitive to various factors, such as input feature representation, data partitioning strategy, and other preprocessing steps. In this section, we aim to optimize these key elements to determine if there is any improvement, mainly when working with CP models.

Except for the hyperparameter tuning, these experiments were carried out on the SR cell line, a challenging dataset identified in our previous work [65]. This cell line showed a suboptimal performance in both CP models (see Table 2A). This approach is based on the idea that if a model shows substantial improvement on a challenging dataset with suboptimal performance, the same adjustments are likely to lead to even more significant improvements on datasets where the model already performs well.

4.3.1. Hyperparameter tuning

Each of the 60 training sets undergoes a hyperparameter tuning process using a grid search to find the best values for predicting the inhibitory activity of molecules on

that cancer cell line. The most important hyperparameters to tune for the RF and XGB models were described in our previous work [65] and listed below for reference.

- RF:
  - $n_estimator$ refers to the number of trees in the forest. Tested values are 250, 500, 750, and 1000.
  - $max_features$ refers to the number of features to consider when looking for the best split. Tested values are 0.2, 0.3, 0.4, 0.6, 0.8, and 0.9.
- XGB
  - $n_estimator$ refers to the number of gradients boosted trees. Equivalent to the number of boosting rounds. Tested values are 250, 500, 750, and 1000.
  - max_depth refers to the maximum tree depth for base learners. Tested values are 5, 6, 7, 8, 9, 10.
  - $cosample_bytree$ refers to the subsample ratio of columns when constructing each tree. Tested values are 0.3, 0.4, 0.6, 0.8, and 0.9.

We employ an 80/10/10 scheme to tune these hyperparameters. This means that 80% of the data was used for training, 10% for validation, and 10% for testing. After finding the best hyperparameter values, we re-trained the model using 90% of the data. As a result, the test set was not used in any way to train or select the final model. The hyperparameter values used for the 60 cell lines were selected based on the value at which most cell lines achieved their best performance.

Our results for the RF model show that increasing the number of trees in the forest ($n_estimator$) leads to better RMSE performance. On the other hand, increasing the number of features ($max_features$) when looking for the best splitting produces the opposite result. For values higher than 0.4, we can observe a constant growth in the RMSE, thus decreasing the performance of the RF model.

In the XGB model, increasing the number of gradient-boosted trees ($n_estimator$) and the subsample ratio of columns when constructing each tree ($cosample_bytree$) leads to a better RMSE result. In the case of increasing the maximum tree depth for base learners (max_depth), after a depth of 9, the RMSE remains constant.

Figure 4.3 shows the grid search results, where the combination of values leading to the best performance of the models is marked with an arrow. These values are n_estimator=1000 and max_features=0.4 for RF; and n_estimator=1000, max_depth=9 and cosample_bytree=0.4 for XGB. This combination of parameters is used for the following experiments.

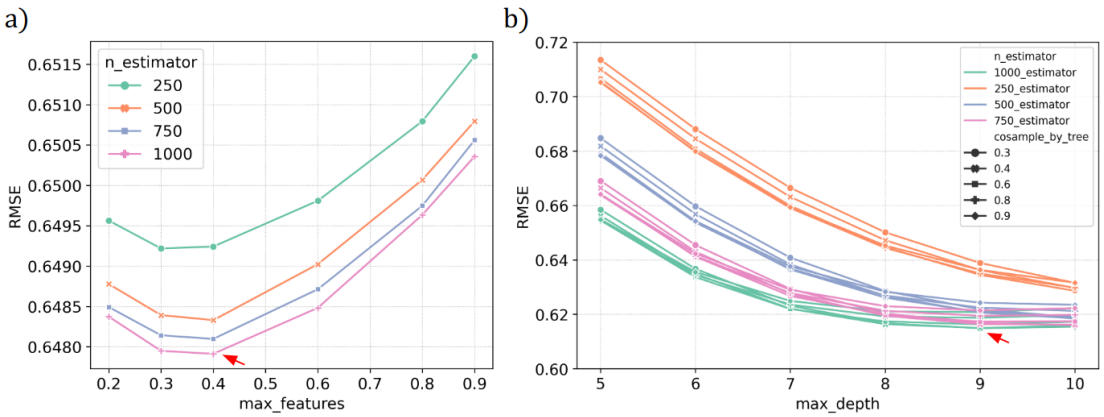


Figure 4.3.: **Grid search results** for a) RF and b) XGB. For both models, a color code represents the "n_estimator" hyperparameter. For the XGB, the markers represent the "cosample_bytree" value evaluated. The red arrow indicates the specific hyperparameter combination, resulting in the optimal model performance regarding the RMSE.

4.3.2. Increasing MFP size leads to more accurate predictions

The impact of feature representation on the performance of ML models in chemoinformatics has been a subject of significant interest [10, 22]. In this experiment, we investigated the impact of varying MFP sizes on the performance of predictive models, both ML and CP, by examining sizes of 256, 1024, and 2048 bits. These sizes represent varying levels of granularity in capturing molecular features, with larger sizes allowing for more detailed representations of molecular structures. In this context, the model referred to as 256 MFP, introduced in Section 4.2, serves as the benchmark for comparing performance across different scenarios.

For both non-CP and CP-valid predictions, results show that increasing the size of the MFP leads to more accurate predictions across the evaluation metrics, including RMSE, Rp, and Rs. Indeed, for non-CP predictions, transitioning from a 256 MFP to a 1024 MFP decreases RMSE from 0.7515 to 0.7203, indicating an improvement of approximately 0.03 pGI_{50} units. Similarly, the improvement continues when transitioning from 256 MFP to 2048 MFP, with RMSE decreasing further to 0.7166, reflecting an improvement of approximately 0.035 pGI_{50} units. This pattern is consistently observed across Rp and Rs, where the improvements are between 0.019 and 0.024 pGI_{50} units. For CP-valid predictions obtained at an 80% confidence level, moving from 256 MFP to 1024 MFP leads to a decrease in RMSE from 0.5130 to 0.4752 and further to 0.4683 for 2048 MFP. The results for the RF model are displayed in Figure 4.4.

However, it is important to note that increasing the MFP size also comes with computational costs, as larger fingerprints require more memory and computational resources for storage and processing. Table 4.1 presents the time needed for generating MFPs and building RF and XGB models. The computation time for generating MFPs shows minimal variance across different sizes, with an approximate increase of 20 seconds

observed as the MFP size escalates. In contrast, the computation time to build the models significantly increases with increasing MFP size, with approximately a 5x increase for the RF model and a 4x increase for the XGB model.

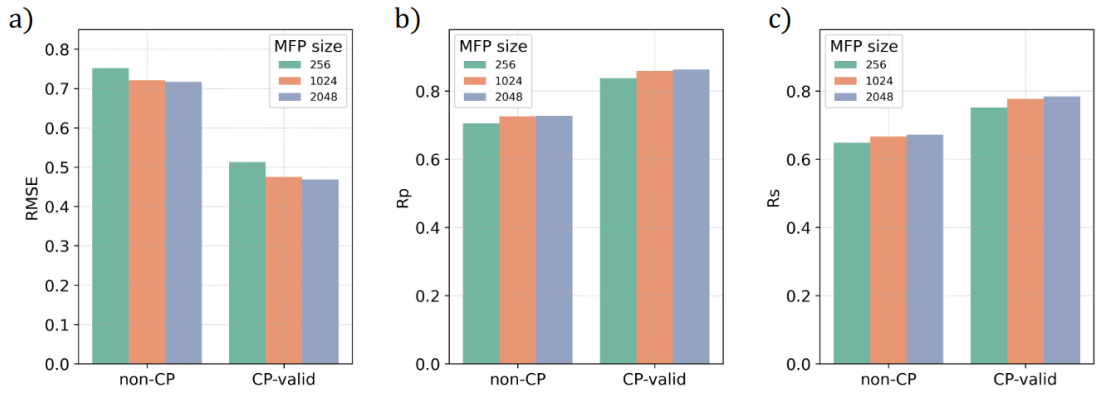


Figure 4.4.: **Increasing the size of the Morgan fingerprints (MFP) leads to more accurate predictions across multiple evaluation metrics**. a) RMSE, b) Rp, and c) Rs values were computed between the observed and predicted pGI_{50} value, using the RF model as the underlying model. CP-valid predictions were computed at an 80% confidence level. Color code refers to the MFP size used.

Table 4.1.: The time to build the model grows exponentially as we increase the size of the MFPs. The table contains the computational times, in minutes, needed for generating MFP (three values: 256, 1024, and 2048 bits) and building models (two models: RF and XGB). MFPs were calculated for 50,555 unique molecules, while model building only included the approximately 41K molecules screened in the SR cell (Figure 3.2-c).

Model	MFP size	Time in minutes to generate MFPs	Time in minutes to build models
RF	256	1.83	0.65
RF	1024	2.01	5.82
RF	2048	2.23	24.42
XGB	256	1.83	0.63
XGB	1024	1.99	3.10
XGB	2048	2.21	13.39

Although larger MFP sizes imply considerable increases in model accuracy, it is essential to note that the performance, as assessed by RMSE, only shows a modest improvement when transitioning from 1024 MFP to 2048 MFP. Indeed, the RMSE decreases by approximately 0.004 pGI_{50} units for non-CP predictions and 0.007 pGI_{50} units for CP-valid predictions (Figure 4.4). This slight improvement in predictive performance does not justify the additional computational cost. Therefore, considering the trade-off between predictive accuracy and computational cost, the 1024 MFP size emerges as the optimal choice. Consequently, this size is selected for subsequent experiments.

4.3.3. Removing molecules with measured $pGI_{50} = 4$ decreases the model performance

An important aspect observed during data exploration was a peak at $pGI_{50} = 4$ (Figure 3.2-b). This peak originates because, according to the NCI-60 screening methodology [100], 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. Therefore, these measurements are less accurate.

Here, we explore different data handling strategies to assess their impact on model performance and generalization capabilities. The three scenarios presented for compounds with measured $pGI_{50} = 4$ are:

- Non-removed. This scenario serves as the benchmark for comparing performance across different scenarios.
- Removed only from the training set. In this scenario, compounds with measured $pGI_{50} = 4$ are removed from the training set but retained in the test set. This allows us to assess how well the model can generalize when evaluated on a slightly different dataset than the one it was built on.
- Removed from both training and test sets. In this scenario, compounds with measured $pGI_{50} = 4$ are removed from both sets, so the model is unaware of these molecules. This scenario allows us to evaluate whether this peak represents noisy data.

Figure 4.5 shows the RMSE, Rp, and Rs results for the non-CP and CP-valid predictions when RF is used as the underlying model. In the baseline scenario, where no molecules are removed, the non-CP RMSE stands at 0.7203. However, upon removing molecules with measured $pGI_{50} = 4$ solely from the training set, we observe a notable increase in RMSE to 0.8559. This suggests that the presence of these molecules in the training data contributes to better generalization of the models to the test set, as their exclusion leads to a degradation in predictive performance, with Rp and Rs decreasing to 0.6854 and 0.5257, respectively. When molecules are removed from both the training and test sets, the non-CP RMSE still exceeds the baseline RMSE. However, the increase is slightly smaller ($\sim$ 0.013 units), reaching an RMSE of 0.7335. This indicates that while the absence of these molecules from both datasets still impacts model performance adversely, the effect is less pronounced than solely removing them from the training set. Therefore, some information from these molecules is still beneficial during the testing phase.

A similar trend occurs for the CP models, where removing molecules with measured $pGI_{50} = 4$ leads to a decrease in model performance, whether from the training set alone or from both the training and test sets. For CP-valid predictions, we also measure the impact of these scenarios in terms of expected error (validity) and uncertainty (efficiency). These results are summarized in Table 4.2. In this case, the expected error is set at 0.2, corresponding to an 80% confidence level. In the baseline scenario where

no molecules are removed, the error rates for both models are lower than the expected error. For the second scenario, when molecules with measured GI50=4 are removed only from the training set, we observe a notable increase in error for both models, with the RF model exhibiting a much larger error increase compared to the XGB model. When molecules with measured $pGI_{50} = 4$ are removed from both the training and test sets, the error remains slightly higher than the baseline error but falls below the expected error.

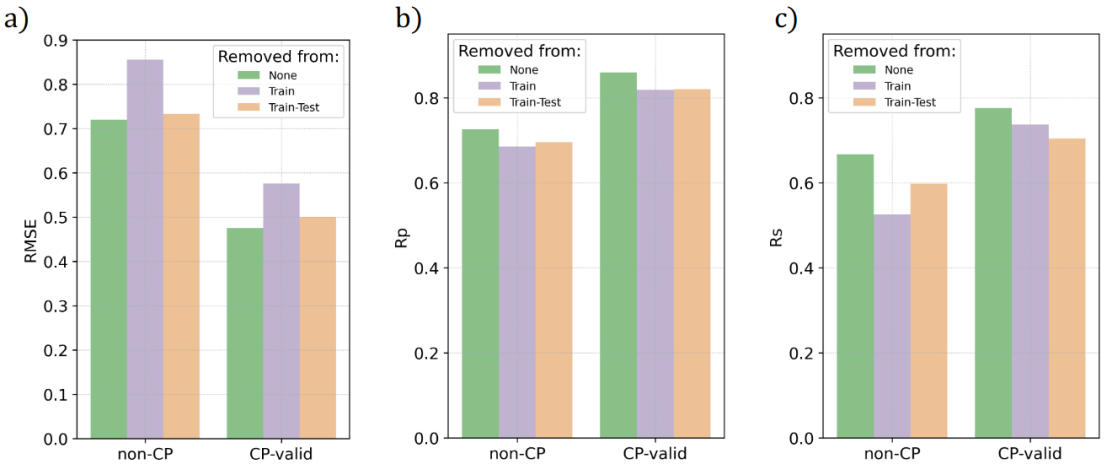


Figure 4.5.: **Removing molecules with measured $pGI_{50} = 4$ decreases the model performance, across multiple evaluation metrics.** a) RMSE, b) Rp, and c) Rs values were computed between the observed and predicted pGI_{50} value, using the RF model as the underlying model. CP-valid predictions were computed at an 80% confidence level. Color code refers to the set from which the compounds with measured $pGI_{50} = 4$ are eliminated. None: Non-removed, Train: Removed only from the training set, and Train-Test: Removed from both training and test sets.

Table 4.2.: Models trained on datasets containing molecules with measured $pGI_{50} = 4$ are more confident in their predictions. The expected error is set at 0.2, corresponding to an 80% confidence level. The results are shown for both models, RF and XGB, considering the three scenarios analyzed (None: Non-removed, Train: Removed only from the training set, and Train-Test: Removed from both training and test sets).

Model	Removed from	Median error	Median uncertainty	Number of test molecules
RF	None	0.1784	1.6784	4064
RF	Train	0.3627	1.8036	4064
RF	Train-Test	0.1965	1.8123	2722
XGB	None	0.1804	1.7589	4064
XGB	Train	0.3115	1.9000	4064
XGB	Train-Test	0.2002	1.8915	2722

Predictions in the baseline scenario exhibit lower uncertainty, with an average uncertainty of 1.68 units for the RF model and 1.76 units for the XGB model. The findings suggest that the absence of molecules with measured $pGI_{50} = 4$ from the training set

significantly impacts model performance and increases uncertainty. Although removing these molecules from the training and test sets mitigates some of these effects, resulting in errors below the expected error level, these errors and uncertainties are not as favorable as those obtained in the baseline scenario. As a result, these types of molecules will be retained for the subsequent experiments.

4.3.4. The importance of optimal data partitioning in cross-validation strategies

Cross-validation is a widely accepted approach for developing AI models that can generalize to new data [71]. By partitioning the dataset into multiple subsets and iteratively training and evaluating the model on different partitions, CV aims to provide more reliable estimates of model performance compared to a single train-test split. However, the data partitioning strategy is crucial for the effectiveness of CV approaches. If the data partitioning strategy is suboptimal, the benefits of CV may be compromised, and the model's ability to generalize to new data may not be improved.

To highlight this issue, we analyzed the impact of using random data partitions in more complex versions of CP, specifically ACP and CCP (see Section 1.5.3). These versions, which combine multiple CPs, allow us to evaluate whether there is an improvement over the baseline CP, which involves only one run. Based on previous analysis, the results reported here use all molecules in the training and test set (i.e., molecules with $pGI_{50} = 4$ were not removed) and use a 1024 MFP size. We kept the default number of ICP models (10 models) for the ACP model, and for the CCP model, we explored 5 and 10 folds (each fold represents an ICP model). We measure whether there is an improvement in RMSE, Rp, and Rs for both non-CP and CP-valid predictions.

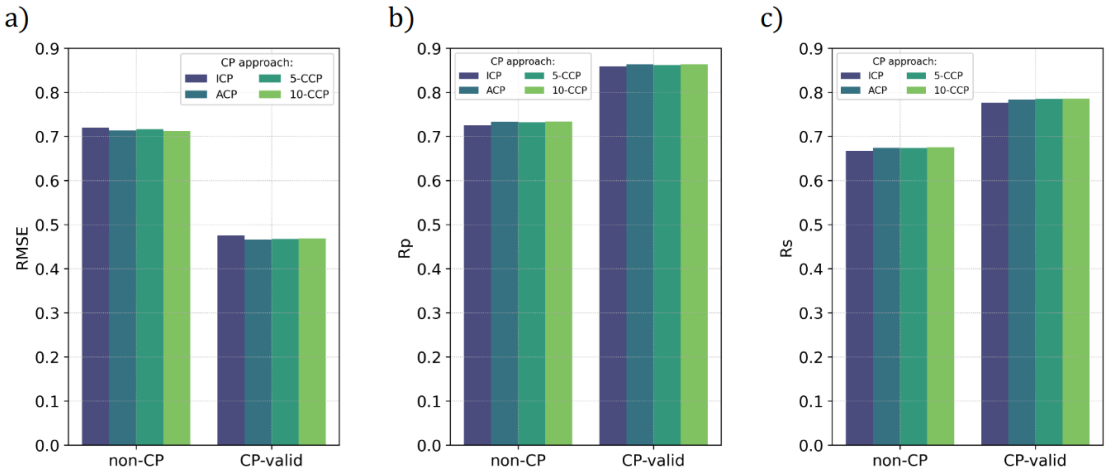


Figure 4.6.: **Data partition strategy is crucial as a CV does not significantly improve over a single-run approach.** a) RMSE, b) Rp, and c) Rs values were computed for the non-CP (left) or CP-valid (right) predictions. CP-valid predictions are obtained at a confidence level of 80%. The color code represents the type of model used, either ICP, ACP, 5-CCP, or 10-CCP. The results are from the RF model.

Our research findings indicate that using a data random partitioning approach does not enhance the performance of the models in any of the three variants, ACP, 5-CCP, or 10-CCP, compared to the standard CP approach (Figure 4.6). Additionally, the long computational times observed in this case suggest that the standard CP model is the most suitable option for achieving optimal results (Figure 4.7). Since the model's performance does not improve significantly over the single-run approach, this may indicate that problems with the partitioning strategy must be addressed.

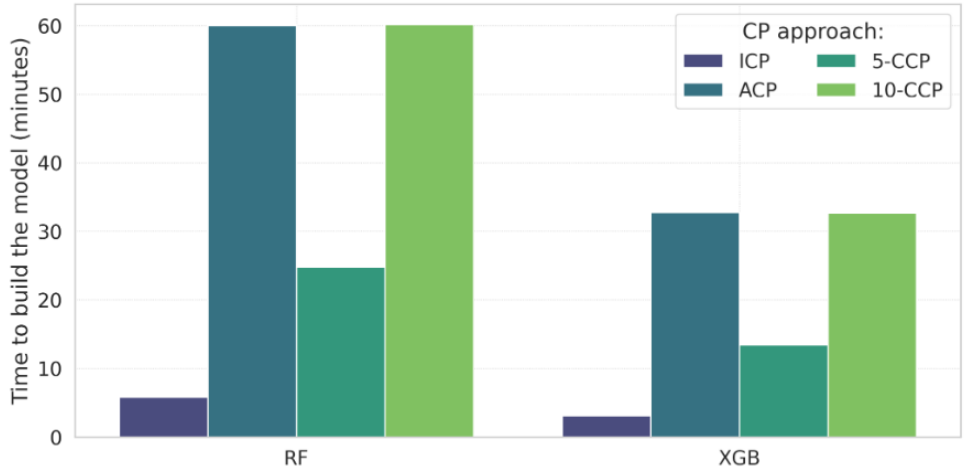


Figure 4.7.: **The ACP and CCP approaches require more computational time to build the models**. The color code represents the type of model used, either ICP, ACP, 5-CCP, or 10-CCP.

When optimally executed, CV can provide valuable information about the robustness and reliability of CP-based models on various datasets. As such, it is essential to optimize the data partitioning process for CV to ensure that the results are accurate and robust across different datasets. The following section will explore the strategies and best practices for optimizing data partitioning in cross-validation.

4.4. Data preparation for a real-world scenario

4.4.1. Leveraging molecular clustering to improve model validation

In this thesis project, a question arises about data partitioning and its impact on pGI_{50} predictions. As discussed in the previous section, cross-validation with random partitioning does not significantly enhance results, underscoring the importance of effective data partitioning. By recognizing this limitation, we investigated molecular clustering as an alternative approach to obtain clusters for model validation strategies. Hence, the subsequent stage of this project involved investigating the most effective

approach to cluster the molecules screened by the NCI-60 panel. The best-performing algorithm can then be used to generate clusters for model validation strategies.

Our research findings have been made publicly available through a peer-reviewed, open-access article, "On the Best Way to Cluster NCI-60 Molecules", published in the Biomolecules Journal:

> Saiveth Hernández-Hernández and Pedro J Ballester. "On the Best Way to Cluster NCI-60 Molecules". In: *Biomolecules* 13.3 (2023), p. 498, https://doi.org/10.3390/biom13030498

In this thesis project, we refer to this paper as the "Clustering paper", which can be found in Appendix B.

Clustering is an unsupervised ML-based algorithm usually employed to discover patterns in datasets. It can be used to cluster biological entities, like small molecules, using different approaches (e.g., hierarchical, distribution-based, and density-based) [11, 74, 115]. Clustering chemical compounds is challenging due to different representations, leading to different clustering outcomes. Indeed, if two compounds have different molecular structures, they can still have similar molecular descriptors or fingerprints, which could result in them being clustered together. Conversely, two compounds with similar molecular structures may have varying molecular descriptors, causing them to be clustered separately. MFPs are an advantageous molecular representation for clustering tasks due to their computational efficiency and robustness to small variations in molecular structure. They are also easily applicable to large datasets and useful in virtual screening. Therefore, based on our previous results, NCI-60 molecules are represented using 1024-bit MFPS (Figure 4.8-a).

Outliers in a data set can significantly impact the performance of clustering algorithms and other AI-based models [63]. As a result, it is common practice to eliminate such instances during data preprocessing [7, 134]. However, traditional statistical definitions that define outlier examples may not apply to chemical compounds. This clustering study introduces a novel definition of outlier molecules based on Tanimoto similarity (Figure 4.8-b) to address this challenge. This definition enables us to leverage the unique characteristics of chemical compounds. Section 3.1B provides the specific results of this approach, which are further discussed in detail in Subsection 4.4.2.

The proposed definition for detecting outlier molecules serves a purpose beyond simply improving clustering results by cleaning the data. Specifically, we leverage this outlier set to establish an external dataset for evaluating our CP models. This approach guarantees that the test set includes molecules that differ significantly from those used to develop the model. A detailed discussion of this approach is presented in Subsection 4.4.3.

Poor compound clustering can negatively impact validation, mainly when testing molecules differ from those in the training set. To address this challenge, we compare the efficacy of three different clustering methods: hierarchical, Taylor-Butina, and

uniform manifold approximation and projection (UMAP)-based (Figure 4.8-c). Section 2.2B explains each clustering algorithm in detail.

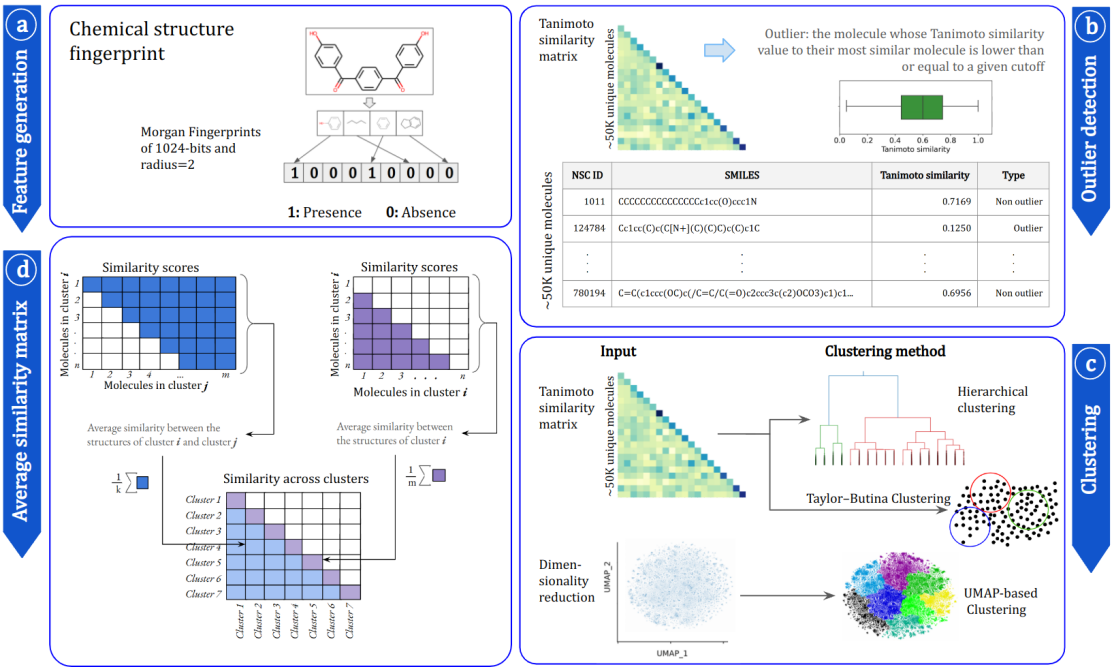


Figure 4.8.: **Leveraging molecular clustering to improve model validation.** a) We utilize MFPs as input for clustering molecules within the NCI-60 panel. b) We propose a novel method to identify outlier molecules and examine the impact of their removal on clustering quality. c) We compare the effectiveness of hierarchical, Taylor-Butina, and UMAP-based clustering techniques to determine the most suitable approach. d) The clustering results are evaluated using established metrics, and the quality of each cluster is assessed using an average similarity matrix.

Our goal was to identify the most effective approach to cluster molecules screened by the NCI-60 panel. The best-performing algorithm can then be used to generate clusters for model validation strategies. The detailed results of these experiments are given in Sections 3.2B to 3.5B and are discussed in Subsection 4.4.4.

The clustering evaluation uses three well-known clustering quality metrics: the Silhouette coefficient, the Calinski–Harabasz score, and the Davies–Bouldin score. Section 2.3B provides a detailed description of each metric. However, as similar compounds may have different descriptors, causing them to cluster separately and vice versa, these metrics may overlook meaningful structural relationships between molecules across clusters. We introduced a new metric that considers molecular structural similarity to address this limitation. This metric comprehensively evaluates intra-cluster cohesion and inter-cluster separation by calculating the average Tanimoto similarity matrix between clusters (Figure 4.8-d). Subsection 4.4.5 offers an in-depth discussion of the contributions of this approach.

Both evaluations showed that while hierarchical and Taylor-Butina methods perform poorly in clustering molecules and are computationally expensive to use in CV strate-

gies, the UMAP-based clustering yields superior quality and is recommended for analyzing the NCI-60 dataset.

4.4.2. Relevance of outlier molecules definition

The first important contribution of our research is the novel method we propose for identifying outlier molecules. The proposed definition of an outlier molecule presented in the clustering paper defined an outlier "as the molecule whose Tanimoto similarity value to their 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" [63].

The proposed definition of outlier molecules represents a novel approach that deviates from the standard definition of outliers. Traditionally, outliers are identified based on statistical measures such as being distant from the mean or median of a dataset [134]. However, when working with molecules, a more tailored definition is required to capture their unique characteristics and interactions. This limitation of traditional methods is evident in Figure 3B-a), where if we followed the standard definition of outliers, there are none in the dataset. Our approach fills this gap, providing a more nuanced understanding of outlier molecules.

The uniqueness of our approach lies in its focus on molecular similarity rather than statistical deviation. By defining an outlier molecule as one whose Tanimoto similarity value to its most similar molecule falls below or equals a predefined cutoff, we emphasize the molecular uniqueness and dissimilarity of these outliers compared to the rest of the dataset. This approach acknowledges the complex and diverse nature of molecular structures. One advantage of this new definition is its direct relevance to the field of drug design and molecular property prediction. This approach identifies outliers based on molecular similarity, making it possible to detect outliers that other, more generic outlier detection methods may have overlooked. Additionally, this approach provides a transparent and interpretable criterion for identifying outliers by using a predefined cutoff for molecular similarity. Researchers can adjust the cutoff based on the specific requirements of their study and the characteristics of the dataset under investigation.

However, it is essential to acknowledge some limitations of this new outlier definition. Indeed, calculating Tanimoto similarity values for all pairs of molecules in a dataset can be computationally expensive, particularly for large datasets [78]. In our case, the calculation of the Tanimoto similarity for each pair of molecules resulted in a total of 1,277,878,735 pairs, which could cause computational cost problems in the future. To prevent this, we only kept the similarity value of each molecule to its closest molecule, excluding itself. We retained these molecules because all the other molecules were even less similar. This way, we only have 50,555 similarity values, which reduces the computational cost for the following steps.

Another possible limitation of our approach is that determining the outlier cutoff value introduces a degree of subjectivity and requires careful consideration. The choice of a cutoff may influence the number and identity of molecules identified as outliers, potentially impacting downstream analyses and interpretations. We address this limitation by calculating the clustering result at different cutoffs listed in Table 1B. We have chosen a clustering method that serves as a basis for determining the best cutoff value for our data. Moreover, we have evaluated the quality of the selected cutoff value by calculating three well-known clustering metrics: the Silhouette coefficient, the Calinski–Harabasz score, and the Davies–Bouldin score. This enables us to determine the impact of different cutoff values and ensures a systematic evaluation of the clustering process. As a result, we found that increasing the outlier cutoff leads to fewer clusters. We also observed that the outlier cutoff of 0.5 best balances the number of clusters and their quality. This is evident from the fact that both the Silhouette coefficient and the Davies-Bouldin score show better values at this particular cutoff. Therefore, an outlier cutoff of 0.5 was utilized for the subsequent experiments.

4.4.3. Outlier detection to create an external validation set

In our work, detecting outlier molecules serves a dual purpose beyond traditional data cleaning methods. While outlier detection commonly aims to identify and remove anomalous data points that may skew analysis results [134], our approach goes further by leveraging these outlier molecules to create an external validation set. Therefore, the external validation set is entirely distinct from the inlier dataset used to train the CP-based model, providing a more rigorous assessment of its predictive capabilities in real-world scenarios.

A widely used method for obtaining external test sets is the scaffold split method [162]. Scaffold splitting has been proposed as a strategy offering a more challenging evaluation of model generalizability [158]. However, this approach may not accurately reflect the actual chemical diversity of the compound library. In addition, molecules with different chemical scaffolds are often similar. Thus, the scaffold split method may not explicitly address the issue of ensuring structural dissimilarity between the external test set and the training data.

In contrast, our approach of leveraging outlier detection to create an external validation set provides a more rigorous and systematic way of ensuring diversity and distinctiveness in the test set. By selecting outlier molecules, our approach guarantees that the external validation set represents a novel chemical space fundamentally different from the training data, replicating real-world scenarios. This enhances the reliability and validity of model evaluation, as it ensures that the model is tested on truly unseen data. Additionally, by explicitly quantifying the dissimilarity between the training and validation sets, our approach offers greater transparency and rigor in model validation compared to other methods.

4.4.4. Understanding the impact of different clustering algorithms on molecular clustering

It is important to note that different clustering algorithms have unique strengths and weaknesses, which depend on the nature of the dataset and its underlying structure. Therefore, this work's next contribution lies in the comprehensive comparison and evaluation of three distinct clustering algorithms, hierarchical, Taylor-Butina, and UMAP-based clustering, by testing them under various conditions.

Hierarchical clustering involves building a binary tree by merging individual molecules stored at the leaves until reaching the root node, which contains all the molecules in the dataset [101]. The choice of which molecules to merge is determined by the Ward linkage criteria. In this study, we examined different Ward linkage cutoffs ranging from 0.5 to 3.0, with a step size of 0.5, which were applied to either all or non-outlier molecules. Our findings indicate that decreasing the Ward linkage cutoff leads to a larger number of clusters, as demonstrated in Table 2B. Indeed, when we move from the cutoff level of 1.5 to 0.5, the number of clusters increases from 2,565 to 33,246. Therefore, using these cutoff levels to generate clusters for model validation strategies is not recommended. Even for the best cutoff, a Ward linkage = 3, the average Tanimoto matrix suggests poor clustering as molecules within the same cluster are similar to those in different clusters. This is evident from the values on the off-diagonal, which are mostly identical to those on the diagonal and, in some cases, even higher (Figure 5B).

Taylor-Butina clustering has the advantage of ensuring that the molecules within the resulting clusters possess a Tanimoto value that is equal to or greater than the specified similarity cutoff [25]. The similarity cutoffs evaluated ranged from 0.35 to 0.99. Table 3B indicates that using smaller similarity cutoff values results in better-defined clusters but at the cost of having a larger number of clusters. A limitation of this algorithm is that it treats molecules that are not labeled at the end of clustering as singletons, even if they have neighboring molecules. Indeed, at a similarity cutoff of 0.9, Table 4B shows that 17% of the clusters obtained from the NCI-60 molecules are singletons. Moreover, 90% of all molecules and 93% of the non-outlier molecules have been assigned to a single cluster. Due to the presence of singletons, the average Tanimoto matrix cannot be calculated for these clusters, so some values are not visible in Figure 6B.

UMAP is a robust algorithm for simplifying complex data by reducing the number of dimensions while preserving its fundamental structure [89]. Unlike traditional linear methods, UMAP utilizes a non-linear approach to learning the underlying structure of the data and creating a low-dimensional embedding that retains the essential topological features of the original data. This new data representation can then be used to perform clustering [115]. Thus, as the first step is dimension reduction, we optimized the main hyperparameters of UMAP (Figure 7B). Unlike previous algorithms that determine the number of clusters based on the selected cutoff, in this case, the

number of clusters should be specified. We narrow the search for the optimal number of clusters using the elbow method (Figure 8B). All clustering quality metrics suggest well-defined clusters (Table 5B). Additionally, Figure 9B shows that molecules from different clusters have lower similarity than molecules from the same cluster, which indicates good clustering results. As for the number of clusters, increasing this number (from 7 to 20) increases the similarity between a cluster and itself from 0.16 to 0.22, for the upper bound (Table 5B).

An advantage of our work is its approach to simultaneously clustering all molecules within the NCI-60 panel rather than performing separate clustering runs for each of the 60 cell lines. This means that each molecule is clustered based on its structural characteristics, independent of its connection to a specific cell line. This approach saves both time and computational resources and ensures that the clustering results are consistent and coherent across the entire dataset. Additionally, this approach enhances scalability and applicability, as it provides a unified framework that can easily incorporate additional molecules or cell lines in future analyses without a significant computational burden. Once the clustering process is complete, each molecule is assigned to a unique cluster. Clustering results can be mapped to a specific cell line using a molecule's NSC ID. The clustering results, exemplified by the UMAP-based method, for each cell line are presented in Figure 4.9.

Since our main objective in this work was to generate clusters that can be used for model validation, our research focused on selecting the most efficient clustering method while keeping the computational cost low, where UMAP-based clustering is positioned as the best option. However, a key advantage of this work is providing detailed results and code for replication[1], enabling other researchers to leverage the clustering outcomes for various tasks beyond model validation. Contrary to clustering studies, which omit or limit the clustering evaluation to a single metric[39, 57, 104, 157], we present three well-known clustering metrics. These metrics are the Silhouette coefficient, the Calinski–Harabasz score, and the Davies–Bouldin score. Each clustering metric captures different aspects of clustering quality, such as compactness, separation, and cluster cohesion. Therefore, by comparing multiple algorithms and metrics, researchers can explore various possibilities and optimize their choices based on the specific requirements of their research objectives, computational resources, and interpretability considerations.

Improving the quality of clustering results can significantly enhance the generalization of predictive models on new compounds. This becomes particularly important when extending the model's scope and application. Hence, it is crucial to be critical while selecting clusters for model validation and employing methods such as leave-one-compound-cluster-out CV strategies. Overall, the findings suggest that the UMAP algorithm is the most effective for clustering the NCI-60 molecules. The number of clusters to be used must be chosen by considering the clustering quality metric se-

[1]The code for the reproduction of the best clustering results is available at https://github.com/Sahet11/nci60_clustering, accessed on 1 February 2024.

lected. Depending on the chosen metric, either 7 or 20 clusters can be used. However, the user should balance this choice with the computational cost they can afford, as more clusters mean more cross-validations.

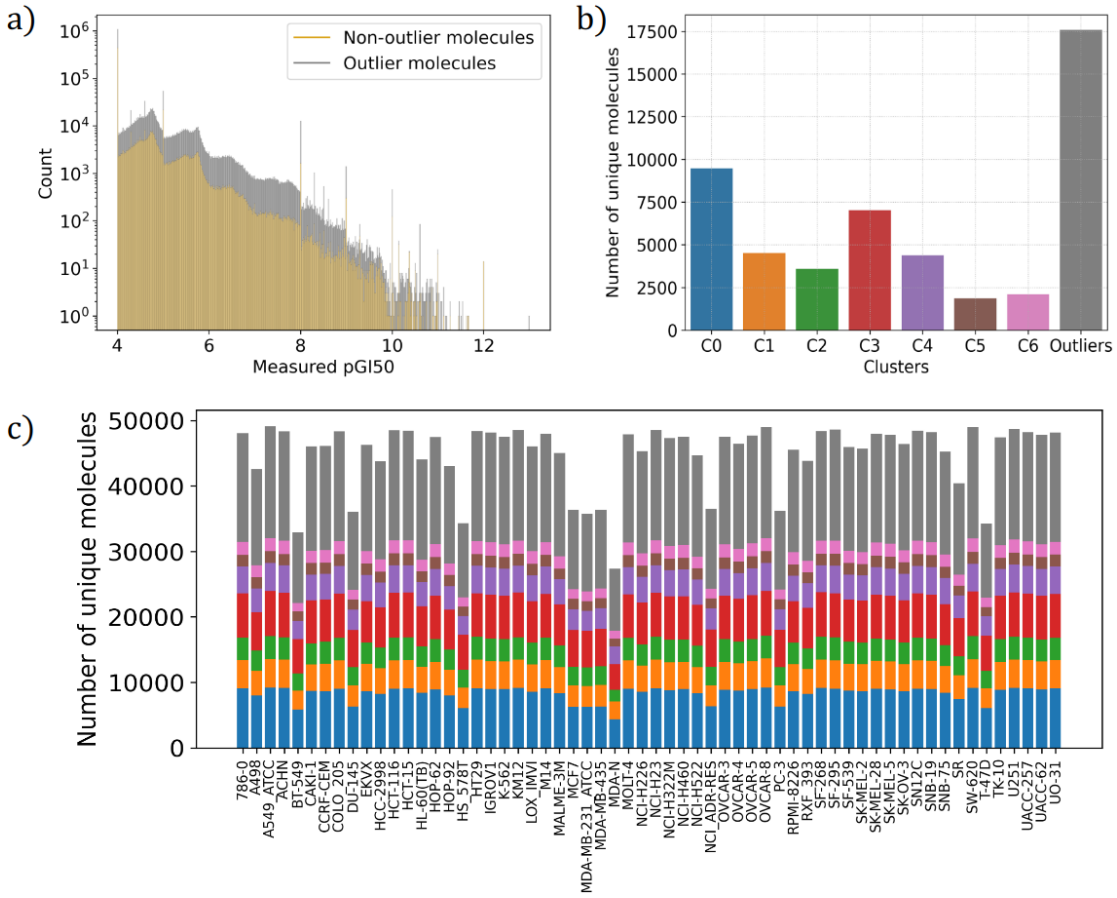


Figure 4.9.: **Description of the NCI-60 dataset after the outlier detection and clustering process.** a) pGI_{50} measurements from non-outlier and outlier molecules across the 60 cell lines for a total of 2,707,434 data points. Outlier molecules were retrieved at an outlier cutoff of 0.5. Potent molecules are rather scarce. b) UMAP-based clustering of the non-outlier molecules results in 7 clusters and provides the best set of clusters. These clusters, along with the outlier molecules, contain the 50,555 unique molecules in the NCI-60 panel. c) Number of unique molecules tested on each cell line broken down by cluster membership (same color code as in b).

4.4.5. Analyzing molecular clustering quality with Tanimoto similarity matrix

To use the generated clusters for model validation strategies, we need to ensure that they have optimal cohesion and separation. Cohesion refers to the closeness of the molecules within a cluster, while separation refers to the distance between molecules from two different clusters. Metrics such as the Silhouette coefficient, the Calinski–Harabasz score, and the Davies–Bouldin score are commonly used to assess clustering performance, primarily focusing on intra-cluster cohesion and inter-cluster

separation. However, in scenarios where molecular similarity is a key consideration, these metrics may not fully capture the nuanced relationships between clusters.

While these traditional metrics can sometimes overlook critical structural relationships between molecules across clusters, we introduced a new metric to address this issue. Based on molecular structural similarity, the average Tanimoto similarity matrix between clusters comprehensively evaluates intra-cluster cohesion and inter-cluster separation. This matrix is calculated as the average Tanimoto similarity between the molecules of cluster i and those in cluster j. The average of these similarity scores represents the position (i, j) of the similarity matrix. In the case of average similarity between molecules from a cluster and itself (position (i, i) of the similarity matrix), the similarity was calculated between two different molecules. This means that values in position (i, i) do not include similarity scores of a molecule with itself [63]. The calculation of this matrix is represented graphically in Figure 4.8-d).

Unlike the Silhouette coefficient, the Calinski–Harabasz score, and the Davies–Bouldin score, evaluating cluster quality based on molecular similarity offers a more nuanced measure of cluster effectiveness, enabling researchers to assess cluster coherence in terms of molecular properties. Additionally, the matrix values are constrained between 0 and 1, aligning with the interpretation of Tanimoto similarity. This presents a notable advantage over the Calinski–Harabasz score and the Davies–Bouldin score metrics, where either a lower or an upper bound is available, but not both. Furthermore, unlike these three metrics, this new metric is straightforward to visualize and interpret, where a well-performing clustering will exhibit diagonal values greater than those outside the diagonal in the matrix. Therefore, this metric comprehensively evaluates cluster efficacy for model validation strategies.

4.5. Integration of the previous steps in a real-world scenario

In this section, we address two scenarios that simulate real-world situations. The initial scenario entails building CP-based models assessed on a test set that differs from the training set (see Subsection 4.4.3 for reference). This test set comprises the outlier molecules found in [63]. As this scenario reflects a retrospective evaluation of the findings, we suggest a method to utilize CP predictions for prospective evaluations in the second scenario.

4.5.1. Retrospective evaluation of CP models on dissimilar test sets

This subsection aims to answer the research question of whether the CP model can generalize well to predict pGI_{50} values for dissimilar compounds in cancer cell lines.

Predicting these values is essential for their practical use in decision-making problems as it could facilitate the identification of potential drug candidates for further evaluation in experimental studies. This generalization ensures that the CP-driven approach remains robust and applicable beyond the training dataset.

The results of this research have been submitted to a peer-reviewed journal, and the preprint, "Conformal prediction of molecule-induced cancer cell growth inhibition challenged by strong distribution shifts" is available at:

> Saiveth Hernández-Hernández, Qianrong Guo, and Pedro J Ballester. "Conformal prediction of molecule-induced cancer cell growth inhibition challenged by strong distribution shifts". In: *Preprint from bioRxiv* (2024), pp. 1–16, https://doi.org/10.1101/2024.03.15.585269.

The full details of this work can be found in Appendix C, which is referred to as the LDMO-CCP paper. In this context, LDMO stands for Leave-dissimilar-molecules-out, and its detailed explanation will be provided later. CCP refers to cross Conformal Prediction (see Section 1.5.3) and represents a more complex version of CP.

The fundamental aim of this study was to create reliable predictive models that could predict the inhibitory activity (pGI_{50}) of molecules in cancer cell lines. In the initial experiments, six different algorithms were evaluated per cell line to determine their efficacy in predicting molecular inhibitory activity (Section 4.2). After careful evaluation, four models were identified as having promising performance and were selected to proceed to the optimization stage. The selected models were fine-tuned during the optimization stage to maximize predictive accuracy and generalization. This involved addressing key aspects such as hyperparameter tuning and feature representation (Section 4.3). Finally, to generate clusters that can be used in model validation strategies, we obtained the best way to cluster molecules in the NCI-60 panel (Section 4.4), [63].

As a result of these previous experiments, we represent molecules using 1024-bit MFPs as this size represented the best balance between performance and computational cost (Figure 4.4 and Table 4.1). Additionally, this work evaluates the impact of integrating information from 7 physicochemical descriptors (Figure 4.10-a).

To address the challenge of model performance when dealing with dissimilar, unseen molecules, we propose a dual-phase approach to building and evaluating CP effectively. The initial phase involves applying the LDMO CV to validate CP models (Figure 4.10-b). The LDMO validation underlines that the test set molecules are not only unseen but also dissimilar to those in the training set. The LDMO-CV approach represents a modified variant of cross-CP, where the folds are obtained through a previous clustering process [63], leading us to train seven CP models per cell line. We trained 60 CP models, one per cell line, using the non-outlier molecules (Figure 4.10-c). Note here that each fold corresponds to a trained model. The CP models built are evaluated using regression, classification, and CP metrics (Figure 4.10-d).

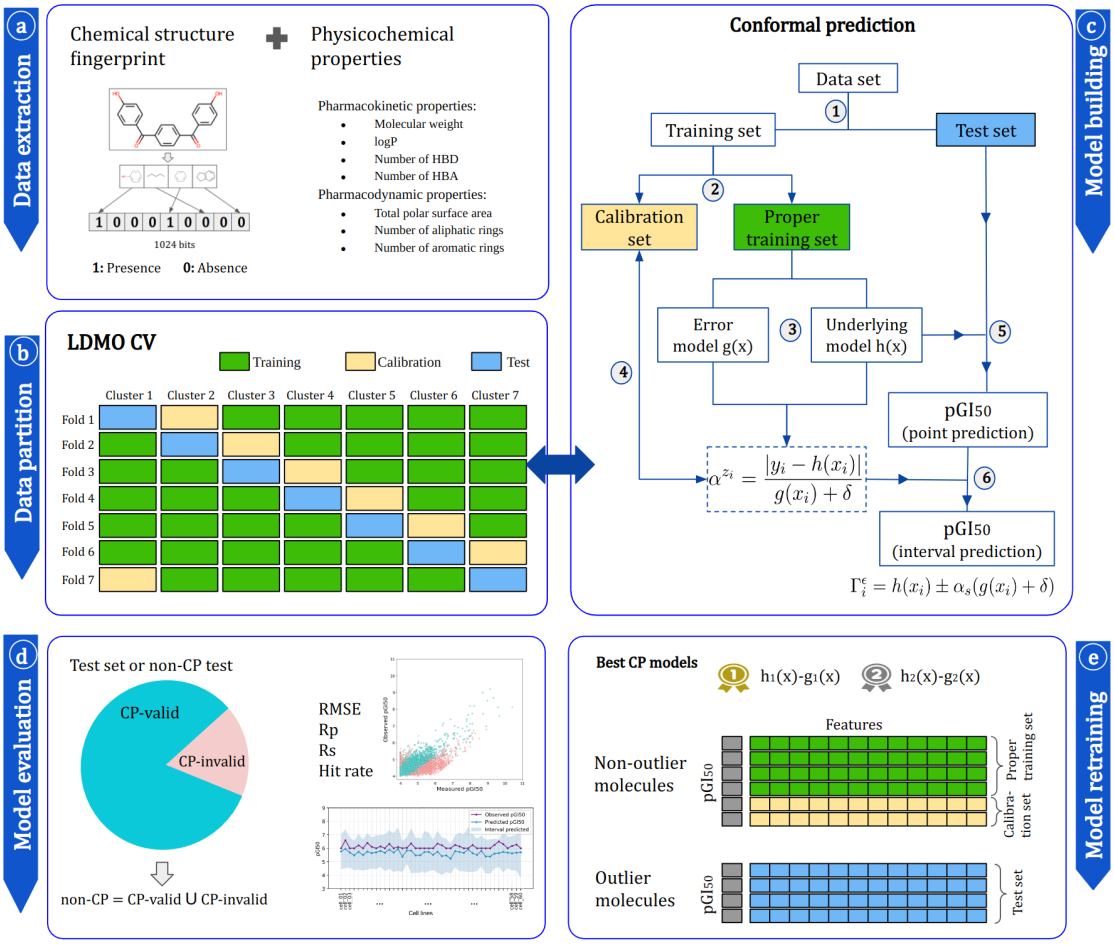


Figure 4.10.: **Summary of LDMO-CCP model development.** For each of the 60 cell lines: a) Generating the features for each molecule using MFPs and PCDs to encode their molecular characteristics. b) For each cell line and model, multiple LDMO-CV validation runs are performed to evaluate how well the considered model can predict the cell activity of molecules with different chemistry. c) For each cell line, the underlying model is trained and tested on each fold of the non-outlier molecules. d) The standard machine learning metrics and the conformal prediction methodology metrics are used to evaluate CP model performance. Note that b) and c) correspond to the first stage of this study. e) In the second stage, the top two CP models identified during the first stage are subsequently re-trained using the non-outlier molecules. Following re-training, these models were evaluated using the outlier molecules to assess their performance on completely new chemistry.

In the subsequent phase, we select the two best-performing LDMO-CCP models from the first stage. These models are then retrained using all the information available from the non-outlier dataset and evaluated on the outlier molecules (Figure 4.10-e). Therefore, the second stage employs a basic version of ICP.

Table 4.3 shows the combinations of the considered CP models, each of them evaluated at four confidence levels: 80%, 85%, 90%, and 95%. Four of the five models listed in this table correspond to the four best models found during the baseline analyses (see Section 4.2). The fifth model, LR-LR, is added to measure the improvement obtained

by moving from linear to nonlinear models.

Table 4.3.: Summary of the LDMO-CCP models built and evaluated. Each CP model is specified by its $h(x) - g(x)$ combination. The algorithms selected are LR, RF, and XGB, where LR serves as a benchmark for comparing the performance of non-linear algorithms. Therefore, some $h(x) - g(x)$ combinations are excluded from this study. RF and XGB models underwent a hyperparameter tuning step to optimize their performance. The LR algorithm utilizes default parameters without additional tuning.

Underlying model h(x)	Error model g(x)		
	LR	RF	XGB
LR	LR-LR	–	–
RF	–	RF-RF	RF-XGB
XGB	–	XGB-RF	XGB-XGB

A total of 4200 models (60 cell lines × 5 CP models × 2 set features × 7-folds) were built and evaluated in the first part of this study. In the second part, a total of 120 models (60 cell lines × 2 CP models) were built and evaluated. The performance of the regression models was evaluated by computing the RMSE, the Rp, and the Rs. The efficiency and validity of CP were also evaluated.

In the upcoming subsections, we will discuss the primary outcomes obtained by building the LDMO-CCP models. These subsections provide comprehensive insight into the development process results and contrast them with random CP models.

4.5.1.1. Retrospective evaluation results in the non-outlier data set

In Subsection 4.3.2, we explored the influence of molecule representation on model performance. To address this, we examined various MFP sizes and selected the one that offered the optimal balance between performance and computational resources. Building upon this investigation, we now extend our analysis to evaluate how incorporating information on physicochemical descriptors affects model performance. Therefore, each model listed in Table 4.3 underwent testing using two distinct sets of features. These feature sets included a 1024-bit MFP and a combination of MFP with seven physicochemical descriptors (MFPs+PCDs).

Incorporating PCD can offer valuable insights into molecular structure and properties that MFP alone may not capture. Additionally, utilizing PCD can address any potential redundancy in the feature space that may arise from MFP encoding similar structural information for various molecules, resulting in redundant features. This additional information can help the model better capture the underlying patterns in the data and make more accurate predictions. As demonstrated in Figure 2C, including PCDs as input features (MFPs+PCDs) results in a statistically significant improvement in the Rs and Rp metrics. Therefore, this representation of molecules is used for the following experiments.

As expected, the nonlinear models have demonstrated better performance than the

linear model. Therefore, only the four nonlinear models will be considered for the upcoming experiments.

Our research assessed four confidence levels: 80%, 85%, 90%, and 95%. These confidence levels correspond to expected errors of 0.2, 0.15, 0.1, and 0.05, respectively. In regression tasks, the expected errors signify the quartile in which the non-conformity score will be selected (see Section 1.5.3 for reference). We represented the uncertainty distribution for each confidence level in Figure 4.11, with dotted lines indicating the requested error value for each level. We calculated the uncertainty by multiplying the predicted error for a given molecule by the non-conformity score chosen at a specific confidence level. To calculate the uncertainty, we multiplied the predicted error for a given molecule by the non-conformity score picked at a particular confidence level, which can be expressed as $uncertainty = 2\alpha_s g(x_i)$.

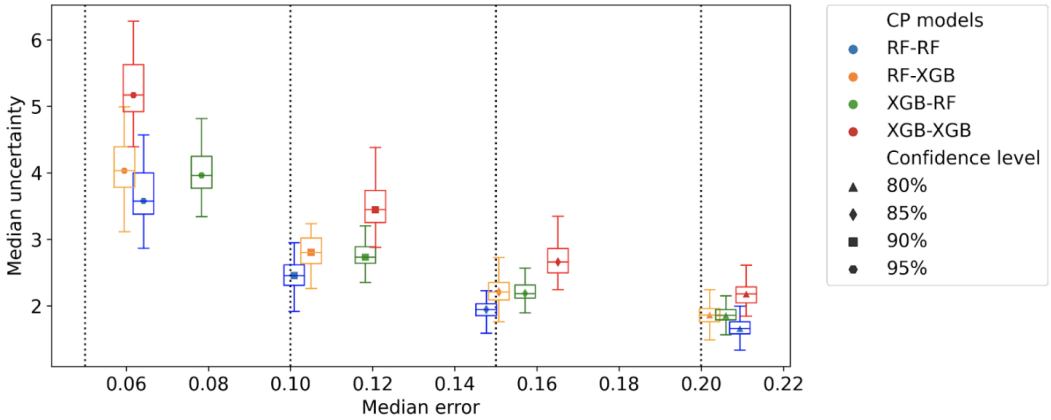


Figure 4.11.: **Trade-off between prediction uncertainty and measured error of CP models.** For each CP model, median error and uncertainty are computed using MFPs+PCDs as input descriptors at different confidence levels. The cell lines' performance is summarized as 7-fold-median uncertainty (y-axis). Thus, boxplots contain 60 points, representing the median uncertainty per cell line. The median of 7-fold-median error values is calculated across 60 cell lines to summarize the error performance. A single point is obtained that represents the median error of the entire group, and each boxplot is positioned along the x-axis based on the corresponding median error obtained. The vertical dotted lines represent the expected error at each confidence level. Color code refers to the four CP models ($h(x) - g(x)$) employed. Markers refer to the requested confidence level.

It is important to note that at higher confidence levels, the non-conformity score tends to be higher, indicating greater dissimilarity between the molecules in the training set and the new molecule. This translates to higher uncertainty in the predictions, regardless of the CP model used. In summary, a high non-conformity score means that we allow for more different values, resulting in a higher degree of uncertainty in the predictions.

Although the four CP models generally tend to have higher errors than expected, their median validity is close to the expected error for all models. Moreover, the CP models that use the XGB algorithm as their error model tend to have higher uncertainty than

those that use the RF algorithm as their error model. For an 80% confidence level, the four CP models achieve the best uncertainty results, with values ranging from 1.6 to 2.1 pGI_{50} units. However, higher confidence levels are associated with higher uncertainty, so the choice of confidence level should be based on the specific task. pGI_{50} predictions at a lower confidence level are more reliable in this case. Therefore, we will use the 80% confidence level for the following experiments.

The best CP models are selected based on several factors calculated with an 80% confidence level. Our selection process is based on models with lower Interquartile Range (IQR), lower median uncertainty, and lower errors across the 60 cell lines. These criteria ensure that the chosen CP model is both consistent and accurate in its predictions at the specified confidence level. Out of the four CP models we evaluated, the RF-RF and XGB-RF models showed the lowest IQR values, measuring 0.179 and 0.156 pGI_{50} units, respectively. The RF-RF model achieved a median uncertainty of 1.66 and a median error of 0.209 pGI_{50} units. In comparison, the XGB-RF model achieved a median uncertainty of 1.86 and a median error of 0.206 pGI_{50} units.

After identifying the two best CP models, we conducted a study to evaluate their predictions at an 80% confidence level (Figure 3C). Our analysis involved comparing these predictions with non-CP predictions, and the results showed that using CP significantly improves prediction accuracy. Specifically, we observed a decrease of approximately 0.3 pGI_{50} units in the median RMSE of CP-valid predictions compared to non-CP predictions. This improvement was also evident in correlation (Rp and Rs), where the CP-valid predictions demonstrated better alignment between the predicted and observed pGI_{50} values.

Useful predictions are achieved despite training on highly dissimilar molecules

To demonstrate the practical use of LDMO-CCP models for molecular screening to inhibit tumor cells, we have specifically targeted SK-OV-3 and TK-10 cell lines. The LDMO-CCP models identified these cell lines as top performers. Metrics on their performance are presented in Table 1C.

To simulate a real-world screening scenario, we employed our CP models to identify the 100 most promising CP-valid molecules with the highest predicted pGI_{50} values for these cell lines. We subsequently refined our search to the three molecules with the highest measured pGI_{50} values. This process is visually represented in Figure 4.12-a), where the 100 CP-valid molecules are enclosed within a red oval region for easy identification, and the top 3 molecules are highlighted within a blue oval region. This process mirrors a 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 that 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 these candidates, we identify the training molecule that exhibits the highest degree of similarity to the best-selected CP-valid molecule (Figure 4.12-b). This selection is made as the other five candidates exhibit a higher dissimilarity to this particular molecule.

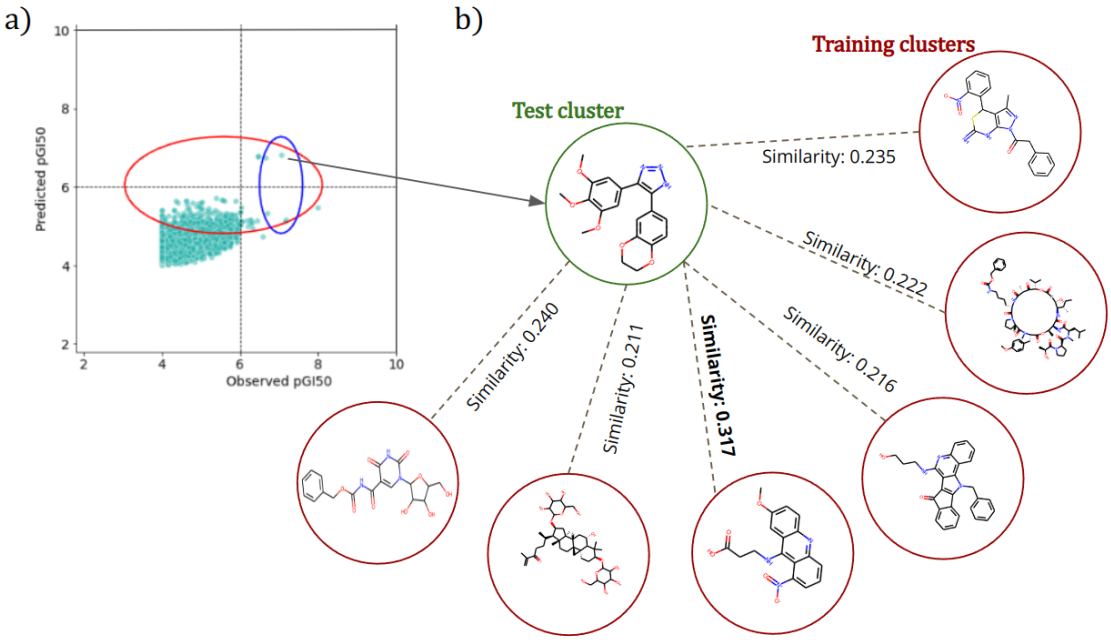


Figure 4.12.: **Dissimilarity measure between the top 3 CP-valid molecules and those observed during training.** a) We initially pinpointed the top 100 CP-valid molecules with the highest predicted pGI_{50} values for the selected cell lines (see Table 1C for reference). These molecules are highlighted within a red oval region for easy identification. Subsequently, we narrowed our focus to the three molecules with the highest measured pGI_{50} values, which are denoted within a blue oval region in the figure. b) We find the closest training molecule for each of the three best CP-valid molecules by computing the Tanimoto similarity. We select the most similar molecule from six candidates (one per cluster), as the other five are more dissimilar. The most similar molecule from the training set to the test molecule is highlighted in bold.

Table 2C shows moderate similarities observed between molecules, ranging from 0.29 to 0.42, regardless of the CP model used (RF-RF or XGB-RF). This suggests that there are subtle resemblances or shared characteristics among the molecules, as shown in Figure 4C. The XGB-RF model identified the three most promising molecules, all of which had predicted pGI_{50} values exceeding 6, categorizing them as active compounds. It is worth noting that the most similar molecule in the training set for all three promising compounds was consistent (NSC-662190), indicating a recurring pattern in their structural features. On the other hand, only one of the three promising molecules identified using the RF-RF model was deemed an active molecule (predicted pGI_{50}=6.8).

4.5.1.2. Outlier set results

Retraining the two best LDMO-CCP models

The previous sections have discussed the CP models and their training and evaluation using a relatively controlled dataset that does not contain any outliers. However, a crucial question is whether these models can generalize well to molecules that are not part of the training set. To address this, we retrain a CP model on the non-outlier dataset and test its performance on the outlier dataset (Figure 4.10-e). This experimental design, which allows for a comprehensive evaluation of the adaptability and performance of a CP model in handling different scenarios it may encounter during training, reinforces the confidence in the models' potential use in various real-world scenarios.

Following the scheme shown in Figure 4.10-e), the two best CP models are retrained using the complete non-outlier molecules dataset. Six of the seven clusters are utilized as the proper training set, while the remaining cluster serves as the calibration set. The uncertainty value determines which cluster will be the calibration set, selecting the cluster with the lowest uncertainty across the 60 models (one per cell line). Following the retraining of the models with non-outlier molecules, we then evaluate their performance on the outlier molecules.

LDMO-CCP models effectively generalize to outlier test sets

We conducted a comparative evaluation to assess the performance of two CP models on both outlier and non-outlier datasets. The benchmark for this analysis was the model's performance within the non-outlier dataset. The performance of each cell line in the non-outlier dataset was summarized through 7-fold-median scores indicating either error or uncertainty. These individual scores were then aggregated by calculating their median across the entire set of 60 cell lines. This calculation resulted in a single value representing the median error exhibited by the entire group of cell lines. In the outlier dataset, the median across the 60 cell lines was straightforwardly calculated to determine the single point encompassing the median error or the median uncertainty of the entire group of cell lines.

We created a calibration plot to illustrate the relationship between the expected error and the actual error achieved by an LDMO-CCP model (Figure 4.13-a). An Error-Uncertainty plot was also generated to show the relationship between the expected error and the associated uncertainty attributed to achieving the requested error (Figure 4.13-b). This comparison was made at four confidence levels: 80%, 85%, 90%, and 95%.

Figure 4.13 shows that when CP models are trained on non-outlier molecules and tested on outlier molecules, they consistently show lower errors than their performance on the non-outlier dataset. Although all CP models are relatively well-calibrated, the models tested on outlier molecules achieve lower errors than expected. This is often perceived as an advantage and highlights the model's ability to generalize to

molecules that differ from its training dataset. Regarding the uncertainty, these results emphasize the model's ability to generate more reliable predictions on the outlier test set, particularly when the RF-RF model is used.

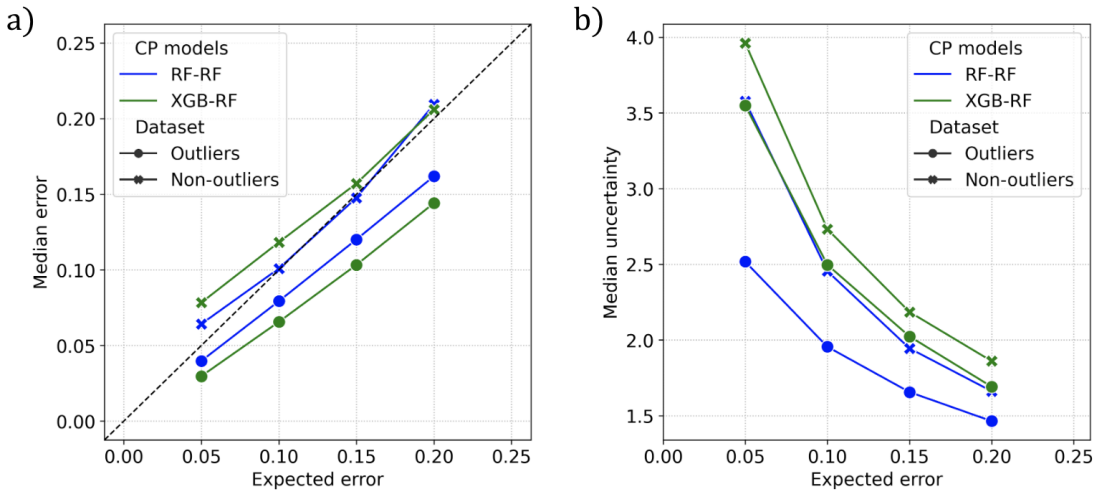


Figure 4.13.: **CP models can effectively generalize to outlier test sets and make low-uncertainty predictions.** Two LDMO-CCP models were compared using outlier and non-outlier datasets at different expected errors. The model's performance within the non-outlier dataset was evaluated through 7-fold-median scores and combined to find their median across all 60 cell lines. In the outlier dataset, a straightforward calculation of the median across the 60 cell lines leads to the single point encompassing the median uncertainty or error of the entire group of cell lines. a) A calibration plot illustrates the expected and actual error relationship. b) Error-Uncertainty plot to illustrate the relationship between the expected error and the associated uncertainty attributed to achieving the requested error. Color code refers to LDMO-CCP models, either RF-RF or XGB-RF. Markers refer to datasets that are either non-outliers or outliers.

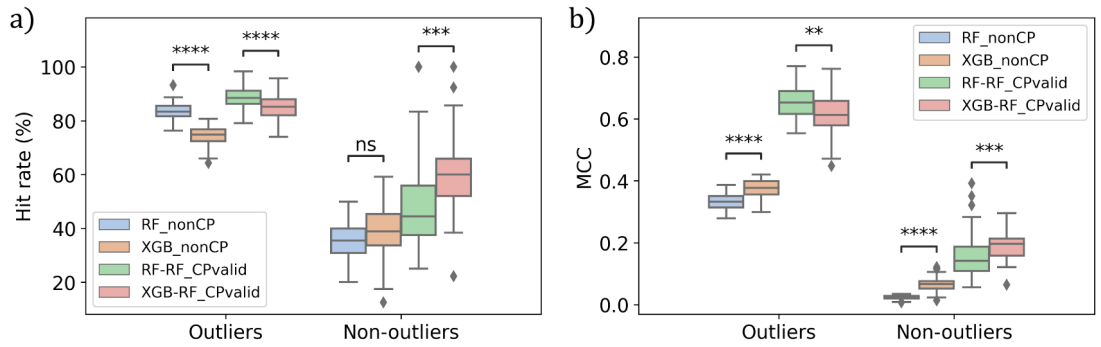


Figure 4.14.: **CP-valid molecules derived from models assessed on the outlier test set yielded higher hit rates.** A comparative analysis evaluated the performance of the two best LDMO-CCP models at an 80% confidence level. The a) hit rate and b) MCC distribution across 60 cell lines were analyzed. Statistical test results for the underlying models were displayed at the top of the boxplots. The color code stands for a key assigned to identify the model used and the type of prediction (nonCP or CP-valid). 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; ****:p ≤1.00e-04.

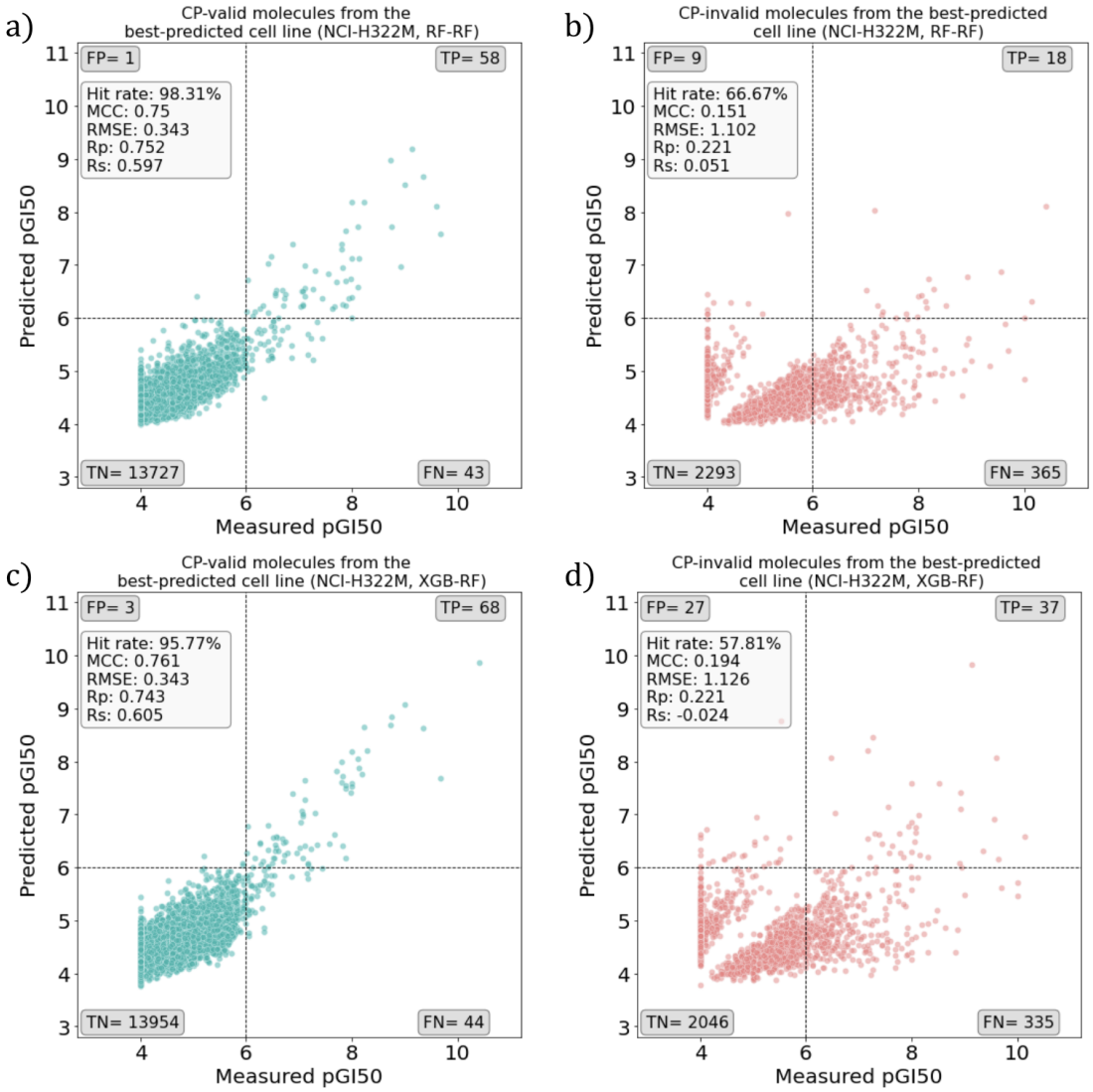


Figure 4.15.: pGI_{50} **predictions of molecule-cell line pairs improve when using LDMO-CCP.** The pGI_{50} value observed and predicted in the best-predicted cell line, obtained by the RF-RF (top) and the XGB-RF (bottom) CP models. The color code indicates the test set molecules, either CP-valid (left) or CP-invalid (right). The vertical and horizontal dotted lines show the threshold for inactive and active molecules ($pGI_{50} \geq 6$). The number of FP, TP, TN, and FN is indicated for CP-valid and CP-invalid predictions. These values are used to compute the hit rate and MCC for each cell line.

Predictions made by CP models were retrieved at an 80% confidence level, and the hit rate was obtained by applying a threshold that distinguished between active and inactive molecules (see Section 3.4 for reference). The hit rate and MCC distribution across 60 cell lines were analyzed, with a color code used to identify the model used and the type of prediction (nonCP or CP-valid). At the top of the boxplots, statistical test results for the underlying models were presented.

Figure 4.14 indicates that focus on CP-valid molecules derived from models assessed on the outlier test set results in higher hit rates. Collectively, these findings suggest

that CP models are not only able to generalize to different data sets but are also able to generate predictions with a low level of uncertainty. This highlights the potential of the top-performing CP models to deliver reliable predictions in real-world applications, particularly those that may involve previously unseen data distributions.

In our experiment, we consider the cell line with the highest hit rate to be the best-performing cell line. Conversely, the cell line with the lowest hit rate is regarded as the worst-performing cell line. According to the analysis, both the RF-RF model and the XGB-RF have identified the NCI-H322M cell line as the best one. However, the hit rate and uncertainty vary in each case. The RF-RF has achieved a higher hit rate and assigns more confidence to the predictions for the molecules tested against the NCS-H322M cell line. Figure 4.15 displays the best-performing cell line obtained by the CP models. In both cases, the subset of CP-valid molecules has higher hit rates than the CP-invalid predictions.

LDMO-CCP models can predict the NCI-60 cell line selectivity of a test molecule

Our goal is to identify the best CP-valid molecule for each cell line using the lowest uncertainty as the criterion. This process is applied to both CP models. For the RF-RF model, we merge the data for each molecule across all 60 cell lines and filter out only the CP-valid and most potent molecules ($pGI_{50} \geq 6$). This results in 9,806 NSC-Cell points for 904 small molecules screened against the 60 cancer cell lines. We narrowed our search to only those NSCs where data is available for at least 50 cell lines, leaving us with 49 NSCs. From these, we selected the best-performing CP-valid molecules based on their uncertainty.

For the XGB-RF analysis, we followed a similar procedure. After the first filter, there were 11,969 NSC-Cell points for 1,471 small molecules screened against the 60 cancer cell lines. We narrowed our search to NSC compounds with data available for at least 50 cell lines, which resulted in the identification of 29 NSC compounds. From these, we selected the best-performing CP-valid molecules.

The selected molecules are NSC-747457 for the RF-RF and NSC-7521 for the XGB-RF, with average uncertainties of 1.81 and 1.3 units, respectively. It is important to highlight that there is an improvement compared to the uncertainty achieved by the selected NSC in the non-outlier test set. Figure 6C displays the relationship between the predicted and measured pGI_{50} values across different cells. It is also important to mention that the NSC-7521 molecule shows a high potency when tested against 58 different cell lines, with measured and predicted pGI_{50} values greater than 8 units. The computation of the average uncertainty for these molecules is based on a similar number of cell lines. Specifically, the average uncertainty in the RF-RF model was determined across 57 cell lines, while for the XGB-RF, it was computed across 58 cell lines.

4.5.2. Comparing CP-based models for Drug Design

Throughout this thesis, we presented four different CP-based approaches for predicting the inhibitory activity of molecules in cancer cells. Each model presents different advantages and disadvantages, summarized in Table 4.4.

The first model examined was a baseline CP, discussed in Section 4.2. In this subsection, we refer to this model as the random CP model. This model serves as a baseline for comparison and is relatively easy to implement [65]. The second model explored is the random ACP model, which comprises 10 aggregated individual CP models. This approach of combining multiple random CP models helps reduce the inherent variability in a single partition. The third model analyzed is the CCP model. This model can provide more reliable performance estimates using a cross-validation strategy. The ACP and CCP models are advanced versions of the CP model. Finally, the LDMO-CCP model was examined [64]. This model addresses traditional cross-validation limitations by ensuring that the test set contains dissimilar molecules to those used for model training.

Table 4.4.: Summary of the advantages and disadvantages of the CP-based models compared in this study. By systematically comparing and evaluating four distinct CP models, this research contributes valuable insights to the field of drug design.

CP model	Advantages	Disadvantages	Ref
Random CP	Provides a basis for comparison and is relatively simple to apply	Overfitting or underfitting on unseen data due to its reliance on a single random partition	[65]
Random ACP	Aggregating multiple random CP models can mitigate the variability inherent in a single partition	Sensitivity to the limitations of the individual CP models it comprises	4.3.4
Random CCP	Cross-validation help in assessing model generalization and can provide more reliable estimates of performance	The traditional CV method can lead to optimistic estimates of performance, especially if there is a high degree of correlation between molecules	4.3.4
LDMO-CCP	LDMO-CCP address the shortcomings of traditional CV by ensuring that the test set comprises molecules dissimilar to those used for model training	Increased complexity due to the need for dissimilarity-based clustering	[64]

Our findings demonstrate that all CP models outperformed the standard ML, even when using the baseline random CP method (Figures 2 A, 5 A, and 6 A). This highlights the utility of incorporating CP to enhance predictive performance and uncertainty quantification.

Notably, the LDMO-CCP method exhibited superior hit rates compared to the random CP (Figure 8A and Figure 4.15). This suggests that incorporating the LDMO approach

into CP can enhance the prediction of the inhibitory activity of molecules on a given cancer cell line. However, it is noteworthy that achieving higher hit rates with the LDMO-CCP model came at the cost of increased error (RMSE) and correlation (Rp), as shown in Figure 7A and Figure 2C. This trade-off between predictive accuracy and reliability is a critical aspect to consider in practical applications of CP.

Moreover, in terms of uncertainty, the LDMO-CCP model generally assigns lower uncertainty to the predictions. Indeed, in the random CP models, the median uncertainty ranges from 1.43 to 2.74 pGI_{50} units, while in the LDMO-CCP, it ranges from 1.6 to 2.1 pGI_{50} units. See Figure 4A and Figure 4.11.

Each approach has its own set of advantages and disadvantages, and the choice depends on factors such as computational resources, dataset characteristics, and desired level of model complexity.

4.5.3. Prospective evaluation of CP models to guide decision-making

The studies examined in the literature review are retrospective evaluations that rely on historical data (Table 2.1). However, conducting prospective evaluations is essential to assess the model's performance in practical applications objectively. In this context, it becomes imperative to determine how to effectively utilize the predictions generated by CP models in the absence of ground truth labels.

Incorporating CP into current decision-making protocols to facilitate potential screenings could improve decision-making by delivering predictions with a user-defined confidence level. This enables the prioritization and efficient allocation of resources towards highly reliable and promising compounds, resulting in more well-informed decisions during the drug discovery process. Therefore, this subsection aims to answer the research question of how CP can be integrated into existing workflows and decision-making processes to support prospective drug screening. The results presented in this section are part of ongoing research.

The term "CP-enhanced" is introduced herein to denote molecules whose pGI_{50} value has been predicted using CP models. The primary objective of this study is to determine the most suitable definition of CP-enhanced by evaluating three proposed criteria:

- C1: Top N molecules with the highest predicted pGI_{50} (benchmark criteria)
- C2: The Top N molecules with a combination of high predicted pGI_{50} and small interval are retrieved from a pool of $3N$ molecules with the highest predicted pGI_{50}.
- C3: Top N molecules with the highest predicted pGI_{50}-low (i.e., the best-left side of the interval).

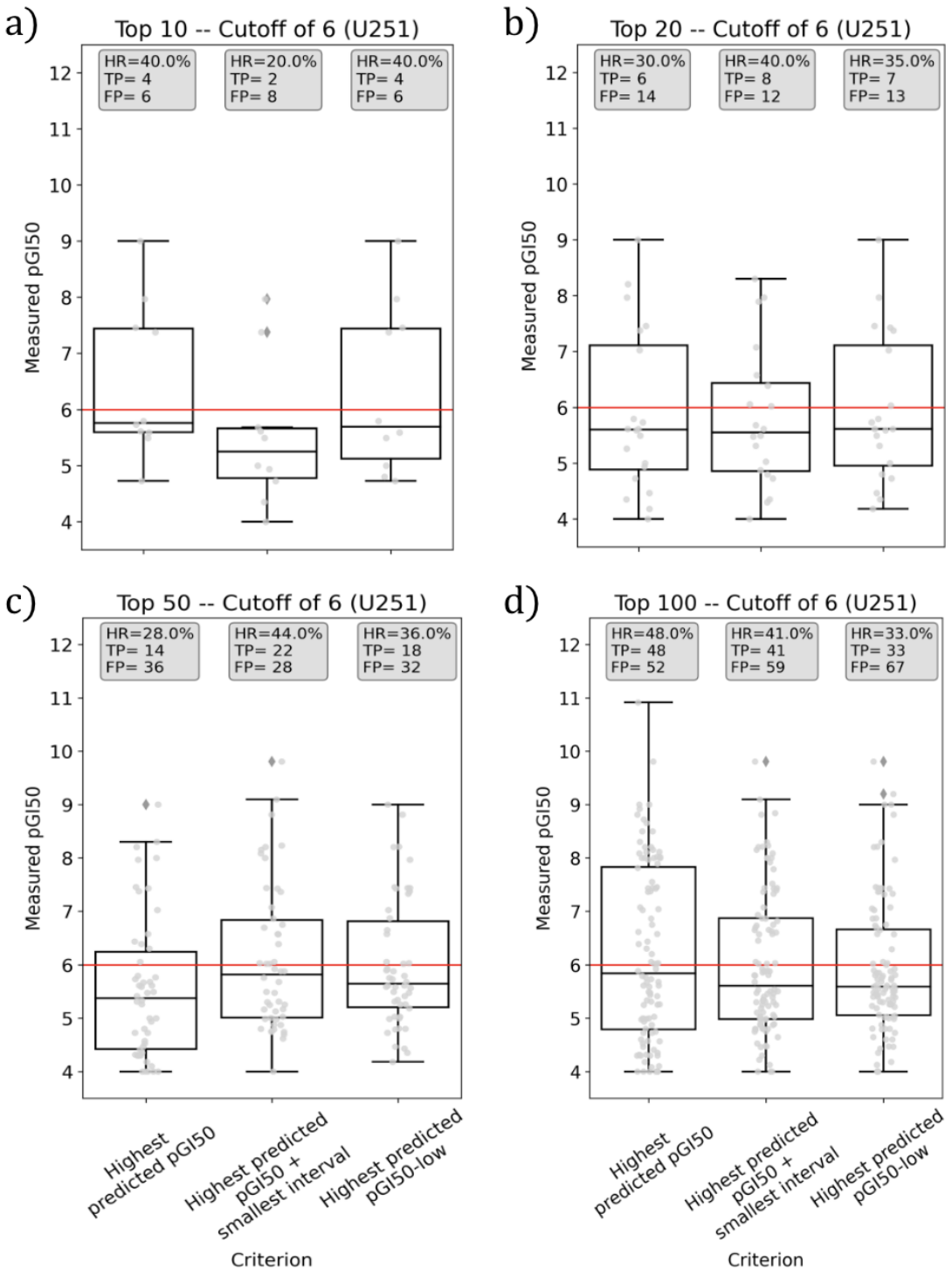


Figure 4.16.: Boxplots for the three criteria selected. From left to right, each boxplot represents the distribution of the top N molecules with the highest predicted pGI_{50}, the top N molecules with the highest predicted pGI_{50} (top M), and with the smallest interval; the top N molecules with the highest pGI_{50}-low (best lower endpoint of the predicted interval). The horizontal line on measured pGI_{50}=6 represents the threshold between active and inactive molecules: every above is true positive, below false positive. a) Top N=10, b) Top N=20, c) Top N=50, d) Top N=100. All metrics are calculated using only the top N molecules. Predictions are from the RF-RF model.

To identify the optimal approach for defining CP-enhanced molecules, we investigated various values of N. Specifically, we tested the values of 10, 20, 50, and 100. Experiments were conducted on the U251 cell line. Each criterion was assessed based on the hit rate, with pGI_{50}=6 being the threshold between active and inactive molecules: every above is true positive, below false positive. Our analysis revealed that N=50 obtains the best hit rate among the three criteria, with the second criterion obtaining the best results (Figure 4.16). Based on Figure 4.17, in a prospective setting on a given cell line, we use the top 50 predictions obtained with C2 (molecules with a combination of high predicted pGI_{50} and small interval) as this method represents the least risk and the highest benefit on average across cancer cell lines.

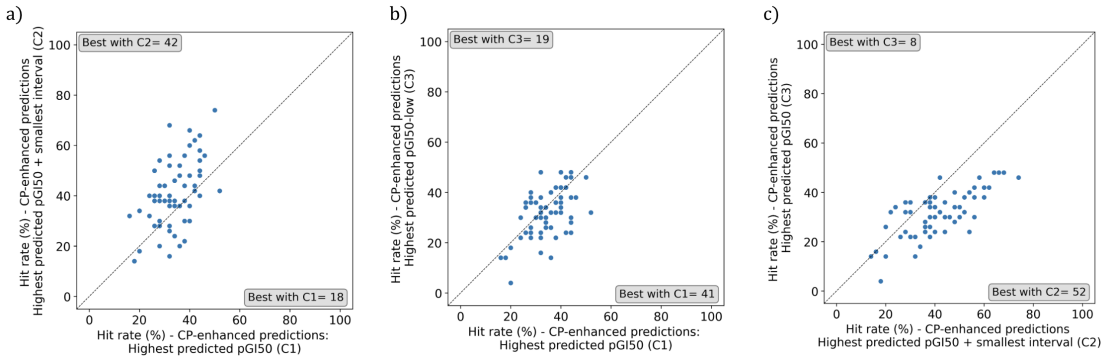


Figure 4.17.: CP-enhanced predictions, defined as the 50 non-outlier molecules with a combination of high predicted pGI_{50} and small interval (criterion 2), provide better hit rates than the other two criteria. Correlation plot between a) criterion 1, 50 molecules with the highest predicted pGI_{50}, and criterion 2, the 50 molecules with a combination of high predicted pGI_{50} and small interval, b) between criterion 2, the 50 molecules with a combination of high predicted pGI_{50} and small interval, criterion 3, 50 molecules with the highest predicted pGI_{50}-low, c) criterion 2, the 50 molecules with a combination of high predicted pGI_{50} and small interval, and criterion 3, 50 molecules with the highest predicted pGI_{50}-low.

5. Conclusions and future work

Contents

5.1. Conclusions 114
5.1.1. General conclusions 114
5.1.2. On the accuracy of CP models to predict pGI_{50} values for different compounds in cancer cell lines 115
5.1.3. On the impact of data split and molecular representation 116
5.1.4. On the CP model generalization to predict pGI_{50} values for dissimilar compounds 117
5.1.5. On the integration of CP into decision-making processes to support prospective drug screening 117
5.2. Future work 118

5.1. Conclusions

5.1.1. General conclusions

This thesis project represents a significant contribution to the field of drug design, particularly to property prediction problems, not only due to the use of ML and CP but also because of the novel methodology we introduce. Our proposed methodology begins with a baseline CP model to measure its accuracy in predicting the inhibitory activity of molecules in cancer cell lines. The CP-based models developed at this early stage outperformed traditional ML-based models by reducing the error (RMSE) and increasing the correlation (Rp) of predictions. We then focus on optimizing key aspects of these models, addressing research gaps in previous works. We focused on finding the best balance between performance and computational cost at this stage. Finally, we evaluated the optimized model's performance in scenarios that simulate real-world situations for retrospective and prospective evaluations. In retrospective evaluations, we found that the optimized CP models can generalize their results to test sets containing molecules different from those observed during training. This comprehensive methodology ensures the reliability and applicability of our predictive CP-based models. These findings will be discussed in more detail in the following sections.

Therefore, we successfully established a proof-of-concept framework after extensive experimentation and analysis of these key methodological components. This frame-

work provides a strong basis for evaluating the efficacy of new compounds in inhibiting or killing tumor cells. Our framework uses the NCI-60 human tumor cell line screen [100], which consists of a diverse panel of 60 different cancer cell lines, providing a comprehensive testing ground. We used this dataset to evaluate the performance of our predictive models across various cancer types, making it a robust tool for cancer research.

Furthermore, by incorporating CP into our predictive modeling framework, we have introduced a layer of uncertainty estimation that enhances the robustness and reliability of our predictions. Our CP models accurately predict molecules' growth inhibition, which provides a valuable tool for identifying promising candidates for further drug development and therapeutic interventions in oncology. This not only facilitates more informed decision-making but also instills confidence in the predictive outcomes generated by our approach. This breakthrough can significantly enhance the effectiveness of cancer treatments by accelerating the development of new drugs.

5.1.2. On the accuracy of CP models to predict pGI_{50} values for different compounds in cancer cell lines

Accurately assessing the reliability of pGI_{50} predictions is essential in determining the effectiveness of CP models. Therefore, during the initial stages of our research, we conducted experiments to evaluate the accuracy of CP models in predicting the inhibitory activity of molecules on cancer cell lines. In addition to comparing the performance of various CP models against each other, we investigated whether CP models could outperform traditional (or standard) ML models.

The results of these experiments were evaluated in terms of error, measured by the RMSE, and correlation, measured by the Rp. By comparing the performance of CP models with standard ML models, we have observed a consistent trend across all analyzed experiments: the CP models consistently outperform the standard ML models. In addition, a primary advantage of leveraging CP models is that an uncertainty is assigned at the instance level. Thus, for each molecule, we can estimate the reliability of the delivered prediction and focus on the most reliable ones.

To complement this comparison, we introduce a novel approach by defining "CP-valid" predictions as the outcomes where, given a confidence level, the true label falls within the predicted interval. Our proposed definition represents a more efficient way to approach more reliable molecules. In contrast to other works, our work formally defines both CP-valid and CP-invalid molecules, offering a targeted approach to identifying promising candidates for further stages of drug development.

Our research findings indicate that when we prioritized CP-valid molecules, the top-performing models achieved a hit rate ranging from 75% to 85%. These results are particularly promising given that the models analyzed in this stage were optimized for hyperparameters and data partitioning, ensuring appropriate training and calibration

set sizes. These findings suggest that CP models show promise in accurately predicting cell activity. Therefore, an optimization phase is needed to improve the efficiency of these models further.

5.1.3. On the impact of data split and molecular representation

The optimization phase was critical to improve the predictive performance of CP models. Building upon promising results achieved by initial CP-based models, we focused on two key areas of optimization: data split and molecular representation.

We investigate two aspects of molecular representation: the MFP size and the incorporation of PCD as part of the input features. Our findings indicate that enlarging the MFP size correlates with decreased error and increased prediction accuracy, as measured by the RMSE, Rp, and Rs. However, it is crucial to note that while increasing the MFP size can yield performance improvements, a threshold exists for its efficacy. Specifically, our analysis reveals that the performance, as measured by RMSE, exhibits only marginal enhancement when transitioning from 1024 MFP to 2048 MFP. Moreover, the computational time required to construct the models escalates significantly with larger MFP sizes, with approximately a five-fold increase in computation time. By integrating the information on the required execution time, we consider an additional criterion generally omitted from these types of studies.

Once the optimal size of the MFPs was found, we extended our investigation to determine the impact of integrating information on the molecules' PCDs. Likewise, our results showed an improvement by incorporating such features. While MFPs capture structural information, PCDs provide complementary information about the chemical properties of molecules, which can help the model generalize better across different datasets or experimental conditions.

The next aspect addressed in our research pertains to data partitioning within the model validation strategy. Initially, we explore the influence of random data partitioning on ACP and CCP, representing more sophisticated CP variants. Despite the potential benefits of ACP and CCP in mitigating variability through aggregation or cross-validation, the reliance on random data partitioning remained a limiting factor, as it fails to capture the underlying patterns in the data. This underscored the need for a more robust data partitioning strategy to enhance the performance and generalizability of complex CP models. Molecular clustering emerged as a promising approach for optimizing data partitioning. Identifying an optimal clustering method within the NCI-60 panel can significantly enhance the efficacy of model validation strategies. Therefore, our study addressed the challenge of poor clustering of compounds, which compromises the validation of CP-based models, particularly when applied to test molecules dissimilar to those in the training set. Despite the significance of clustering, to the best of our knowledge, limited evidence exists regarding its application in CP frameworks for predicting molecular growth inhibition.

Our goal for this stage was to find a cost-efficient clustering method for model validation. We found UMAP-based clustering to be the best option. Our work also includes detailed results and code for replication, which can be used for various tasks beyond model validation.

5.1.4. On the CP model generalization to predict pGI_{50} values for dissimilar compounds

A fundamental aspect of our study centers around the ability of CP models to predict the pGI_{50} values of new compounds accurately. This ability is essential for their practical use in decision-making problems. If the model can accurately predict outcomes for unseen data, it could facilitate the identification of potential drug candidates for further evaluation in experimental studies. This generalization ensures that the CP-driven approach remains robust and applicable beyond the training dataset.

Using the LDMO-CV approach, we constructed and evaluated LDMO-CCP models on clusters of non-outlier molecules, leading to models that generally perform well on unseen molecules with characteristics different from those in the training set. This is crucial as real-world compound libraries contain many diverse molecules beyond those previously studied on the cell lines.

We found that prediction uncertainty rises as the requested confidence level increases, resulting in higher prediction errors for dissimilar molecules. By training CP models on non-outlier molecules and testing them on outlier molecules, we demonstrated their ability to handle distribution shifts between training and test data.

Our study highlights the utility of predictions with associated confidence intervals, as they improve decision-making by enabling the selection of novel molecules with smaller prediction errors for further exploration and experimental validation. Additionally, these CP models can predict the selectivity of dissimilar molecules across NCI-60 cell lines. We identified reliable and potent molecules by focusing on predictions with 80% confidence. Notably, NSC:747457 and NSC:7521 (Proscillaridin) emerged as promising candidates with predicted cell line selectivity, showcasing the potential of this approach for prioritizing molecules in drug discovery efforts.

5.1.5. On the integration of CP into decision-making processes to support prospective drug screening

Although several studies have demonstrated the advantages of using CP in combination with other virtual screening methods, individual prospective validations of CP models have yet to be reported [110, 150]. Prospective evaluations are essential to obtain a more accurate assessment of the model's performance in practical applications. Our approach goes beyond retrospective evaluation and includes prospective evaluation. This dual evaluation strategy validates our models' performance on historical

data and assesses their predictive capabilities on unseen data, which more accurately reflects real-world scenarios.

Our initial findings suggest that it may be possible to employ CP predictions in the absence of ground truth labels. We introduced the concept of "CP-enhanced" predictions and explored three potential criteria to establish them. Our preliminary results indicate that the criterion which involves selecting the top N molecules with a combination of high predicted pGI_{50} and a narrow interval from a pool of 3N molecules with the highest predicted pGI_{50} is the most promising method to define P-enhanced predictions. However, additional research is required to evaluate the applicability of this definition across various scenarios.

5.2. Future work

While our results showcase the effectiveness of the integrated ML and CP, some areas remain for improvement and refinement. In future work, we aim to enhance the predictive capabilities of our models by exploring alternative methods for feature generation. While the current approach combines MFPs and PCDs to represent molecules, exploring alternative methods could enhance feature extraction. Specifically, we propose to investigate the graph learning-based adaptive machine (GLAM) [83] method to generate features from molecular structures. By representing molecules as graphs and leveraging graph-based techniques, we can learn representations directly from molecule-cell activities and their chemical structures, capturing nonlinear relationships and patterns that other feature descriptors may miss.

Future work involves developing a hybrid framework integrating Bayesian principles into CP. Given ongoing efforts combining frequentist and Bayesian methods [91, 141, 152], exploring this fusion with CP, which follows frequentist principles, is a natural progression to enhance predictive robustness and uncertainty estimation.

In addition, future research endeavors will focus on elucidating the factors contributing to the varying activity of molecules across different cell lines. This may involve developing cell line-specific predictive models or incorporating additional contextual information into our existing modeling framework.

While the NCI-60 panel has served as a valuable resource for our current experiments, future research endeavors will benefit from exploring additional datasets to further validate and generalize our findings. For instance, datasets such as the Cancer Cell Line Encyclopedia (CCLE) or the Genomics of Drug Sensitivity in Cancer (GDSC) offer comprehensive collections of cell lines with associated genomic, pharmacological, and clinical annotations.

Bibliography

[1] Zaigham Abbas and Sakina Rehman. “An overview of cancer treatment modalities”. In: *Neoplasm* 1 (2018), pp. 139–157.

[2] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, et al. “A review of uncertainty quantification in deep learning: Techniques, applications and challenges”. In: *Information fusion* 76 (2021), pp. 243–297.

[3] Chayan Acharya, Andrew Coop, James E Polli, et al. “Recent advances in ligand-based drug design: relevance and utility of the conformationally sampled pharmacophore approach”. In: *Current computer-aided drug design* 7.1 (2011), pp. 10–22.

[4] Bilal Ahmed, Muhammad Imran Qadir, and Saba Ghafoor. “Malignant Melanoma: Skin Cancer- Diagnosis, Prevention, and Treatment”. In: *Critical Reviews™ in Eukaryotic Gene Expression* 30.4 (2020).

[5] B. Alberts, A. Johnson, J. Lewis, et al. “Cancer as a Microevolutionary Process”. In: *Molecular biology of the cell. 4th edn.* Ed. by Nigel Chaffey. Oxford University Press, 2023. URL: https://www.ncbi.nlm.nih.gov/books/NBK26891/.

[6] Mohamed Alloghani, Dhiya Al-Jumeily, Jamila Mustafina, et al. “A systematic review on supervised and unsupervised machine learning algorithms for data science”. In: *Supervised and unsupervised learning for data science* (2020), pp. 3–21.

[7] JAS Almeida, LMS Barbosa, AACC Pais, et al. “Improving hierarchical cluster analysis: A new method with outlier detection and automatic clustering”. In: *Chemometrics and Intelligent Laboratory Systems* 87.2 (2007), pp. 208–217.

[8] Jonathan Alvarsson, Staffan Arvidsson McShane, Ulf Norinder, et al. “Predicting with confidence: using conformal prediction in drug discovery”. In: *Journal of Pharmaceutical Sciences* 110.1 (2021), pp. 42–49.

[9] American Cancer Society. *All About Cancer (accessed: 26.01.2024)*. URL: https://www.cancer.org/cancer.html. (accessed: 26.01.2024).

[10] Xin An, Xi Chen, Daiyao Yi, et al. “Representation of molecules for drug response prediction”. In: *Briefings in Bioinformatics* 23.1 (2022), bbab393.

[11] Tyler WH Backman, Yiqun Cao, and Thomas Girke. “ChemMine tools: an online service for analyzing and clustering small molecules”. In: *Nucleic acids research* 39.suppl_2 (2011), W486–W491.

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

[13] Pedro J Ballester. *Machine learning for molecular modelling in drug design*. 2019.

[14] Pedro J Ballester, Rick Stevens, Benjamin Haibe-Kains, et al. *Artificial intelligence for drug response prediction in disease models*. 2022.

[15] Thomas A Ban. “The role of serendipity in drug discovery”. In: *Dialogues in clinical neuroscience* 8.3 (2006), pp. 335–344.

[16] Maximilian Beckers, Noé Sturm, Finton Sirockin, et al. "Prediction of Small-Molecule Developability Using Large-Scale In Silico ADMET Models". In: *Journal of medicinal chemistry* 66.20 (2023), pp. 14047–14060.

[17] Valentina Bellini, Marco Cascella, Franco Cutugno, et al. "Understanding basic principles of Artificial Intelligence: A practical guide for intensivists". In: *Acta Bio Medica: Atenei Parmensis* 93.5 (2022).

[18] Grazyna Biala, Ewa Kedzierska, Marta Kruk-Slomka, et al. "Research in the field of drug design and development". In: *Pharmaceuticals* 16.9 (2023), p. 1283.

[19] Vincent Blay, Bhairavi Tolani, Sunita P Ho, et al. "High-throughput screening: today's biochemical and cell-based approaches". In: *Drug Discovery Today* 25.10 (2020), pp. 1807–1821.

[20] Nicolas Bosc, Francis Atkinson, Eloy Felix, et al. "Large scale comparison of QSAR and conformal prediction methods and their applications in drug discovery". In: *Journal of cheminformatics* 11.1 (2019), pp. 1–16.

[21] Leo Breiman. "Random forests". In: *Machine learning* 45 (2001), pp. 5–32.

[22] Nathan Brown. "Chemoinformatics—an introduction for computer scientists". In: *ACM Computing Surveys (CSUR)* 41.2 (2009), pp. 1–38.

[23] Fred Bunz. *Principles of cancer genetics*. Vol. 1. Springer, 2008.

[24] Paulo EP Burke, Claudia B de L Campos, Luciano da F Costa, et al. "A biochemical network modeling of a whole-cell". In: *Scientific reports* 10.1 (2020), p. 13303.

[25] Darko Butina. "Unsupervised data base clustering based on daylight's fingerprint and Tanimoto similarity: A fast and automated way to cluster small and large data sets". In: *Journal of Chemical Information and Computer Sciences* 39.4 (1999), pp. 747–750.

[26] Cancer Research UK. *Types of cancer (accessed: 29.01.2024)*. URL: https://www.cancerresearchuk.org/about-cancer/what-is-cancer/how-cancer-starts/types-of-cancer. (accessed: 29.01.2024).

[27] Antonino Carbone. "Cancer classification at the crossroads". In: *Cancers* 12.4 (2020), p. 980.

[28] Lars Carlsson, Martin Eklund, and Ulf Norinder. "Aggregated conformal prediction". In: *Artificial Intelligence Applications and Innovations: AIAI 2014 Workshops: CoPA, MHDW, IIVC, and MT4BD, Rhodes, Greece, September 19-21, 2014. Proceedings 10*. Springer. 2014, pp. 231–240.

[29] Giulia Caron and Giuseppe Ermondi. *Molecular descriptors for polarity: the need for going beyond polar surface area*. 2016.

[30] Claudio N Cavasotto and Juan I Di Filippo. "Artificial intelligence in the early stages of drug discovery". In: *Archives of biochemistry and biophysics* 698 (2021), p. 108730.

[31] Paris Charilaou and Robert Battat. "Machine learning models and over-fitting considerations". In: *World Journal of Gastroenterology* 28.5 (2022), p. 605.

[32] Ming Chen, Jiangwen Zhang, Katia Sampieri, et al. "An aberrant SREBP-dependent lipogenic program promotes metastatic prostate cancer". In: *Nature genetics* 50.2 (2018), pp. 206–218.

[33] Wayne E Childers, Khaled M Elokely, and Magid Abou-Gharbia. "The resurrection of phenotypic drug discovery". In: *ACS Medicinal Chemistry Letters* 11.10 (2020), pp. 1820–1828.

[34] Geoffrey Cooper. "The Development and Causes of Cancer". In: *The Cell: A Molecular Approach. 2nd edition*. Ed. by Geoffrey Cooper. Sinauer Associates, 2000. URL: https://www.ncbi.nlm.nih.gov/books/NBK9839/.

[35] Isidro Cortes-Ciriano and Andreas Bender. "Reliable prediction errors for deep neural networks using test-time dropout". In: *Journal of chemical information and modeling* 59.7 (2019), pp. 3330–3339.

[36] Isidro Cortés-Ciriano and Andreas Bender. "Concepts and applications of conformal prediction in computational drug discovery". In: *arXiv preprint arXiv:1908.03569* (2019).

[37] Isidro Cortés-Ciriano, Nicholas C Firth, Andreas Bender, et al. "Discovering highly potent molecules from an initial set of inactives using iterative screening". In: *Journal of chemical information and modeling* 58.9 (2018), pp. 2000–2014.

[38] Isidro Cortés-Ciriano, Ctibor Škuta, Andreas Bender, et al. "QSAR-derived affinity fingerprints (part 2): modeling performance for potency prediction". In: *Journal of Cheminformatics* 12.1 (2020), p. 41.

[39] Isidro Cortés-Ciriano, Gerard JP van Westen, Guillaume Bouvier, et al. "Improved large-scale prediction of growth inhibition patterns using the NCI60 cancer cell line panel". In: *Bioinformatics* 32.1 (2016), pp. 85–95.

[40] Francisco Sahli Costabal, Kristen Matsuno, Jiang Yao, et al. "Machine learning in drug development: Characterizing the effect of 30 drugs on the QT interval using Gaussian process regression, sensitivity analysis, and uncertainty quantification". In: *Computer Methods in Applied Mechanics and Engineering* 348 (2019), pp. 313–333.

[41] Raúl Cruz-Barbosa, Saiveth Hernández-Hernández, and Luis Enrique Sucar. "Mass segmentation of mammograms using Markov models associated with constrained clustering". In: *Medical & Biological Engineering & Computing* 58 (2020), pp. 2475–2495.

[42] Bhanuranjan Das, Anurag TK Baidya, Alen T Mathew, et al. "Structural modification aimed for improving solubility of lead compounds in early phase drug discovery". In: *Bioorganic & Medicinal Chemistry* 56 (2022), p. 116614.

[43] Ashwani K. Dhingra, Bhawna Chopra, Sakshi Bhardwaj, et al. "Chapter 8 - Virtual screening". In: *Computational Approaches in Drug Discovery, Development and Systems Pharmacology*. Ed. by Rupesh Kumar Gautam, Mohammad Amjad Kamal, and Pooja Mittal. Academic Press, 2023, pp. 223–236. ISBN: 978-0-323-99137-7. DOI: https://doi.org/10.1016/B978-0-323-99137-7.00007-1. URL: https://www.sciencedirect.com/science/article/pii/B9780323991377000071.

[44] Jay F Dorsey, Ryan D Salinas, Mai Dang, et al. "Cancer of the central nervous system". In: *Abeloff's clinical oncology*. Elsevier, 2020, pp. 906–967.

[45] Irini Doytchinova. "Drug design—Past, present, future". In: *Molecules* 27.5 (2022), p. 1496.

[46] DrugBank. *Glossary | DrugBank Help Center (accessed: 01.08.2023)*. URL: https://dev.drugbank.com/guides/terms. (accessed: 01.04.2022).

[47] Jintong Du, Jing Guo, Dongwei Kang, et al. "New techniques and strategies in drug discovery". In: *Chinese Chemical Letters* 31.7 (2020), pp. 1695–1708.

[48] Martin Eklund, Ulf Norinder, Scott Boyer, et al. "The application of conformal prediction to the drug discovery process". In: *Annals of Mathematics and Artificial Intelligence* 74 (2015), pp. 117–132.

[49] Leif W Ellisen and Daniel A Haber. "Basic Principles of Cancer Genetics". In: *Principles of Clinical Cancer Genetics: A Handbook from the Massachusetts General Hospital* (2010), pp. 1–22.

[50] K Esfahani, L Roudaia, NA Buhlaiga, et al. "A review of cancer immunotherapy: from the past, to the present, to the future". In: *Current Oncology* 27.s2 (2020), pp. 87–97.

[51] Xiaomin Fang, Lihang Liu, Jieqiong Lei, et al. "Geometry-enhanced molecular representation learning for property prediction". In: *Nature Machine Intelligence* 4.2 (2022), pp. 127–134.

[52] Jean-Loup Faulon, Donald P Visco, and Ramdas S Pophale. "The signature molecular descriptor. 1. Using extended valence sequences in QSAR and QSPR studies". In: *Journal of chemical information and computer sciences* 43.3 (2003), pp. 707–720.

[53] J Ferlay, M Ervik, F Lam, et al. *Global Cancer Observatory: Cancer Today. Lyon, France: International Agency for Research on Cancer. From: https://gco.iarc.who.int/today, (accessed: 25.01.2024)*. 2024.

[54] Isabella Fornacon-Wood, Hitesh Mistry, Corinne Johnson-Hart, et al. "Understanding the differences between Bayesian and frequentist statistics". In: *International journal of radiation oncology, biology, physics* 112.5 (2022), pp. 1076–1082.

[55] Jerome H Friedman. "Greedy function approximation: a gradient boosting machine". In: *Annals of statistics* (2001), pp. 1189–1232.

[56] Leah Frye, Sathesh Bhat, Karen Akinsanya, et al. "From computer-aided drug discovery to computer-driven drug discovery". In: *Drug Discovery Today: Technologies* 39 (2021), pp. 111–117.

[57] Sonny Gan, David A Cosgrove, Eleanor J Gardiner, et al. "Investigation of the use of spectral clustering for the analysis of molecular data". In: *Journal of chemical information and modeling* 54.12 (2014), pp. 3302–3319.

[58] Ghita Ghislat, Taufiq Rahman, and Pedro J Ballester. "Recent progress on the prospective application of machine learning to structure-based virtual screening". In: *Current opinion in chemical biology* 65 (2021), pp. 28–34.

[59] Pablo Gómez-Sacristán, Saw Simeon, Viet-Khoa Tran-Nguyen, et al. "Inactive-enriched machine-learning models exploiting patent data improve structure-based virtual screening for PDL1 dimerizers". In: *Journal of Advanced Research* (2024).

[60] Rajarshi Guha and Egon Willighagen. "A survey of quantitative descriptions of molecular structure". In: *Current topics in medicinal chemistry* 12.18 (2012), pp. 1946–1956.

[61] Emily Hargrave-Thomas, Bo Yu, and Jóhannes Reynisson. "Serendipity in anticancer drug discovery". In: *World journal of clinical oncology* 3.1 (2012), p. 1.

[62] KJ Harrington and P Nenclares. "The biology of cancer". In: *Medicine* 51.1 (2022), pp. 1–6.

[63] Saiveth Hernández-Hernández and Pedro J Ballester. "On the Best Way to Cluster NCI-60 Molecules". In: *Biomolecules* 13.3 (2023), p. 498.

[64] Saiveth Hernández-Hernández, Qianrong Guo, and Pedro J Ballester. "Conformal prediction of molecule-induced cancer cell growth inhibition challenged by strong distribution shifts". In: *Preprint from bioRxiv* (2024), pp. 1–16.

[65] Saiveth Hernández-Hernández, Sachin Vishwakarma, and Pedro Ballester. "Conformal prediction of small-molecule drug resistance in cancer cell lines". In: *Conformal and Probabilistic Prediction with Applications*. PMLR. 2022, pp. 92–108.

[66] Seyed Aghil Hooshmand, Sadegh Azimzadeh Jamalkandi, Seyed Mehdi Alavi, et al. "Distinguishing drug/non-drug-like small molecules in drug discovery using deep belief network". In: *Molecular Diversity* 25 (2021), pp. 827–838.

[67] James P Hughes, Stephen Rees, S Barrett Kalindjian, et al. "Principles of early drug discovery". In: *British journal of pharmacology* 162.6 (2011), pp. 1239–1249.

[68] Eyke Hüllermeier and Willem Waegeman. "Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods". In: *Machine learning* 110.3 (2021), pp. 457–506.

[69] Giulio Imbalzano, Andrea Anelli, Daniele Giofré, et al. "Automatic selection of atomic fingerprints and reference configurations for machine-learning potentials". In: *The Journal of chemical physics* 148.24 (2018).

[70] Peter Imming, Christian Sinning, and Achim Meyer. "Drugs, their targets and the nature and number of drug targets". In: *Nature reviews Drug discovery* 5.10 (2006), pp. 821–834.

[71] Andrada E Ivanescu, Peng Li, Brandon George, et al. "The importance of prediction model validation and assessment in obesity and nutrition research". In: *International journal of obesity* 40.6 (2016), pp. 887–894.

[72] Yankang Jing, Yuemin Bian, Ziheng Hu, et al. "Deep learning for drug design: an artificial intelligence paradigm for drug discovery in the big data era". In: *The AAPS journal* 20 (2018), pp. 1–10.

[73] Supratik Kar and Jerzy Leszczynski. "Open access in silico tools to predict the ADMET profiling of drug candidates". In: *Expert opinion on drug discovery* 15.12 (2020), pp. 1473–1487.

[74] Md Rezaul Karim, Oya Beyan, Achille Zappa, et al. "Deep learning-based clustering approaches for bioinformatics". In: *Briefings in bioinformatics* 22.1 (2021), pp. 393–415.

[75] Renata R Kazakova and Patrick Masson. "Quantitative Measurements of Pharmacological and Toxicological Activity of Molecules". In: *Chemistry* 4.4 (2022), pp. 1466–1474.

[76] Timothy J Key, Naomi E Allen, Elizabeth A Spencer, et al. "The effect of diet on risk of cancer". In: *The Lancet* 360.9336 (2002), pp. 861–868.

[77] Nathalie Kliemann, Fernanda Rauber, Renata Bertazzi Levy, et al. "Food processing and cancer risk in Europe: results from the prospective EPIC cohort study". In: *The Lancet Planetary Health* 7.3 (2023), e219–e232.

[78] Anton Kocheturov, Panos M Pardalos, and Athanasia Karakitsiou. "Massive datasets and machine learning for computational biomedicine: trends and challenges". In: *Annals of Operations Research* 276 (2019), pp. 5–34.

[79] Stanley E Lazic and Dominic P Williams. "Quantifying sources of uncertainty in drug discovery predictions with probabilistic models". In: *Artificial Intelligence in the Life Sciences* 1 (2021), p. 100004.

[80] Matthew D Lee, Mohammed Elsayed, Sumit Chopra, et al. "A No-Math Primer on the Principles of Machine Learning for Radiologists". In: *Seminars in Ultrasound, CT and MRI*. Vol. 43. 2. Elsevier. 2022, pp. 133–141.

[81] Paul Leeson. "Chemical beauty contest". In: *Nature* 481.7382 (2012), pp. 455–456.

[82] Kerstin Lenhof, Lea Eckhart, Lisa-Marie Rolli, et al. "Reliable Anti-Cancer Drug Sensitivity Prediction and Prioritization". In: *Preprint from Research Square* (2023), pp. 1–33.

[83] Yuquan Li, Chang-Yu Hsieh, Ruiqiang Lu, et al. "GLAM: An adaptive graph learning method for automated molecular interactions and properties predictions". In: (2022).

[84] Henrik Linusson, Ulf Norinder, Henrik Boström, et al. "On the calibration of aggregated conformal predictors". In: *Conformal and probabilistic prediction and applications*. PMLR. 2017, pp. 154–173.

[85] David J Livingstone. "The characterization of chemical structures using molecular properties. A survey". In: *Journal of chemical information and computer sciences* 40.2 (2000), pp. 195–209.

[86] Yu-Chen Lo, Stefano E Rensi, Wen Torng, et al. "Machine learning in chemoinformatics and drug discovery". In: *Drug discovery today* 23.8 (2018), pp. 1538–1546.

[87] Ricardo Macarron, Martyn N Banks, Dejan Bojanic, et al. "Impact of high-throughput screening in biomedical research". In: *Nature reviews Drug discovery* 10.3 (2011), pp. 188–195.

[88] Andrey Malinin and Mark Gales. "Predictive uncertainty estimation via prior networks". In: *Advances in neural information processing systems* 31 (2018).

[89] Leland McInnes, John Healy, and James Melville. "Umap: Uniform manifold approximation and projection for dimension reduction". In: *arXiv preprint arXiv:1802.03426* (2018).

[90] Inc Merriam-Webster. *Webster's ninth new collegiate dictionary*. Vol. 10. Merriam-Webster, 1990.

[91] Lewis H Mervin, Simon Johansson, Elizaveta Semenova, et al. "Uncertainty quantification in drug design". In: *Drug discovery today* 26.2 (2021), pp. 474–489.

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

[93] Hrishikesh Mishra, Nitya Singh, Tapobrata Lahiri, et al. "A comparative study on the molecular descriptors for predicting drug-likeness of small molecules". In: *Bioinformation* 3.9 (2009), p. 384.

[94] Thomas M Moerland, Joost Broekens, Aske Plaat, et al. "Model-based reinforcement learning: A survey". In: *Foundations and Trends® in Machine Learning* 16.1 (2023), pp. 1–118.

[95] John G Moffat, Joachim Rudolph, and David Bailey. "Phenotypic screening in cancer drug discovery—past, present and future". In: *Nature reviews Drug discovery* 13.8 (2014), pp. 588–602.

[96] Richard C Mohs and Nigel H Greig. "Drug discovery and development: Role of basic biological research". In: *Alzheimer's & Dementia: Translational Research & Clinical Interventions* 3.4 (2017), pp. 651–657.

[97] Vladimir Nasteski. "An overview of the supervised machine learning methods". In: *Horizons. b* 4 (2017), pp. 51–62.

[98] National Cancer Institute. *About Cancer (accessed: 25.01.2024)*. URL: https://www.cancer.gov/about-cancer. (accessed: 25.01.2024).

[99] National Center for Advancing Translational Sciences. *Glossary (accessed: 29.01.2024)*. URL: https://toolkit.ncats.nih.gov/glossary/drug/. (accessed: 29.01.2024).

[100] 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. (accessed: 01.04.2022).

[101] Frank Nielsen and Frank Nielsen. "Hierarchical clustering". In: *Introduction to HPC with MPI for Data Science* (2016), pp. 195–211.

[102] NonConformist. *NonConformist v2.1.0 python package (accessed: 01.04.2022)*. URL: http://donlnz.github.io/nonconformist/. (accessed: 01.04.2022).

[103] Ulf Norinder, Lars Carlsson, Scott Boyer, et al. "Introducing conformal prediction in predictive modeling. A transparent and flexible alternative to applicability domain de-

termination". In: *Journal of chemical information and modeling* 54.6 (2014), pp. 1596–1603.

[104] Rebecca Nugent and Marina Meila. "An overview of clustering applied to molecular biology". In: *Statistical methods in molecular biology* (2010), pp. 369–404.

[105] Claire O'Connor and Jill Adams. "Cancer as a Microevolutionary Process". In: *Essentials of Cell Biology*. Ed. by Zhaohua Tang and Ivor Hickey. Cambridge, 2010. URL: https://www.nature.com/scitable/topicpage/cell-division-and-cancer-14046590/.

[106] Jake S O'Donnell, Michele WL Teng, and Mark J Smyth. "Cancer immunoediting and resistance to T cell-based immunotherapy". In: *Nature reviews Clinical oncology* 16.3 (2019), pp. 151–167.

[107] Olga Obrezanova, Gábor Csányi, Joelle MR Gola, et al. "Gaussian processes: a method for automatic QSAR modeling of ADME properties". In: *Journal of chemical information and modeling* 47.5 (2007), pp. 1847–1857.

[108] Ganesh Ravindra Padalkar, Shivani Dinkar Patil, Mukta Mallikarjun Hegadi, et al. "Drug discovery using generative adversarial network with reinforcement learning". In: *2021 International Conference on Computer Communication and Informatics (ICCCI)*. IEEE. 2021, pp. 1–3.

[109] Viswanadha Vijaya Padma. "An overview of targeted cancer therapy". In: *BioMedicine* 5 (2015), pp. 1–6.

[110] Conor Parks, Zied Gaieb, and Rommie E Amaro. "An analysis of proteochemometric and conformal prediction machine learning protein-ligand binding affinity models". In: *Frontiers in molecular biosciences* 7 (2020), p. 93.

[111] Debleena Paul, Gaurav Sanap, Snehal Shenoy, et al. "Artificial intelligence in drug discovery and development". In: *Drug discovery today* 26.1 (2021), p. 80.

[112] Antonio Peón, Stefan Naulaerts, and Pedro J Ballester. "Predicting the reliability of drug-target interaction predictions with maximum coverage of target space". In: *Scientific reports* 7.1 (2017), p. 3820.

[113] Chayanit Piyawajanusorn, Linh C Nguyen, Ghita Ghislat, et al. "A gentle introduction to understanding preclinical data for cancer pharmaco-omic modeling". In: *Briefings in Bioinformatics* 22.6 (2021), bbab312.

[114] Jaroslaw Polanski. "Unsupervised Learning in Drug Design from Self-Organization to Deep Chemistry". In: *International Journal of Molecular Sciences* 23.5 (2022), p. 2797.

[115] Denis N Prada Gori, Manuel A Llanos, Carolina L Bellera, et al. "iRaPCA and SOMoC: Development and validation of web applications for new approaches for the clustering of small molecules". In: *Journal of Chemical Information and Modeling* 62.12 (2022), pp. 2987–2998.

[116] S Prasanna and RJ Doerksen. "Topological polar surface area: a useful descriptor in 2D-QSAR". In: *Current medicinal chemistry* 16.1 (2009), pp. 21–41.

[117] Anita Rácz, Dávid Bajusz, and Károly Héberger. "Effect of dataset size and train/test split ratios in QSAR/QSPR multiclass classification". In: *Molecules* 26.4 (2021), p. 1111.

[118] RDKit. *Open-source cheminformatics*. URL: https://www.rdkit.org. (accessed: 01.01.2024).

[119] William C Reinhold, Margot Sunshine, Hongfang Liu, et al. "CellMiner: a web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the NCI-60 cell line set". In: *Cancer research* 72.14 (2012), pp. 3499–3511.

[120] Vinaitheerthan Renganathan. “Overview of frequentist and bayesian approach to survival analysis”. In: *Applied Medical Informatics* 38.1 (2016), pp. 25–38.

[121] Timothy J Ritchie and Simon JF Macdonald. “The impact of aromatic ring count on compound developability–are too many aromatic rings a liability in drug design?” In: *Drug discovery today* 14.21-22 (2009), pp. 1011–1020.

[122] Timothy J Ritchie, Simon JF Macdonald, Robert J Young, et al. “The impact of aromatic ring count on compound developability: further insights by examining carbo-and hetero-aromatic and-aliphatic ring types”. In: *Drug Discovery Today* 16.3-4 (2011), pp. 164–171.

[123] David Rogers and Mathew Hahn. “Extended-connectivity fingerprints”. In: *Journal of chemical information and modeling* 50.5 (2010), pp. 742–754.

[124] Sara E Rosenbaum. *Basic pharmacokinetics and pharmacodynamics: An integrated textbook and computer simulations*. John Wiley & Sons, 2016.

[125] Pol Rovira and Jürgen Rehm. “Estimation of cancers caused by light to moderate alcohol consumption in the European Union”. In: *European Journal of Public Health* 31.3 (2021), pp. 591–596.

[126] Christopher J Roy and William L Oberkampf. “A comprehensive framework for verification, validation, and uncertainty quantification in scientific computing”. In: *Computer methods in applied mechanics and engineering* 200.25-28 (2011), pp. 2131–2144.

[127] Harriet Rumgay, Kevin Shield, Hadrien Charvat, et al. “Global burden of cancer in 2020 attributable to alcohol consumption: a population-based study”. In: *The Lancet Oncology* 22.8 (2021), pp. 1071–1080.

[128] Francisco J Samaniego. *A comparison of the Bayesian and frequentist approaches to estimation*. Vol. 24. Springer, 2010.

[129] Gisbert Schneider. “Prediction of drug-like properties”. In: *Madame curie bioscience database [Internet]*. Landes Bioscience, 2013.

[130] John R Searle. “Minds, brains, and programs”. In: *Behavioral and brain sciences* 3.3 (1980), pp. 417–424.

[131] Vidushi Sharma, Sharad Wakode, and Hirdesh Kumar. “Structure-and ligand-based drug design: Concepts, approaches, and challenges”. In: *Chemoinformatics and bioinformatics in the pharmaceutical sciences* (2021), pp. 27–53.

[132] Robert P Sheridan, Wei Min Wang, Andy Liaw, et al. “Extreme gradient boosting as a method for quantitative structure-activity relationships”. In: *Journal of chemical information and modeling* 56.12 (2016), pp. 2353–2360.

[133] Robert H Shoemaker. “The NCI60 human tumour cell line anticancer drug screen”. In: *Nature Reviews Cancer* 6.10 (2006), pp. 813–823.

[134] Md Nazmul Kabir Sikder and Feras A Batarseh. “Outlier detection using AI: a survey”. In: *AI Assurance* (2023), pp. 231–291.

[135] Natesh Singh, Philippe Vayer, Shivalika Tanwar, et al. “Drug discovery and development: introduction to the general public and patient groups”. In: *Frontiers in Drug Discovery* 3 (2023), p. 1201419.

[136] Satendra Singh, Dev Bukhsh Singh, Budhayash Gautam, et al. “Pharmacokinetics and pharmacodynamics analysis of drug candidates”. In: *Bioinformatics*. Elsevier, 2022, pp. 305–316.

[137] Vikash Singh, Michael Pencina, Andrew J Einstein, et al. “Impact of train/test sample regimen on performance estimate stability of machine learning in cardiovascular imaging”. In: *Scientific reports* 11.1 (2021), pp. 1–8.

[138] Sandeep Sinha and Divya Vohora. "Drug discovery and development: An overview". In: *Pharmaceutical medicine and translational clinical research* (2018), pp. 19–32.

[139] Vishal B Siramshetty, Pranav Shah, Edward Kerns, et al. "Retrospective assessment of rat liver microsomal stability at NCATS: data and QSAR models". In: *Scientific Reports* 10.1 (2020), p. 20713.

[140] C Škuta, Isidro Cortés-Ciriano, Wim Dehaen, et al. "QSAR-derived affinity fingerprints (part 1): fingerprint construction and modeling performance for similarity searching, bioactivity classification and scaffold hopping". In: *Journal of Cheminformatics* 12.1 (2020), pp. 1–16.

[141] Nigel Stallard, John Whitehead, and Simon Cleall. "Decision-making in a phase II clinical trial: a new approach combining Bayesian and frequentist concepts". In: *Pharmaceutical Statistics: The Journal of Applied Statistics in the Pharmaceutical Industry* 4.2 (2005), pp. 119–128.

[142] Lathrop T Stedman. "Stedman's medical dictionary25th edBaltimore". In: *MD: Lippincott Williams & Wilkins* (2000).

[143] Tomaž Stepišnik, Blaž Škrlj, Jörg Wicker, et al. "A comprehensive comparison of molecular feature representations for use in predictive modeling". In: *Computers in Biology and Medicine* 130 (2021), p. 104197.

[144] Christine Stewart, Christine Ralyea, and Suzy Lockwood. "Ovarian cancer: an integrated review". In: *Seminars in oncology nursing*. Vol. 35. 2. Elsevier. 2019, pp. 151–156.

[145] Jeffrey Strovel, Sitta Sittampalam, Nathan P Coussens, et al. "Early drug discovery and development guidelines: for academic researchers, collaborators, and start-up companies". In: *Assay guidance manual [Internet]* (2016).

[146] Jiangming Sun, Lars Carlsson, Ernst Ahlberg, et al. "Applying mondrian cross-conformal prediction to estimate prediction confidence on large imbalanced bioactivity data sets". In: *Journal of chemical information and modeling* 57.7 (2017), pp. 1591–1598.

[147] Fredrik Svensson, Natalia Aniceto, Ulf Norinder, et al. "Conformal regression for quantitative structure–activity relationship modeling—quantifying prediction uncertainty". In: *Journal of Chemical Information and Modeling* 58.5 (2018), pp. 1132–1140.

[148] Vladimir Svetnik, Andy Liaw, Christopher Tong, et al. "Random forest: a classification and regression tool for compound classification and QSAR modeling". In: *Journal of chemical information and computer sciences* 43.6 (2003), pp. 1947–1958.

[149] Viet-Khoa Tran-Nguyen, Muhammad Junaid, Saw Simeon, et al. "A practical guide to machine-learning scoring for structure-based virtual screening". In: *Nature Protocols* 18.11 (2023), pp. 3460–3511.

[150] Alzbeta Tuerkova, Brandon J Bongers, Ulf Norinder, et al. "Identifying novel inhibitors for hepatic organic anion transporting polypeptides by machine learning-based virtual screening". In: *Journal of Chemical Information and Modeling* 62.24 (2022), pp. 6323–6335.

[151] Vyshnavi Vennelakanti, Helena W Qi, Rimsha Mehmood, et al. "When are two hydrogen bonds better than one? Accurate first-principles models explain the balance of hydrogen bond donors and acceptors found in proteins". In: *Chemical Science* 12.3 (2021), pp. 1147–1162.

[152] Steffen Ventz, Giovanni Parmigiani, and Lorenzo Trippa. "Combining Bayesian experimental designs and frequentist data analyses: motivations and examples". In: *Applied Stochastic Models in Business and Industry* 33.3 (2017), pp. 302–313.

[153] Victoria Volodina and Peter Challenor. "The importance of uncertainty quantification in model reproducibility". In: *Philosophical Transactions of the Royal Society A* 379.2197 (2021), p. 20200071.

[154] Vladimir Vovk. "Cross-conformal predictors". In: *Annals of Mathematics and Artificial Intelligence* 74 (2015), pp. 9–28.

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

[156] Dingyan Wang, Jie Yu, Lifan Chen, et al. "A hybrid framework for improving uncertainty quantification in deep learning-based QSAR regression modeling". In: *Journal of cheminformatics* 13.1 (2021), pp. 1–17.

[157] Lin Wang, Minghu Jiang, Yinghua Lu, et al. "A comparative study of clustering methods for molecular data". In: *International Journal of Neural Systems* 17.06 (2007), pp. 447–458.

[158] Yuyang Wang, Jianren Wang, Zhonglin Cao, et al. "Molecular contrastive learning of representations via graph neural networks". In: *Nature Machine Intelligence* 4.3 (2022), pp. 279–287.

[159] Mary C White, Dawn M Holman, Jennifer E Boehm, et al. "Age and cancer risk: a potentially modifiable relationship". In: *American journal of preventive medicine* 46.3 (2014), S7–S15.

[160] Todd P Whitehead, Catherine Metayer, Joseph L Wiemels, et al. "Childhood leukemia and primary prevention". In: *Current problems in pediatric and adolescent health care* 46.10 (2016), pp. 317–352.

[161] Magdalena Wiercioch and Johannes Kirchmair. "DNN-PP: A novel Deep Neural Network approach and its applicability in drug-related property prediction". In: *Expert Systems with Applications* 213 (2023), p. 119055.

[162] Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, et al. "MoleculeNet: a benchmark for molecular machine learning". In: *Chemical science* 9.2 (2018), pp. 513–530.

[163] Jingbo Yang, Yiyang Cai, Kairui Zhao, et al. "Concepts and applications of chemical fingerprint for hit and lead screening". In: *Drug Discovery Today* 27.11 (2022), p. 103356.

[164] Jie Yu, Dingyan Wang, and Mingyue Zheng. "Uncertainty quantification: Can we trust artificial intelligence in drug discovery?" In: *Iscience* (2022).

[165] Jin Zhang, Ulf Norinder, and Fredrik Svensson. "Deep learning-based conformal prediction of toxicity". In: *Journal of chemical information and modeling* 61.6 (2021), pp. 2648–2657.

[166] Juan Zhang, Junping Yin, and Ruili Wang. "Basic framework and main methods of uncertainty quantification". In: *Mathematical Problems in Engineering* 2020 (2020), pp. 1–18.

[167] Li Zhou and Margarita Sordo. "Expert systems in medicine". In: *Artificial intelligence in medicine*. Elsevier, 2021, pp. 75–100.

[168] Zhenpeng Zhou, Steven Kearnes, Li Li, et al. "Optimization of molecules via deep reinforcement learning". In: *Scientific reports* 9.1 (2019), p. 10752.

ANNEXES

A. COPA paper

Saiveth Hernández-Hernández, Sachin Vishwakarma, and Pedro Ballester. "Conformal prediction of small-molecule drug resistance in cancer cell lines". In: *Conformal and Probabilistic Prediction with Applications*. PMLR. 2022, pp. 92–108

Proceedings of Machine Learning Research 179:1–17, 2022 Conformal and Probabilistic Prediction and Applications

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

© 2022 S. Hernández-Hernández, S. Vishwakarma & P. Ballester.

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.

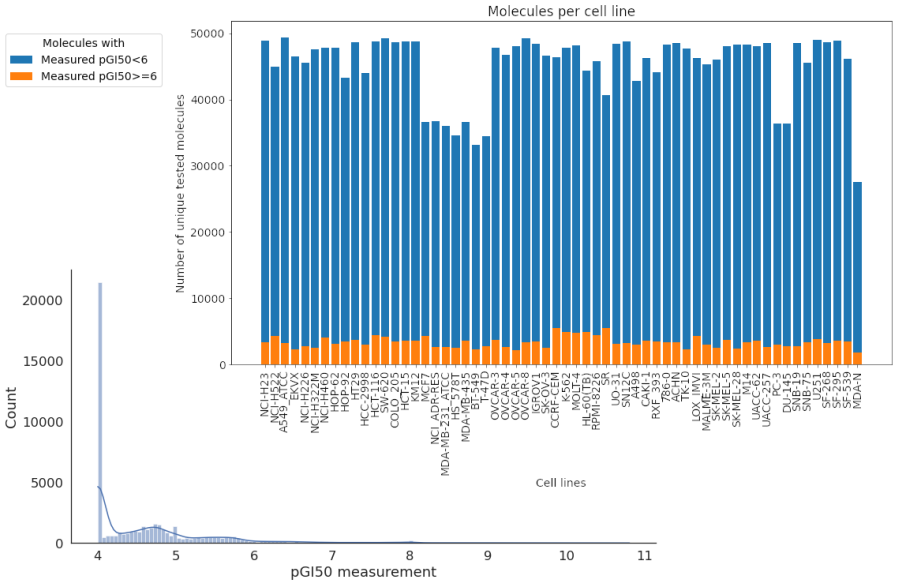


Figure 1: **Each cell line has abundant data, although potent molecules are rather scarce**. Distribution of pGI_{50} measurements in the 50,846 unique NSC IDs (bottom). Distribution of the number of unique molecules tested per cell line (top). The most potent molecules ($pGI_{50} \geq 6$) for each cell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

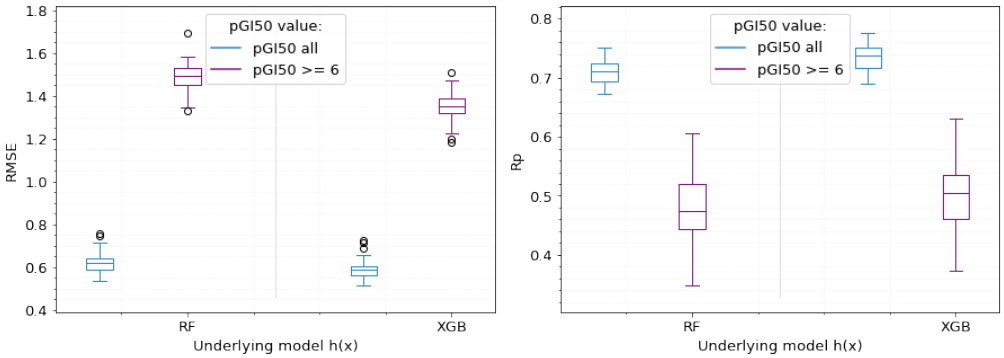
The performance of the regression models was evaluated for the 60 test sets. Each CP model in Table 1 was evaluated using the validity and efficiency. The underlying model $h(x)$ and the error model $g(x)$ are standard supervised learning models (Subsection 2.3), for which we compute the root mean square 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.

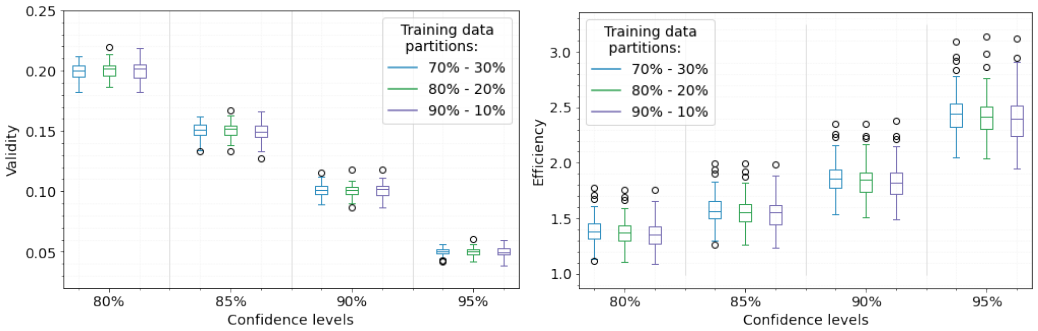
3.2. Conformal prediction

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).

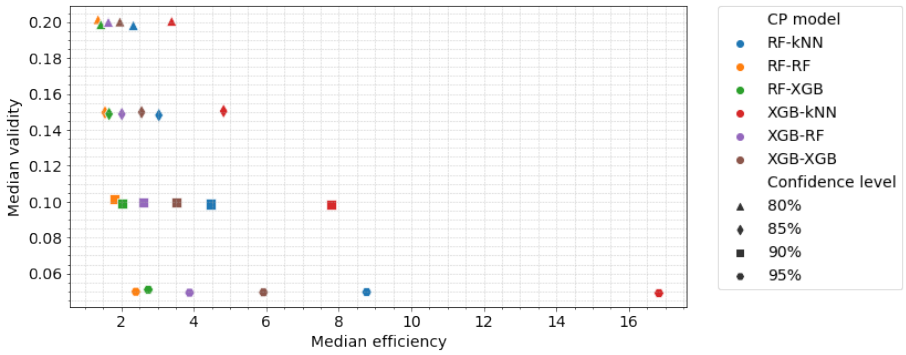
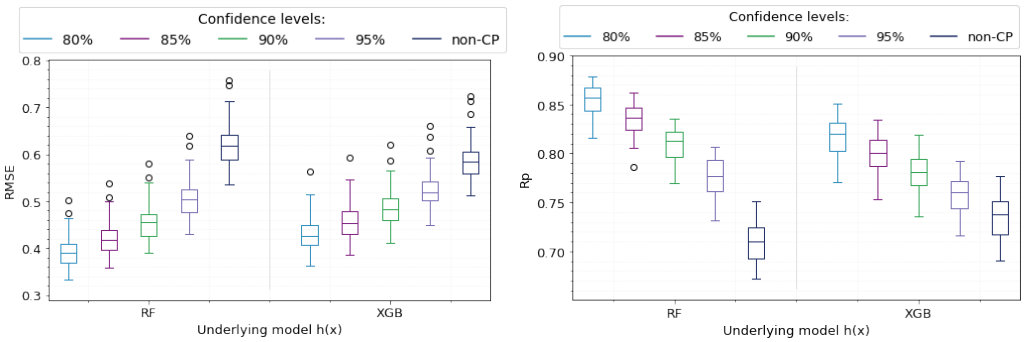


Figure 4: **CP models have substantially different efficiency within a given confidence level.** Median efficiency vs median validity across 60 test sets (one per cell line), at four confidence levels. Color code refers to the six CP models ($h(x)$-$g(x)$) employed. Markers refers to the requested confidence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



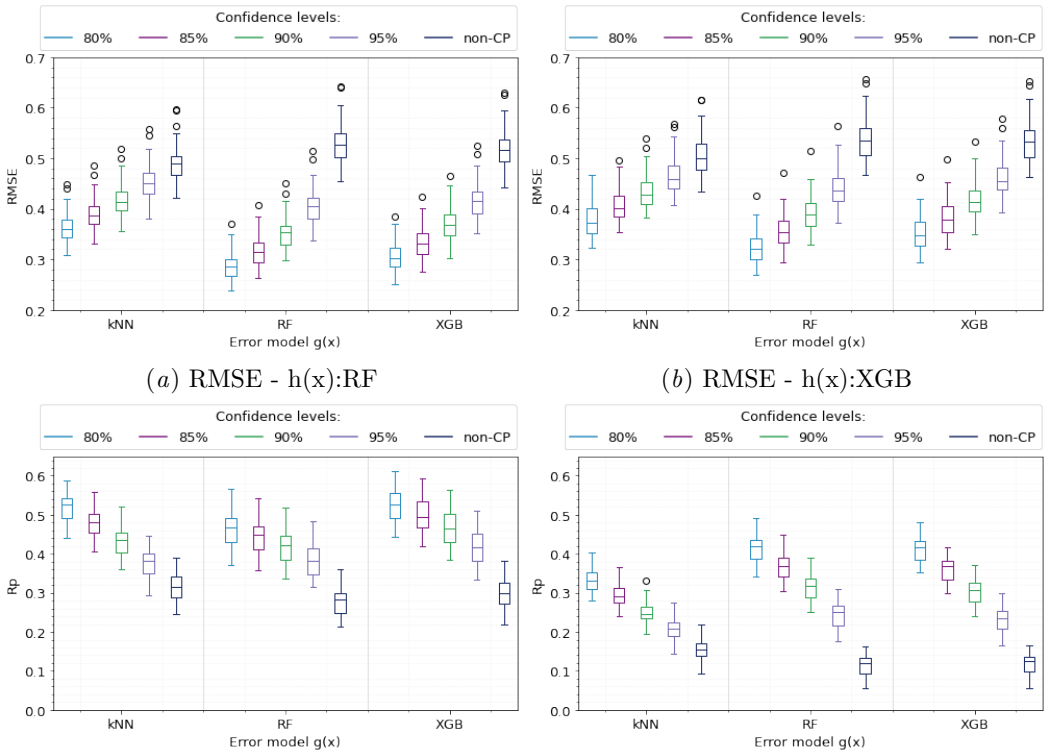
(*a*) RMSE (*b*) Rp

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

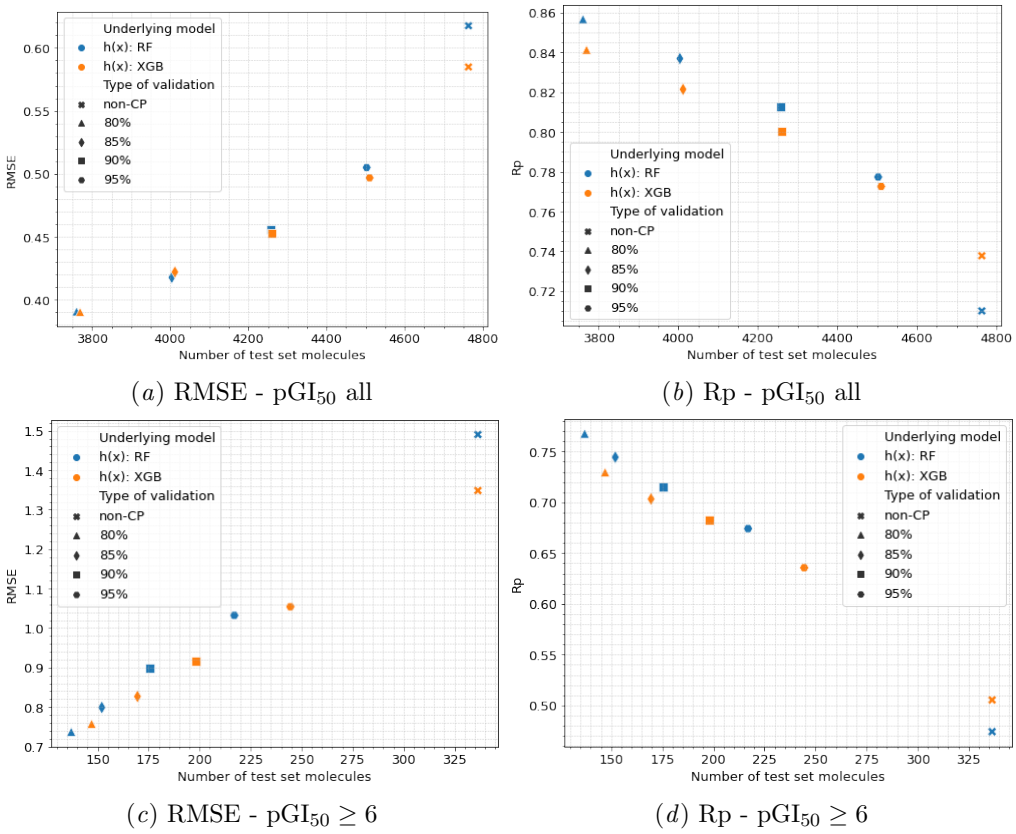
(*c*) Rp - h(x):RF

(*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.

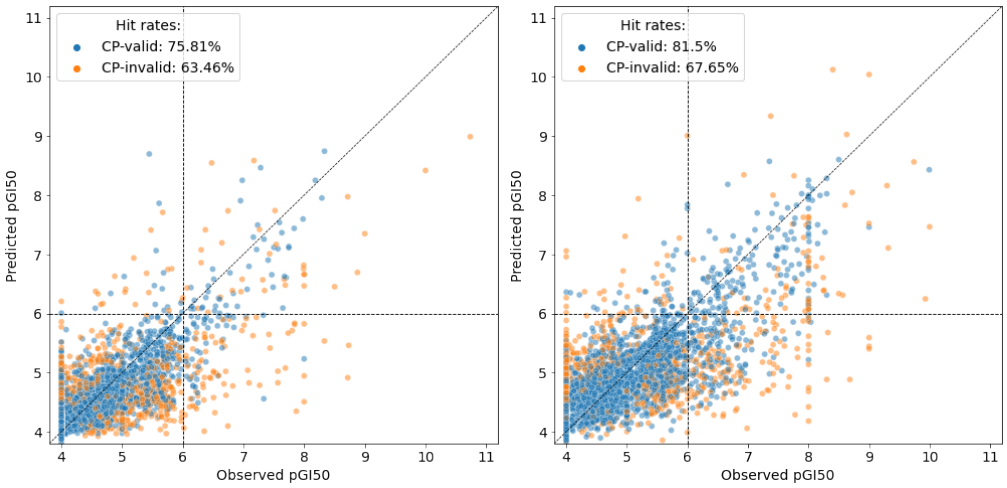
Table 2: Best and worst predicted cell line for each underlying model, for predictions that are non-CP and CP-valid. The confidence level in the CP is 80%. The total column represents the number of test set molecules used to calculate these performance metrics.

Model h(x)	Cell line	All test set					CP-valid test set		
		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).

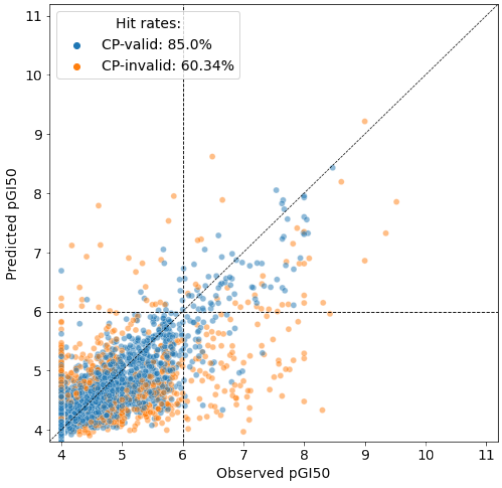
4. Conclusions

- 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

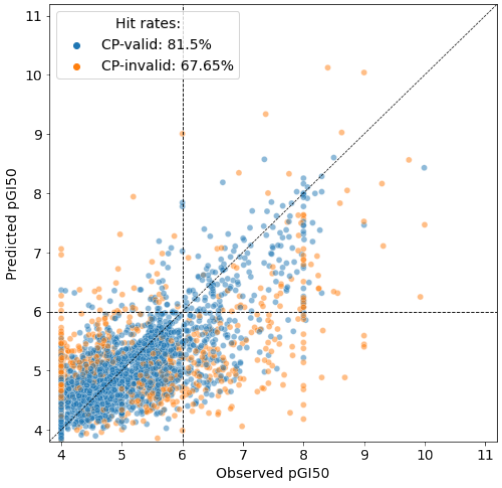


(*a*) Best with RF-RF: OVCAR-5

(*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.

B. Clustering paper

Saiveth Hernández-Hernández and Pedro J Ballester. "On the Best Way to Cluster NCI-60 Molecules". In: *Biomolecules* 13.3 (2023), p. 498, https://doi.org/10.3390/biom13030498

biomolecules

MDPI

Article

On the Best Way to Cluster NCI-60 Molecules

Saiveth Hernández-Hernández [1] and Pedro J. Ballester [2,*]

[1] Cancer Research Center of Marseille (INSERM U1068, Institut Paoli-Calmettes, Aix-Marseille Université UM105, CNRS UMR7258), 13009 Marseille, France
[2] Department of Bioengineering, Imperial College London, London SW7 2AZ, UK
* Correspondence: p.ballester@imperial.ac.uk

Abstract: Machine learning-based models have been widely used in the early drug-design pipeline. To validate these models, cross-validation strategies have been employed, including those using clustering of molecules in terms of their chemical structures. However, the poor clustering of compounds will compromise such validation, especially on test molecules dissimilar to those in the training set. This study aims at finding the best way to cluster the molecules screened by the National Cancer Institute (NCI)-60 project by comparing hierarchical, Taylor–Butina, and uniform manifold approximation and projection (UMAP) clustering methods. The best-performing algorithm can then be used to generate clusters for model validation strategies. This study also aims at measuring the impact of removing outlier molecules prior to the clustering step. Clustering results are evaluated using three well-known clustering quality metrics. In addition, we compute an average similarity matrix to assess the quality of each cluster. The results show variation in clustering quality from method to method. The clusters obtained by the hierarchical and Taylor–Butina methods are more computationally expensive to use in cross-validation strategies, and both cluster the molecules poorly. In contrast, the UMAP method provides the best quality, and therefore we recommend it to analyze this highly valuable dataset.

Keywords: NCI-60 panel; small molecules; clustering; model validation

check for updates

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

Academic Editors: Lukasz Kurgan and Zhiliang Ji

Received: 16 January 2023
Revised: 2 March 2023
Accepted: 6 March 2023
Published: 8 March 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In the past decades, artificial intelligence (AI) has been used to develop predictive models with a wide range of applications in biomedicine and healthcare [1]. In particular, machine learning (ML)—a subarea of AI—has become an important component in early drug discovery, for instance, by developing quantitative structure–activity relationship (QSAR) models [2]. The use of ML-based models in drug discovery has been possible due to the availability of preclinical data that can be reused to build and validate predictive models. 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 [3].

Although ML-based models have been extensively used, several challenges remain to be overcome along the drug-design pipeline, one of them related to model performance on unseen compounds. Indeed, many articles fail to account for the nearly inevitable reduction in predictive ability that may occur when something that is a useful predictor in one data set is not as useful in another dataset [4]. Models that perform well on an independent data set can be achieved using model validation strategies such as bootstrap or k-fold cross-validation (CV) [4]. More demanding model validation strategies include asymmetric validation embedding (AVE) [5], leave-dissimilar-target-out (LDTO) CV [6], leave-one-cell-line-out (LOCO) CV, leave-one-tissue-out (LOTO) CV, leave-one-compound-cluster-out (LOCCO) CV [7], and those using other similarity metrics between training and test data instances [8]. These complex model-validation methods pose additional

challenges [9], particularly the LOCCO CV approach, which inherits the challenges of any clustering method.

Clustering methods are widely explored unsupervised ML-based algorithms whose aim is to discover the underlying structure or patterns existing in a given dataset. The clustering of biological entities, such as small molecules, can be performed using different approaches, including hierarchical clustering (HC), distribution-based clustering, and density-based clustering [10]. For example, several clustering algorithms, including hierarchical, Taylor–Butina, and UMAP clustering, have been compared on 29 data sets with between 100 and 5000 small molecules [11]. In addition, hierarchical clustering has been used to cluster molecules from the PubChem database [12], and Taylor–Butina clustering has been used to cluster molecules from the MolPort database [13].

A clustering analysis in a virtual chemical database can be used to quantify the diversity of compounds, which is relevant in several areas of chemistry such as in high-throughput screening (HTS) [14] and in QSAR-based virtual screening predictions [13]. In addition, the resulting clustering of compounds could lead to a comprehensive understanding of the underlying mechanism of action (MOA) of the drugs [15]. However, clustering chemical compounds also poses a challenge in terms of their representation. This is because chemical compounds can be represented in different ways, which can result in different clustering outcomes. It is possible for two compounds with different molecular structures to have comparable molecular descriptors or fingerprints, leading to them being clustered together using those representations. Similarly, two compounds with similar molecular structures may have different molecular descriptors or fingerprints, resulting in them being clustered separately using those representations. Structural fingerprints, such as the 1024-bit Morgan fingerprint, offer several advantages over other molecular representations for clustering tasks. For instance, they are computationally efficient and can be easily applied to large datasets. Moreover, Morgan fingerprints are robust to small variations in the molecular structure, making them a useful tool in virtual screening, where slight modifications to the molecular structure may occur due to synthetic or computational alterations.

In this study, we aim to evaluate and compare three different methods for clustering the NCI-60 molecules to determine the best-performing algorithm that can be used to generate clusters for model validation, ensuring that the resulting clusters are optimal for use in LOCCO-CV strategies or other model validation strategies. For this, we compare the hierarchical, Taylor–Butina, and UMAP clustering algorithms while tuning their essential hyperparameters. Additionally, we provide a definition of outlier molecules that aligns with the clustering problem we are addressing, allowing us to optimize outlier computation. By applying this definition, we investigate the impact of removing outlier molecules from the NCI-60 panel before clustering. To evaluate the clustering performance, we compute the silhouette coefficient, the Calinski–Harabaz score, and the Davies–Bouldin score for each method. Finally, alongside the clustering metrics, we compute a similarity matrix to achieve a better understanding of the molecules per cluster. The results show that, overall, removing outlier molecules results in better clustering. Moreover, the results show that UMAP clustering outperforms hierarchical and Taylor–Butina clustering.

2. Materials and Methods

The methodology of this study comprises three stages: data representation, clustering, and clustering evaluation. These stages are summarized in Figure 1 and are explained in detail in the following subsections. An important part of this study is to evaluate the impact of removing outlier molecules from the data set. Therefore, this study analyzes two scenarios per clustering method. In the first scenario, all of the unique molecules were clustered. In the second scenario, outlier molecules were removed before clustering. The outlier detection method is explained in Section 3.1.

It is important to note that the outlier detection method and clustering algorithms were applied to the entire set of molecules in the NCI-60 panel and not on a per-cell line

basis. Although the clustering output provides information on molecule NSC ID, SMILES representations, and their assigned clusters, it is possible to map the clustering results to each individual cell line.

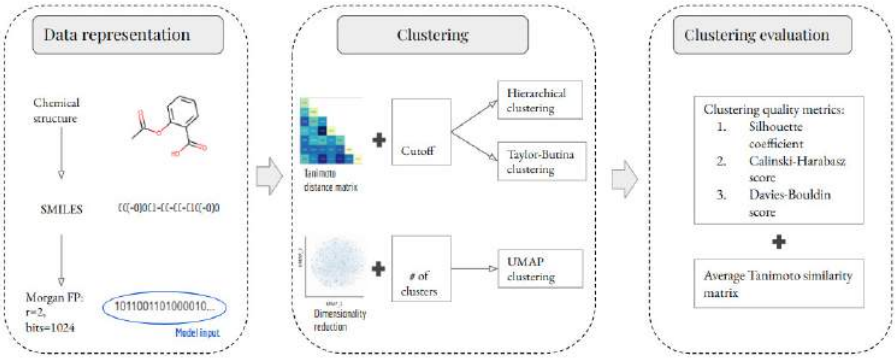


Figure 1. Schematic representation of the development of ML-based models for clustering small molecules. Chemical structures and their SMILES were retrieved from the NCI-60 dataset. SMILES were subsequently preprocessed and used to calculate the MFPs (radius 2 and 1024 bits). The MFPs were used to build and evaluate three clustering methods. Three clustering quality metrics were used to evaluate the clustering performance of each developed model.

2.1. NCI-60 Dataset

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 [3]. These 60 different human tumor cell lines comprise 9 cancer types: leukemia; melanoma; and cancers of the lung, colon, brain, ovary, breast, prostate, and kidney. All of the compounds that are screened have been 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 [16]. Each compound submitted to the NCI-60 panel for testing and evaluation is identified with a unique registration number called the National Service Center (NSC) ID.

Data quality is crucial in the development of AI/ML models during drug discovery. Therefore, data cleansing is necessary to ensure high-quality data is used to generate these models. To achieve this, we follow the preprocessing stage described by [17]. At the data-representation stage, each molecule is represented by a 1024-bit Morgan fingerprint (MFP) [18] with a radius of 2, which indicates the presence or absence of a particular substructure in the molecule. We chose this fingerprint for several reasons. Firstly, previous research has shown that the choice of fingerprint and metric has little effect on downstream predictions, such as target prediction based on molecular similarity [19]. Furthermore, our choice led to practically the best performance with respect to other fingerprints and metrics, suggesting that it is a near-optimal choice. In [20], the authors showed that the (extended-connectivity) fingerprints of radius 2 and 3 are among the best-performing fingerprints when ranking diverse structures by similarity. Previous research has shown that structural (or rule-based) fingerprints, such as E3FP, Morgan, or topological, should be considered for similarity-based clustering [21]. Finally, the 1024-bit MFP size has been successfully used in retrospective studies [22], for potency prediction [23], similarity searching, and bioactivity classification [24]. Therefore, in this study, the 1024-bit MFPs will be used as features to build the clustering models.

After the preprocessing stage, there remain around 2.7 M data points (50,555 small molecules screened against 60 cancer cell lines), which represent a matrix completeness of 89.25%. Figure 2 shows the distribution of the small molecules per cell line. We also analyze the Tanimoto similarity distribution of these small molecules (Figure 3a).

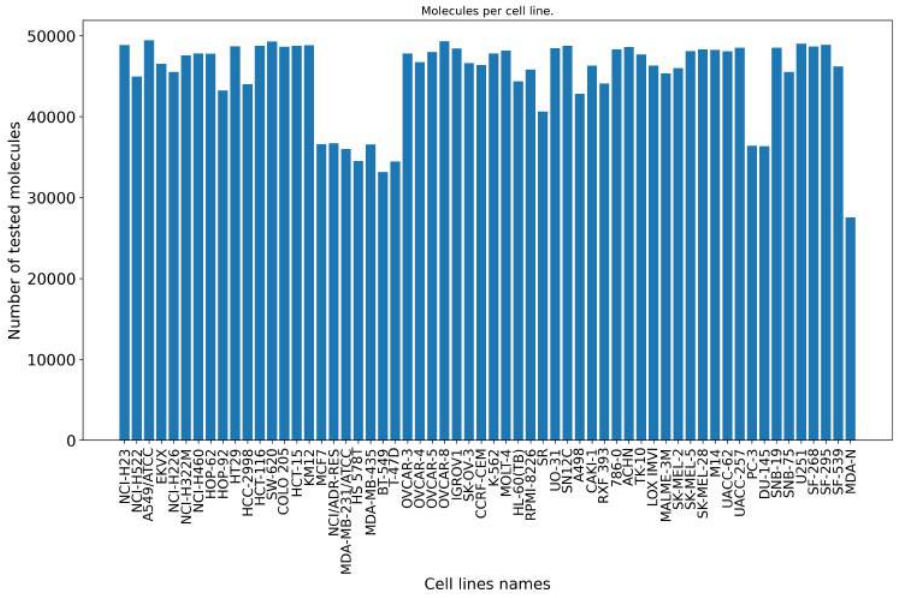
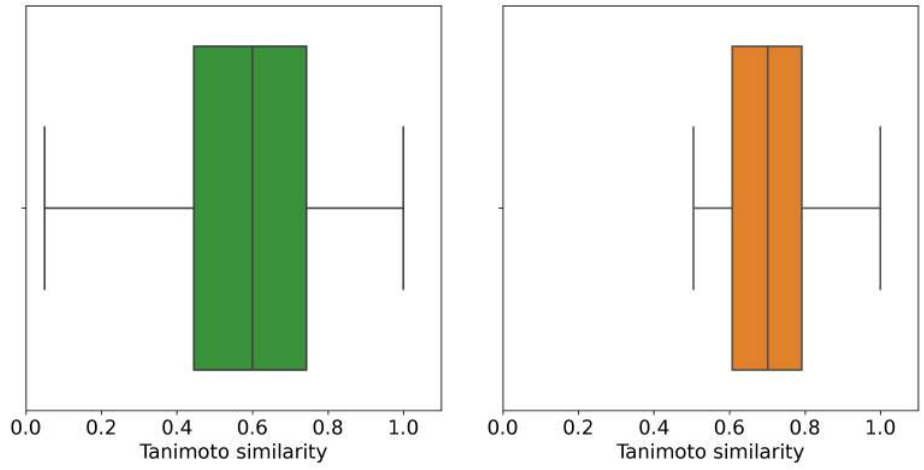


Figure 2. Number of tested molecules per NCI-60 cell line. On average, 45,124 small molecules are screened against each of the 60 cell lines from the NCI-60 panel. Each barplot represents the number of unique molecules (vertical axis) retrieved per cell line (horizontal axis) after the preprocessing stage.



(**a**) (**b**)

Figure 3. Distribution of the chemical similarity of molecules. (**a**) Before the outlier detection method (all 50,555 molecules); (**b**) after the outlier detection method, where 32,971 non-outlier molecules are retrieved at an outlier cutoff of 0.5. Each point in the corresponding boxplot is a molecule with its similarity to its closest molecule (other than itself). In total, 50% of molecules, either all of the molecules or non-outlier molecules, have their most similar molecule with similarity values less than 0.6 and 0.7, respectively.

2.2. Clustering Methods

Hierarchical clustering consists in building a binary merge tree, starting from the molecules stored at the leaves and merging them until reaching the root of the tree that contains all the molecules of the dataset [25]. The linkage criteria determine the metric used for the merge strategy, for example, single linkage, complete linkage, or Ward linkage. In the context of clustering chemical structure databases, the Ward linkage is commonly

used [26], which minimizes the sum of squared differences within all clusters. The graphical representation of the binary merge tree representing the resulting hierarchical clustering is called a dendrogram. The python implementation of this algorithm uses the RDkit library [27], where the first step is to calculate the similarity for each pair of molecules. Then, a distance matrix containing $1 - similarity$ for each pairwise similarity value is created. This distance matrix is the model input (Figure 1). Finally, the linkage criteria used in this study is the Ward linkage.

Taylor–Butina clustering is an algorithm based on exclusion spheres at a given Tanimoto level [28]. The way the clusters are built allows all of the molecules belonging to each cluster to have a Tanimoto value above or equal to the similarity cutoff used. At each iteration, the molecules are visited and labeled, either as a cluster centroid or as a cluster member. A disadvantage of this algorithm is that at the end of the clustering, molecules that have not been labeled are considered as singletons, even if they have neighbors. The reason is that their neighbors have been attracted by a better centroid. With this approach, we benefit from fast clustering since in each iteration only unlabeled molecules are compared, and we avoid the formation of highly heterogeneous clusters [28]. The python implementation of the Taylor–Butina algorithm employs the RDkit [27] library. The distance matrix is calculated in the same way as in hierarchical clustering (Figure 1); then, based on the similarity cutoff given, each molecule is assigned to a cluster id.

The uniform manifold approximation and projection (UMAP) is a non-linear dimensionality reduction algorithm that seeks to learn the manifold structure of the data and find a low-dimensional embedding while preserving the essential topological structure of that manifold [29]. While UMAP has been used for dimensionality reduction [30], it has also been used for clustering [11]. UMAP has four basic parameters to control the impact on the resulting embedding. These are *n_neighbors*, which controls how UMAP balances local versus global structure in the data; *min_dist*, which controls how tightly UMAP is allowed to pack points together; *n_components*, which allows the user to determine the dimensionality of the reduced dimension space we will be embedding the data into; and *metric*, which controls how distance is computed in the ambient space of the input data. After the dimensionality reduction is completed, this new representation of the molecules is clustered using the *AgglomerativeClustering* function from the Scikit-learn library [31].

2.3. Clustering Evaluation

The last step is the evaluation of the clustering algorithms. In this study, we consider three unsupervised metrics to evaluate the clustering quality results, the silhouette coefficient, the Calinski–Harabasz score, and the Davies–Bouldin score. These metrics are available in the Scikit-learn library [31].

To calculate the silhouette coefficient [32], we first calculate the mean intra-cluster distance a_i for each molecule i in the cluster C_I as follows:

$$a_i = \frac{1}{|C_I| - 1} \sum_{j \in C_I, i \neq j} d(i, j), \tag{1}$$

where $d(i, j)$ is the distance between molecules i and j in the cluster C_I. Next, we calculate the mean inter-cluster distance b_i for each molecule i to some cluster C_J as follows:

$$b_i = \min_{J \neq I} \frac{1}{|C_J|} \sum_{j \in C_J} d(i, j), \tag{2}$$

where $d(i, j)$ is the distance between the molecule i to all molecules in C_J, with $C_J \neq C_I$.

Then, the silhouette coefficient is defined for each molecule using the mean intra-cluster distance a_i and the mean inter-cluster distance b_i as:

$$SC(i) = \frac{b_i - a_i}{\max(a_i, b_i)}. \tag{3}$$

In summary, the silhouette coefficient is calculated using the mean distance between a given molecule and all other molecules in the same cluster (a_i), and the mean distance between a given molecule and all other molecules in the next nearest-cluster (b_i) [15]. The SC function implemented in the Scikit-learn library returns the mean silhouette coefficient over all molecules. The best value is 1, and the worst value is -1. Values near 0 indicate overlapping clusters. Negative values generally indicate that a molecule has been assigned to the wrong cluster as a different cluster is more similar.

The Calinski–Harabasz score [33], also known as the variance ratio criterion, is defined as the ratio of the sum of inter-clusters dispersion and of within-cluster dispersion for all clusters, where dispersion is defined as the sum of distances squared. This score is calculated for k clusters as follows:

$$CH = \frac{tr(B_k)}{tr(W_k)} \times \frac{n_E - k}{k - 1}, \tag{4}$$

where the $tr(B_k)$ is the trace of the between-cluster dispersion matrix, and $tr(W_k)$ is the trace of the within-cluster dispersion matrix, defined by:

$$W_k = \sum_{q=1}^{k} \sum_{x \in C_q} (x - c_q)(x - c_q)^T, \tag{5}$$

$$B_k = \sum_{q=1}^{k} n_q (c_q - c_E)(c_q - c_E)^T, \tag{6}$$

with C_q the set of points in cluster q, c_q the center of cluster q, c_E the center of cluster E, and n_q the number of points in cluster q. A higher Calinski–Harabasz score relates to a model with better-defined clusters.

Finally, the Davies–Bouldin score [34] is defined as the average similarity measure of each cluster C_i with its most similar cluster C_j:

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{i \neq j} R_{ij}, \tag{7}$$

where R_{ij} is defined as:

$$R_{ij} = \frac{s_i + s_j}{d_{ij}}, \tag{8}$$

and s_i is the average distance between each molecule of cluster and the centroid of that cluster, and d_{ij} is the distance between cluster centroids i and j. The minimum score is zero, with lower values indicating better clustering.

We complement this evaluation by calculating a matrix with the average Tanimoto similarity between the molecules of cluster i and those in cluster j. The average of these similarity scores represents the position (i, j) of the final matrix. In the case of average similarity between molecules from a cluster and itself (position (i, i)), the similarity was calculated between two different molecules. This means that values in position (i, i) do not include similarity scores of a molecule with itself. Note that the resulting matrix is a symmetric matrix (the similarity scores between cluster i and cluster j are equal to similarity scores between cluster j and cluster i); thus, for display purposes, the similarity matrix is shown as a lower triangular matrix.

3. Results

3.1. Outlier Detection

One of the aspects to consider when using ML-based algorithms, in particular those used for clustering, is that many of them are sensitive to outliers and fluctuations in the density of data points [35]. In this study, we defined an outlier as the molecule whose Tanimoto similarity value to their most similar molecule is lower than or equal to a given cutoff. Thus, an outlier molecule is different from any other molecule in the set at the predefined outlier cutoff.

To remove outlier molecules, we first calculated the Tanimoto similarity for each pair of molecules (1,277,878,735 pairs). Then, we retained the similarity value of each molecule to its closest molecule (other than itself). In this way, we only have 50,555 similarity values. We retain those molecules because all the other molecules will be even less similar. The next step was to find the best outlier cutoff based on the three clustering quality metrics. We used the hierarchical clustering (with *Ward linkage* = *3*) as the baseline algorithm for this experiment and evaluated the outlier cutoffs from 0.2 to 0.5, with a step size of 0.1.

Table 1 shows that as we increase the outlier cutoff, we obtain fewer clusters. This may be because at each outlier cutoff we have fewer molecules to cluster, and they are more similar to each other. Additionally, increasing the outlier cutoff generally leads to all three metrics improving their results with respect to the previous outlier cutoff. Moreover, the outlier cutoff of 0.5 seems to be the best tradeoff between the number of clusters and their quality since the silhouette coefficient and the Davies–Bouldin score achieve better values at this outlier cutoff. Therefore, the outlier cutoff of 0.5 was used for the following experiments. Figure 3 shows the distribution of similarity values before and after the outlier detection method.

Table 1. Hierarchical clustering (with Ward linkage = 3) to identify the best outlier cutoff. Increasing the outlier cutoff, and thus removing more molecules, leads to better quality clustering according to the three selected metrics. Comparison of outlier cutoffs based on the number of molecules removed and the clustering metrics. The best value for each metric is highlighted in bold. For readability, clustering quality results were rounded to either two or three decimal places.

Outlier Cutoff	Molecules Removed	# Clusters	Silhouette Coefficient	Calinski–Harabasz Score	Davies–Bouldin Score
0.2	151	6	−0.011	5.10	87.72
0.3	2096	6	−0.013	4.65	79.91
0.4	8994	5	−0.006	**5.45**	84.81
0.5	17,584	4	**−0.007**	5.32	**67.15**

3.2. Hierarchical Clustering

To perform hierarchical clustering of molecules we have to specify the Ward linkage cutoff to be used. We evaluated the impact of different cutoffs on the three selected clustering quality metrics, as well as on the number of clusters obtained. The Ward linkage cutoffs explored ranged from 0.5 to 3.0, with a step size of 0.5. These cutoffs were applied to either all molecules or non-outlier molecules and are reported in Table 2. By using the Ward linkage as the merging criterion, we are requesting that at each step the hierarchical clustering algorithm has to find the pair of clusters that leads to the minimum increase in total intra-cluster variance after merging. Table 2 shows that stricter (smaller) Ward linkage cutoffs result in a larger number of clusters.

Regarding the clustering quality metrics, the silhouette coefficient suggests that some molecules have been assigned to the wrong cluster as negative values are obtained at each cutoff. The Calinski–Harabasz score suggests that better-defined clusters are obtained at higher cutoff values, 2.5 and 3.0 for instance. In contrast, the Davies–Bouldin score suggests that better-defined clusters are obtained at smaller cutoff values but at the cost

of a larger number of clusters. This behavior is repeated for all molecules as well as for non-outlier molecules.

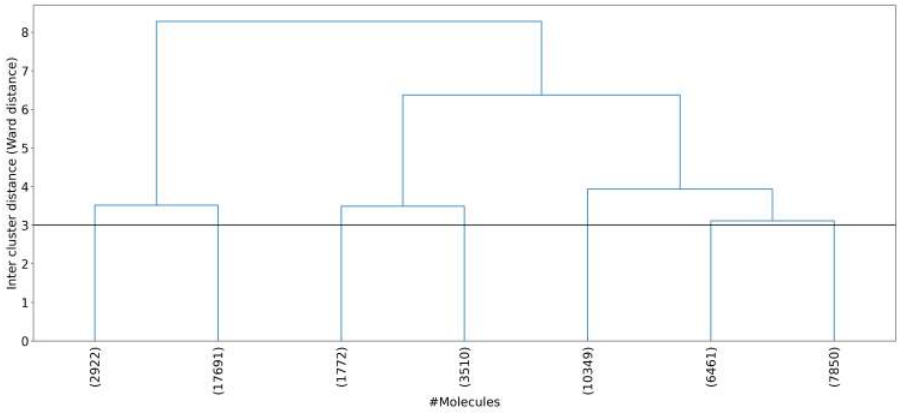
Considering that the main purpose of this clustering problem is to use the clusters for ML model validation, the results in Table 2 suggest that a cutoff value between 2.0 and 3.0 generates a number of clusters that is less computationally expensive to use in a cross-validation strategy. In particular, the Ward linkage cutoff of 3.0 achieves the best clustering quality results, and it improves between all molecules and non-outlier molecules. Indeed, at this cutoff the silhouette coefficient improves by 50%, the Calinski–Harabasz score improves by 10%, and the Davies–Bouldin score improves by 13% when using non-outlier molecules. Therefore, we chose the Ward linkage cutoff of 3.0 for the following experiments.

Table 2. Clustering quality depending on cutoff and outlier removal. Overall, an improvement in each of the clustering quality metrics is observed when outlier molecules are removed. Clustering quality metrics were calculated for each Ward linkage cutoff evaluated in hierarchical clustering. From top to bottom, each cutoff is stricter than the previous one. Here, 32,971 non-outlier molecules were retrieved at an outlier cutoff of 0.5. The best value for each metric for either all molecules or non-outlier molecules is highlighted in bold. For readability, clustering quality results were rounded to either two or three decimal places.

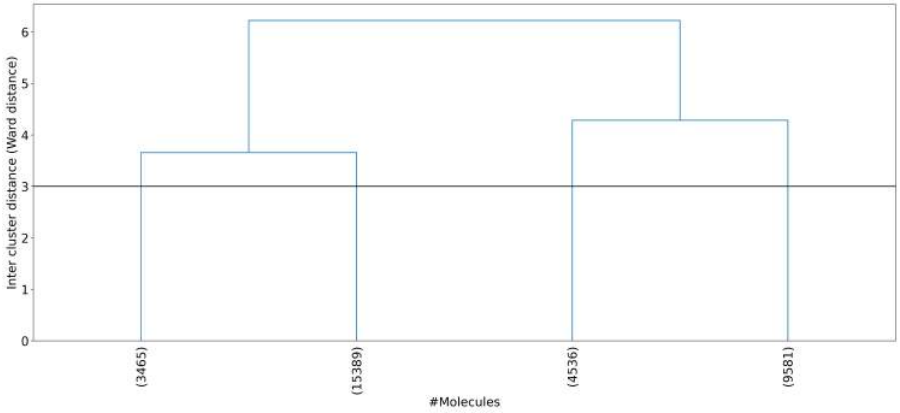
	All Molecules				Non-Outlier Molecules			
Ward Linkage Cutoff	**# Clusters**	**Silhouette Coefficient**	**Calinski–Harabasz Score**	**Davies–Bouldin Score**	**# Clusters**	**Silhouette Coefficient**	**Calinski–Harabasz Score**	**Davies–Bouldin Score**
3.0	7	**−0.014**	**4.84**	77.19	4	**−0.007**	**5.32**	67.15
2.5	9	**−0.014**	4.01	80.19	7	−0.009	3.50	69.94
2.0	89	−0.022	1.77	28.63	88	−0.030	1.96	21.45
1.5	2565	−0.106	1.28	7.28	1766	−0.103	1.37	6.97
1.0	17,958	−0.257	1.05	2.59	11,436	−0.260	1.07	2.62
0.5	33,246	−0.305	1.01	**1.43**	20,723	−0.339	1.01	**1.53**

The next step was to analyze the number of molecules per cluster. Figure 4 shows the dendrogram for both all 50,555 molecules and 32,971 non-outlier molecules. When all of the molecules are clustered (Figure 4a), one cluster concentrates about 35% of the molecules (17,691 molecules in cluster 2), whereas when non-outlier molecules are clustered (Figure 4b), one cluster concentrates about 46% of the molecules (15,389 molecules in cluster 2). Even with this high concentration of molecules in a single cluster, the remaining clusters have enough information as the smallest clusters obtained in all molecules and non-outlier molecules concentrate 3.5% and 10.5% of the molecules, respectively.

To complement the analysis of the clustering quality, we now calculate the average similarity matrix for both all molecules and non-outlier molecules. This matrix was calculated as explained in Section 2.3. In a good clustering result, molecules from different clusters should have much lower similarity than molecules from the same cluster. In the average similarity matrix, this means that the similarity values on the diagonal must always be greater than the off-diagonal values. Figure 5 shows that molecules from the same cluster are, on average, very similar to molecules from different clusters since the off-diagonal values are mostly equal to the diagonal values, and in some cases even greater than the diagonal values. This suggests a poor clustering of molecules by using hierarchical clustering (with Ward linkage = 3), even when outlier molecules are removed. This assumption is consistent with the values of the clustering quality metrics reported in Table 2.



(**a**)



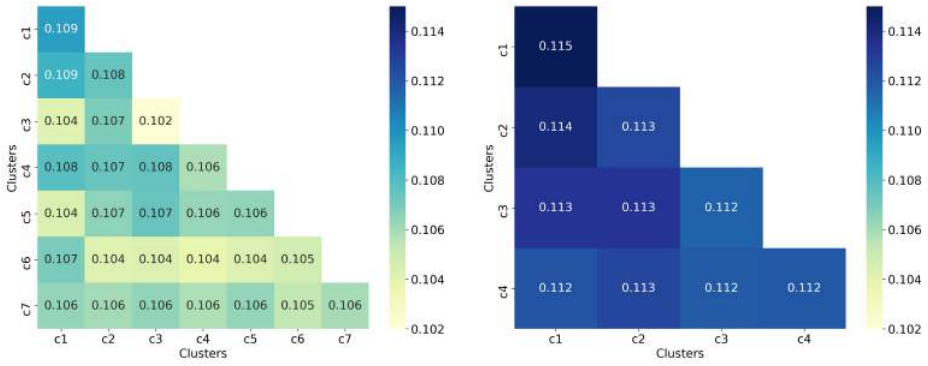
(**b**)

Figure 4. Dendrogram representation of hierarchical clustering results. (**a**) Clustering of all 50,555 molecules results in 7 clusters at a Ward linkage cutoff of 3. (**b**) Clustering of 32,971 non-outlier molecules results in 4 clusters at a Ward linkage cutoff of 3. For each plot, the horizontal axis shows the number of molecules per cluster, and the vertical axis shows the Ward linkage between any two clusters. Outlier molecules were removed using an outlier cutoff of 0.5 (see Section 3.1 for more details about outlier detection). The average cluster size is 7229 molecules (std = 5517) when all molecules are clustered and 8243 molecules (std = 5460) when non-outlier molecules are clustered.

3.3. Taylor–Butina Clustering

One of the advantages of Taylor–Butina clustering is that molecules in the resulting clusters will have a Tanimoto value greater than or equal to the established similarity cutoff. In this study, we analyzed similarity cutoffs 0.35, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 0.97, and 0.99. These cutoffs were applied to either all molecules or non-outlier molecules and evaluated in terms of the three clustering quality metrics.

Table 3 shows that smaller similarity cutoffs result in a larger number of clusters. As for the clustering quality metrics, the silhouette coefficient and the Davies–Bouldin score suggest that better-defined clusters are obtained with smaller similarity cutoff values but at the cost of a higher number of clusters. This trend is observed for all molecules as well as for non-outlier molecules. The number of clusters should also be considered as it impacts the number of cross-validations and hence the computational cost. Therefore, if we consider the number of clusters then the Calinski–Harabasz score suggests that the similarity cutoff of 0.9 is the best option. Overall, the three clustering quality metrics improve between all molecules and non-outlier molecules.



(**a**) (**b**)

Figure 5. Hierarchical clustering with Ward linkage = 3. (**a**) Clustering of all 50,555 molecules results in 7 clusters. (**b**) The clustering of 32,971 non-outlier molecules results in 4 clusters. Outlier molecules were removed using an outlier cutoff of 0.5 (see Section 3.1 for more details about outlier detection). The matrix is calculated as the average Tanimoto similarity between the molecules of cluster *i* and those in cluster *j*, and the average similarity between molecules of a cluster and itself (see Section 2.3 for more details on how this matrix is calculated). The matrices have been adjusted to the same scale to facilitate clustering comparison between all molecules and non-outlier molecules. For readability, similarity scores were rounded to three decimal places. The average similarity between a cluster and itself (diagonal) is close to the average similarity between two different clusters (off-diagonal), indicating poor clustering of both all molecules and non-outlier molecules.

Table 3. Better-defined clusters are obtained when using smaller similarity cutoff values, at the cost of a larger number of clusters. Clustering quality metrics were calculated for each similarity cutoff evaluated in the Taylor–Butina clustering. Non-outlier molecules were retrieved at an outlier cutoff of 0.5. The best value for each metric for either all molecules or non-outlier molecules is highlighted in bold. For readability, clustering quality results were rounded to either two or three decimal places.

	All Molecules				Non-Outlier Molecules			
Similarity Cutoff	**# Clusters**	**Silhouette Coeffi-cient**	**Calinski–Harabasz Score**	**Davies–Bouldin Score**	**# Clusters**	**Silhouette Coeffi-cient**	**Calinski–Harabasz Score**	**Davies–Bouldin Score**
0.35	30,590	**0.111**	6.10	**0.65**	16,860	**0.120**	7.09	**0.72**
0.50	19,513	0.086	5.27	0.93	9077	0.092	7.06	1.06
0.60	12,797	0.029	4.95	1.20	5996	0.042	6.83	1.30
0.70	5872	−0.067	4.96	1.66	2984	−0.046	6.54	1.74
0.80	1074	−0.168	6.64	2.61	634	−0.155	7.38	2.63
0.90	64	−0.196	**9.03**	3.78	44	−0.201	**7.48**	3.83
0.95	21	−0.204	2.85	3.27	11	−0.203	2.25	2.73
0.97	12	−0.214	1.40	2.51	5	−0.202	0.76	1.96
0.99	4	−0.215	0.65	2.12	1	–	–	–

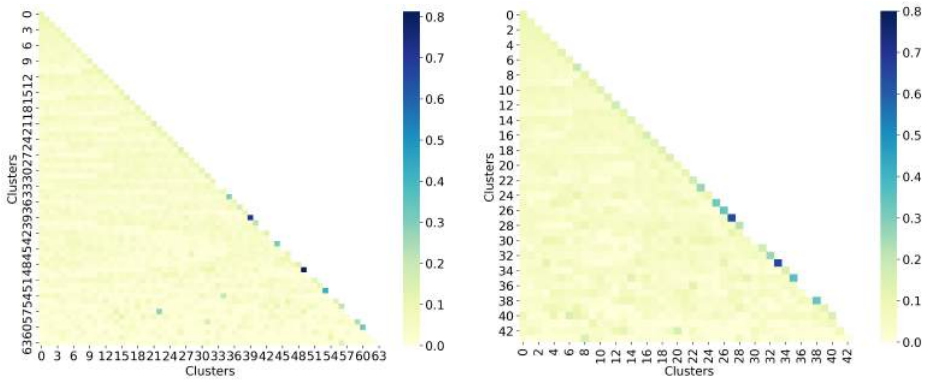
Since one of the drawbacks of the Taylor–Butina algorithm is the possibility of obtaining clusters that are singletons, the next step in this analysis is to evaluate the number of molecules per cluster. Based on results from Table 3, we choose the similarity cutoff of 0.9 to perform this experiment. Indeed, Table 4 shows that when all molecules are clustered, 17% of the clusters obtained are singletons. This value decreases to 10% in the non-outlier molecules. Moreover, 90% of all molecules and 93% of the non-outlier molecules have been assigned to a single cluster.

The last step in the clustering quality analysis is to calculate the average similarity matrix for both all molecules and non-outlier molecules. To analyze these matrices, which are calculated as the average within-cluster and between-cluster similarity, the sizes of the clusters must be taken into account. Indeed, since both cases (all molecules and non-outlier molecules) have singletons, it is not possible to calculate the average within-cluster similarity, so these values are missing on the matrix diagonal. For the average between-cluster similarity, it can be the similarity between two molecules. Overall, Figure 6 shows

that molecules from different clusters have lower similarity than molecules from the same cluster, notably for non-outlier molecules. However, these values may be biased by the number of molecules per cluster.

Table 4. Descriptive statistics of the cluster sizes obtained by the Taylor–Butina algorithm. This algorithm was applied to all molecules and non-outlier molecules, showing that singletons were obtained in both cases. The similarity cutoff used to obtain the clusters in Taylor–Butina was set to 0.9 (see Table 3 for more details about the similarity cutoff selected). Non-outlier molecules were retrieved at an outlier cutoff of 0.5.

	All Molecules	**Non-Outlier Molecules**
Number of clusters	64	44
mean	790	749
std	5703	4617
min	1	1
10%	1	1
17%	1	2
20%	2	2
25%	3	3
50%	7	13
75%	52	30
max	45,652	30,669



(**a**) (**b**)

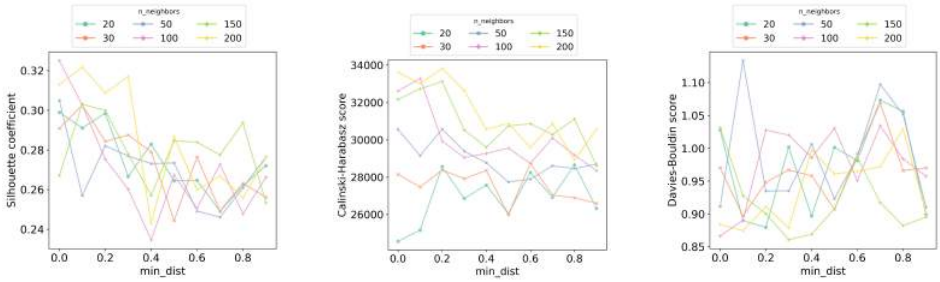
Figure 6. Taylor-Butina clustering with similarity cutoff = 0.90. (**a**) Clustering of all 50,555 molecules results in 64 clusters, and (**b**) clustering of 32,971 non-outlier molecules results in 44 clusters. Outlier molecules were removed using an outlier cutoff of 0.5 (see Section 3.1 for more details about outlier detection). The matrix is calculated as the average Tanimoto similarity between the molecules of cluster *i* and those in cluster *j*, and the average similarity between molecules of a cluster and itself (see Section 2.3 for more details on how this matrix is calculated). The matrices have been adjusted to the same scale to facilitate clustering comparison between all molecules and non-outlier molecules. The average similarity between one cluster and itself (diagonal) is substantially higher than the average similarity between two different clusters (off-diagonal). However, these values may be biased by the number of molecules per cluster.

3.4. UMAP Clustering

UMAP has four basic hyperparameters (*n_neighbors*, *min_dist*, *n_components*, and *metric*) that control the dimensionality reduction result and one that controls the clustering (*n_clusters*). In this study, MFPs are embedded in two dimensions (*n_components* = 2) and the Jaccard *metric* is used to calculate the distance between MFPs. For the remaining hyperparameters (*n_neighbors*, *min_dist*, and *n_clusters*), we implemented a two-step hyperparameter tuning process to find their optimal values. In the first step, we fixed the number of clusters and tune the values of the number of neighbors (*n_neighbors*) and the distance (*min_dist*). Based on the

results obtained with hierarchical clustering, when all molecules are clustered, the number of clusters was set to 7. The evaluation of this experiment was done in terms of the three clustering quality metrics. Once these hyperparameters have been set, the second step of hyperparameter tuning uses these values and focuses on finding the optimal number of clusters by using the *elbow* method [36]. The elbow method fits the clustering model for a range of *n_clusters* values. If the data are very clustered, the optimal number of clusters is given by the point of inflection on the curve (i.e., the elbow); otherwise, the elbow will be unclear.

In the first step of the hyperparameter tuning process, we performed a grid search to look for the best *n_neighbors* and *min_dist* hyperparameters. The values of *n_neighbors* evaluated are 20, 30, 50, 100, 150, and 200, while for *min_dist* the values evaluated are in the range of 0.0 to 0.9, with a step size of 0.1. Figure 7 shows the results of each of the clustering quality metrics obtained at each hyperparameter combination. Overall, the results suggest that using 100 or 200 neighbors UMAP is able to balance local versus global structure in the molecules in a better way than with fewer neighbors. Additionally, at these numbers of neighbors, the best *min_dist* is 0.0, which means molecules represented in the embedded space are close to each other. Since the combination (100, 0.0) is the best in two of the three clustering quality metrics, these values were used in the second step of the hyperparameter tuning.



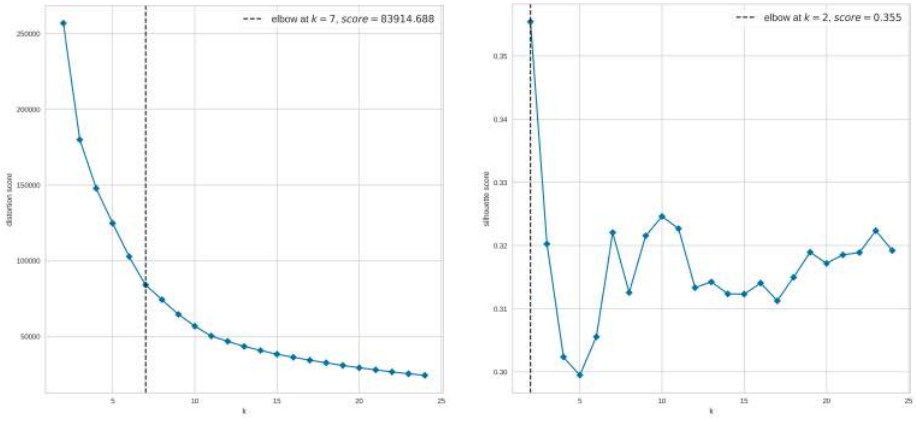
(**a**) (**b**) (**c**)

Figure 7. Increasing the number of neighbors leads to recovering better-defined clusters. In the first step of the hyperparameter tuning process for UMAP clustering, a grid search was performed to find the best combination of the number of neighbors (*n_neighbors*) and distance (*min_dist*). The number of clusters used for this search was set to 7. Each combination of (*min_dist*, *n_neighbors*) values was evaluated considering the clustering quality metrics (**a**) silhouette coefficient, (**b**) Calinski–Harabasz score, and (**c**) Davies–Bouldin score. For each plot, the horizontal axis shows the distance values (*min_dist*) evaluated, the vertical axis shows the corresponding metric values, and the color code identifies the number of neighbors (*n_neighbors*) evaluated.

Given the values $n_neighbors = 100$ and $min_dist = 0.0$, we now look for the optimal number of clusters to be used. Since our evaluation of clustering quality goes beyond the three clustering metrics, we use the elbow method to narrow the search for the number of clusters rather than to find the optimal number. Here, we explored the range of clusters from 2 to 25. We also compared two scoring metrics to evaluate the clusters. These scoring metrics are the distortion, defined by the sum of squared distances between each observation and its closest centroid, and the silhouette metric, defined by the mean ratio of intra-cluster and nearest-cluster distance.

Figure 8a suggests that the optimal number of clusters is 7 when using the distortion score. However, Figure 8b suggests that our data are not highly clustered as the elbow is not clearly defined when the Silhouette score is used. Instead, the method returns as the optimal number of clusters the value at which the highest Silhouette score was obtained; in this case, the suggested optimal number is 2 clusters. Since we are using the elbow method to narrow the search for the optimal number of clusters, from Figure 8 we analyze clusters 6, 7, and 8. In addition, to find out whether a larger number of clusters improves the clustering results, we also explored the UMAP clustering when 20 groups are requested.

These values were applied to either all molecules or non-outlier molecules and evaluated in terms of the three clustering quality metrics.



(**a**) (**b**)

Figure 8. Elbow method to narrow the search for the optimal number of clusters when using UMAP clustering. In the second step of the hyperparameter tuning process for UMAP clustering, the elbow method was explored, covering a range of clusters from 2 to 25, and using (**a**) the distortion metric and (**b**) the silhouette metric. For each plot, the horizontal axis shows the number of clusters (*k*) evaluated, and the vertical axis shows the corresponding metric.

With the best combination of hyperparameters for UMAP found, we evaluated the clustering results in terms of the clustering quality metrics obtained at each number of clusters. Table 5 shows these results, where, in summary, all clustering quality metrics suggest well-defined clusters. Even when all metrics reach their best value with different numbers of clusters, the difference in the results with respect to the best value is small. In fact, the difference of the best silhouette coefficient with respect to the other values is between 3–6%; for the Calinski–Harabasz score, this difference ranges from 0.5% to 12%; and for the Davies–Bouldin score the difference ranges from 2% to 4%. For non-outlier molecules, the difference between the best and worst value for each of the clustering metrics is between 4.5 and 11% for the silhouette coefficient, between 6 and 12% for the Calinski–Harabasz score, and between 1 and 5.5% for the Davies–Bouldin score.

The next step was to analyze the number of molecules per cluster when 7 and 20 clusters are requested. These two values were selected considering the optimal number of clusters obtained with the elbow method (Figure 8a) and the results of the clustering metrics (Table 5). Table 6 shows the descriptive statistics of the cluster sizes, for both all 50,555 molecules and 32,971 non-outlier molecules. In general, the size of the clusters is well distributed (mean cluster size is close to median cluster size) for all molecules as well as for non-outlier molecules. Moreover, each of the 7 clusters concentrates between 5 and 22% of all molecules, and between 4 and 24% of non-outlier molecules. However, when 20 clusters are required, the smallest clusters in each case concentrate 1.19% and 0.96% of the molecules, respectively. This should be taken into account when performing cross-validation since having few molecules could affect the model performance.

The last step in the clustering quality analysis is to calculate the average similarity matrix for both all molecules and non-outlier molecules. Overall, Figure 9 shows that molecules from different clusters have lower similarity than molecules from the same cluster, which indicates good clustering results. This is consistent with the clustering quality metrics shown in Table 5. Moreover, the similarity values between a cluster and itself (diagonal) improve when only non-outlier molecules are clustered. As for the number of clusters, increasing this number (from 7 to 20) increases the similarity between a cluster and itself from 0.16 to 0.22, for the upper bound.

Table 5. Clustering quality metrics suggest well-defined clusters at each selected number of clusters. Clustering quality metrics were calculated for each number of clusters evaluated in the UMAP algorithm. Non-outlier molecules were retrieved at an outlier cutoff of 0.5. The best value for each metric for either all molecules or non-outlier molecules is highlighted in bold. For readability, clustering quality results were rounded to either two or three decimal places.

	All Molecules			Non-Outlier Molecules		
Number of Clusters	**Silhouette Coefficient**	**Calinski Harabasz Score**	**Davies Bouldin Score**	**Silhouette Coefficient**	**Calinski Harabasz Score**	**Davies Bouldin Score**
6	0.306	30,410.88	0.90	0.318	20,683.06	**0.85**
7	**0.325**	32,621.97	0.87	0.295	21,420.86	0.89
8	0.314	34,263.14	**0.86**	0.302	22,194.48	0.90
20	0.315	**34,464.21**	0.88	**0.333**	**23,609.54**	0.86

Table 6. Descriptive statistics of the cluster sizes obtained by the UMAP clustering. This algorithm was applied to all molecules and non-outlier molecules show that the size of the clusters is well distributed in both cases. Non-outlier molecules were retrieved at an outlier cutoff of 0.5. The number of clusters required was set to either 7 or 20 (see Table 5 for more details about the number of clusters selected).

	All Molecules		Non-Outlier Molecules	
# of clusters	7	20	7	20
mean	7222	2528	4710	1649
std	3575	1032	2149	846
min	2664	605	1811	318
25%	4444	1981	3241	884
50%	6960	2486	4684	1524
75%	10,522	3202	5961	2391
max	11,000	4364	8072	3058

3.5. Comparison of Clustering Algorithms

Finally, we compare the best clustering results obtained with each of the three methods explored in this study. In the case of UMAP clustering, we compare the results when 7 and 20 clusters are required. This comparison is performed for all molecules as well as for non-outlier molecules. Additionally, we include the run time (in minutes) taken by each method. Table 7 summarizes the results of this comparison.

When all molecules are clustered, the UMAP model obtains the best results on all metrics, whether using 7 or 20 clusters. Indeed, the silhouette coefficient improves between 0.339 and 0.521 units with respect to hierarchical and Taylor–Butina clustering. With respect to the Calinski–Harabaz score, although there is no upper limit for this score, the improvement is evident, with more than 34K units. As for the Davies–Bouldin score, the improvement is between 3–76 units. Regarding the run time, the Taylor–Butina clustering consumes 10x more time than the UMAP clustering, which is the fastest method. Similar behavior occurs when non-outlier molecules are clustered. That is, the UMAP clustering achieves the best results in terms of the three quality metrics as well as the run time.

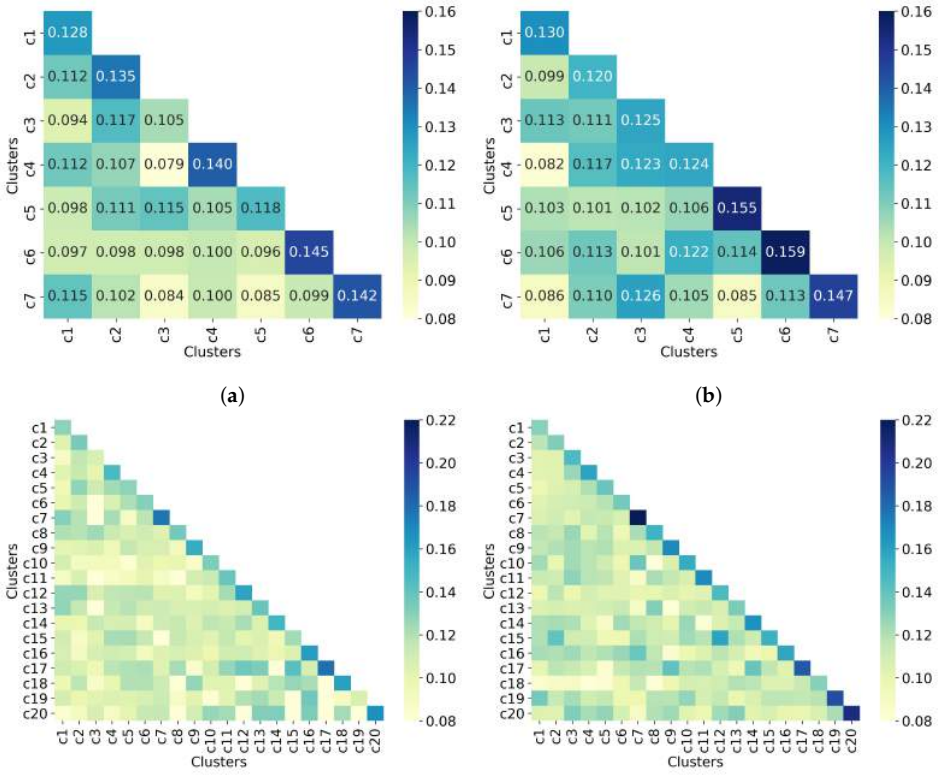


Figure 9. UMAP clustering with *n_neighbors* = 100 and *min_dist* = 0.0. Clustering of (**a**) all 50,555 molecules into 7 clusters, (**b**) 32,971 non-outlier molecules into 7 clusters, (**c**) all 50,555 molecules into 20 clusters, and (**d**) 32,971 non-outlier molecules into 20 clusters. Outlier molecules were removed using an outlier cutoff of 0.5 (see Section 3.1 for more details about outlier detection). The matrix is calculated as the average Tanimoto similarity between the molecules of cluster *i* and those in cluster *j*, and the average similarity between molecules of a cluster and itself (see Section 2.3 for more details on how this matrix is calculated). The matrices have been adjusted to the same scale to facilitate clustering comparison between all molecules and non-outlier molecules. For readability, similarity scores were rounded to three decimal places when seven clusters were requested. When 20 clusters were requested, the similarity scores are not displayed. The average similarity between one cluster and itself (diagonal) is substantially higher than the average similarity in different clusters (off-diagonal), indicating well-defined clusters in both all molecules and non-outlier molecules.

In general, removing outlier molecules prior to clustering leads to an improvement in hierarchical and UMAP clustering (with 20 clusters). In the case of the Taylor–Butina clustering, there is no improvement by performing this step; in fact, the silhouette coefficient decreases by 2.5%, the Calinski–Harabasz score decreases by 17%, and the Davies–Bouldin score decreases by 1.3%. The absence of improvement in UMAP with seven clusters when outliers are removed may be due to the fact that according to the elbow method (Figure 8a), this is the optimal number of clusters, and therefore an improvement could be difficult to achieve.

Table 7. UMAP clustering outperforms results obtained by hierarchical and Taylor–Butina clustering. Comparison of the clustering quality metrics for each clustering algorithm. Non-outlier molecules were retrieved at an outlier cutoff of 0.5. The best value for each metric for either all molecules or non-outlier molecules is highlighted in bold. For readability, clustering quality results were rounded to either two or three decimal places.

	All Molecules					Non-Outlier Molecules				
Clustering Method	**Number of Clusters**	**Silhouette Coefficient**	**Calinski Harabasz Score**	**Davies Bouldin Score**	**Time Took**	**Number of Clusters**	**Silhouette Coefficient**	**Calinski Harabasz Score**	**Davies Bouldin Score**	**Time Took (min)**
HC	7	−0.014	4.84	77.19	25	4	−0.007	5.32	67.15	3
TB	64	−0.196	9.03	3.78	40	44	−0.201	7.48	3.83	6
UMAP	7	**0.325**	32,621.97	**0.87**	11	7	0.295	21,420.86	0.89	2
UMAP	20	0.315	**34,464.21**	0.88	4	20	**0.333**	**23,609.54**	**0.86**	2

4. Discussion

The main aim of this study is to provide an optimal clustering of molecules from the NCI-60 panel, which can be used to generate clusters for model validation. As model validation is a technique to assess model generalization, high-quality clustering results could improve the generalization of ML-based models. However, in studies using clustering methods that derive clusters to be used in cross-validation (LOCCO-CV, for example), the analysis of clustering quality is usually omitted and at best restricted to a single metric. To provide a comprehensive comparison of the clustering results obtained by hierarchical, Taylor–Butina, and UMAP clustering, in this study we show three well-known clustering metrics, along with the similarity matrix. The latter provides information on the structure of each cluster obtained.

The results show that the clustering quality metrics vary from method to method (Tables 2, 3 and 5). Using different cutoffs in hierarchical and Taylor–Butina clustering leads to results that are computationally expensive to use under a cross-validation strategy. Even at cutoff values where there is a trade-off between the number of clusters obtained and the quality of the clustering, according to the similarity matrix (Figures 5 and 6), the molecules in different clusters are similar to each other. This inter-cluster similarity may affect the model performance since similar molecules are present in the training set and the test set. This highlights the importance of complementing the usual clustering metrics by calculating the average similarity matrix.

In methods such as hierarchical or Taylor–Butina clustering, the number of clusters is not an input but a consequence of the selected cutoff level. In the case of UMAP clustering, the number of clusters must be provided by the user. Since this parameter can affect the clustering quality, we address this problem by using the elbow method. However, instead of using it to find the optimal number of clusters, we use it to narrow the search for this number. The results suggest that we can generally get higher-quality clustering if we request a higher number of clusters. However, we require a compromise between the number of clusters and clustering quality as more clusters imply more cross-validations and thus more expensive computation.

In addition to the optimal number of clusters, the elbow method also provides insights into the clusterability of the data set (Figure 8). In this case, this analysis suggests that the molecules in the NCI-60 panel are not very clustered as the inflection point of the curve is not clearly observed (Figure 8b). This has a greater impact on hierarchical and Taylor–Butina clustering, in addition to the similarity distribution of molecules (Figure 3), since these methods also require a cutoff that determines the behavior of the clusters (intra-cluster distance or similarity between molecules).

Hierarchical, Taylor–Butina, and UMAP clustering were also tested to evaluate if the removal of outlier molecules improves clustering quality. We define outlier molecules and evaluate the results according to the desired clustering. In this case, the results demonstrate a benefit when non-outlier molecules are clustered using hierarchical and

UMAP clustering. This is not the case for the Taylor–Butina clustering, where a decrease in the metrics is observed (Table 7). This suggests that Taylor–Butina is not suitable for clustering molecules having a distribution as shown in Figure 3 since in both scenarios (all molecules or non-outlier molecules) we obtain clusters that concentrate most of the molecules or singletons. While the adopted outlier detection method improves clustering, we believe that a comprehensive search for an optimal method using high-quality packages such as PyOD [37] is likely to result in further improvement.

In summary, there are many factors influencing clustering problems, particularly the clustering of molecules. Models that are not properly validated are susceptible to reduced performance in unseen compounds; therefore, poor clustering of molecules can lead to the poor estimation of model generalizability. Since results showed different clustering quality metrics with respect to the method used, special care must be taken in this task.

5. Conclusions

- In this paper, we provide a comparison of hierarchical, Taylor–Butina, and UMAP clustering to find the best method to cluster NCI-60 molecules. We also evaluated the impact of removing outlier molecules before clustering for each of these methods.
- Results suggest that the most effective way to cluster the NCI-60 molecules is by using the UMAP algorithm either with 7 or 20 clusters depending on the clustering quality metric selected. In addition, the choice of the number of clusters to be used must be balanced with the computational cost that the user can afford since more clusters imply more cross-validations.
- All three clustering methods have been tested on other datasets; however, we are not aware of any previous studies that have clustered all or part of the molecules from the NCI-60 panel nor of studies comparing results when outlier molecules are removed.
- High-quality clustering results could improve model generalization on unseen compounds, which is important when the scope and use of the predictive model need to be extended. Therefore, special care must be taken in the clusters to be used in model validation, such as LOCCO-CV strategies.
- In this study, the 1024-bit Morgan fingerprint was utilized as a descriptor to characterize the molecules in the NCI-60 panel. Although this is likely to be a near-optimal choice given past comparative studies, future work could investigate other descriptors to find out whether they improve the clustering of the NCI-60 molecules further.
- Another interesting topic for future work is to investigate how clustering is affected by the inclusion of large proportions of assumed inactive molecules (decoys) generated for each NCI-60 molecule with potent activity as it wis important that decoys are contained in the same cluster as the active they were generated from [38].

Author Contributions: Conceptualization, S.H.-H. and P.J.B.; methodology, S.H.-H. and P.J.B.; software, S.H.-H.; validation, S.H.-H. and P.J.B.; formal analysis, S.H.-H.; investigation, S.H.-H.; resources, P.J.B.; data curation, S.H.-H.; writing—original draft preparation, S.H.-H.; writing—review and editing, P.J.B.; visualization, S.H.-H.; supervision, P.J.B.; project administration, P.J.B.; and funding acquisition, P.J.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by the National Council of Sciences and Technology of Mexico (CONACYT) grant number 775584.The Wolfson Foundation and the Royal Society for a Royal Society Wolfson Fellowship awarded to P.J.B.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The code for the reproduction of the best clustering results is available at https://github.com/Sahet11/nci60_clustering, accessed on 1 February 2023.

Acknowledgments: We thank the anonymous reviewers for their feedback.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial intelligence
ML	Machine learning
QSAR	Quantitative structure–activity relationships
AVE	Asymmetric validation embedding
HTS	High-throughput screening
CV	Cross-validation
LDTO	Leave-dissimilar-target-out
LOTO	Leave-one-tissue-out
LOCCO	Leave-one-compound-cluster-out
LOCO	Leave-one-cell-line-out
MOA	Mechanism of action
NCI	National Cancer Institute
NSC	National Service Center
MFP	Morgan fingerprint
SMILES	Simplified molecular-input line-entry system
UMAP	Uniform manifold approximation and projection
HC	Hierarchical clustering
TB	Taylor–Butina
SC	Silhouette coefficient
CH	Calinski–Harabasz score
DB	Davies–Bouldin score

References

1. Ballester, P.J.; Stevens, R.; Haibe-Kains, B.; Huang, R.S.; Aittokallio, T. Artificial intelligence for drug response prediction in disease models. *Brief. Bioinform.* **2022**, *23*, bbab450.
2. Ballester, P.J. Machine learning for molecular modelling in drug design. *Biomolecules* **2019**, *9*, 216. [CrossRef]
3. Shoemaker, R.H. The NCI60 human tumour cell line anticancer drug screen. *Nat. Rev. Cancer* **2006**, *6*, 813–823. [CrossRef] [PubMed]
4. Ivanescu, A.E.; Li, P.; George, B.; Brown, A.W.; Keith, S.W.; Raju, D.; Allison, D.B. The importance of prediction model validation and assessment in obesity and nutrition research. *Int. J. Obes.* **2016**, *40*, 887–894. [CrossRef] [PubMed]
5. Wallach, I.; Heifets, A. Most ligand-based classification benchmarks reward memorization rather than generalization. *J. Chem. Inf. Model.* **2018**, *58*, 916–932. [CrossRef]
6. Ballester, P.J.; Mitchell, J.B. Comments on "leave-cluster-out cross-validation is appropriate for scoring functions derived from diverse protein data sets": Significance for the validation of scoring functions. *J. Chem. Inf. Model.* **2011**, *51*, 1739–1741. [CrossRef]
7. Cortés-Ciriano, I.; Van Westen, G.J.; Bouvier, G.; Nilges, M.; Overington, J.P.; Bender, A.; Malliavin, T.E. Improved large-scale prediction of growth inhibition patterns using the NCI60 cancer cell line panel. *Bioinformatics* **2016**, *32*, 85–95. [CrossRef]
8. Li, H.; Peng, J.; Sidorov, P.; Leung, Y.; Leung, K.S.; Wong, M.H.; Lu, G.; Ballester, P.J. Classical scoring functions for docking are unable to exploit large volumes of structural and interaction data. *Bioinformatics* **2019**, *35*, 3989–3995. [CrossRef] [PubMed]
9. Tran-Nguyen, V.K.; Ballester, P.J. Beware of Simple Methods for Structure-Based Virtual Screening: The Critical Importance of Broader Comparisons. *J. Chem. Inf. Model.* **2023**. [CrossRef]
10. Karim, M.R.; Beyan, O.; Zappa, A.; Costa, I.G.; Rebholz-Schuhmann, D.; Cochez, M.; Decker, S. Deep learning-based clustering approaches for bioinformatics. *Brief. Bioinform.* **2021**, *22*, 393–415. [CrossRef]
11. Prada Gori, D.N.; Llanos, M.A.; Bellera, C.L.; Talevi, A.; Alberca, L.N. iRaPCA and SOMoC: Development and Validation of Web Applications for New Approaches for the Clustering of Small Molecules. *J. Chem. Inf. Model.* **2022**, *62*, 2987–2998. [CrossRef] [PubMed]
12. Backman, T.W.; Cao, Y.; Girke, T. ChemMine tools: An online service for analyzing and clustering small molecules. *Nucleic Acids Res.* **2011**, *39*, W486–W491. [CrossRef] [PubMed]
13. Ferreira, L.T.; Borba, J.V.; Moreira-Filho, J.T.; Rimoldi, A.; Andrade, C.H.; Costa, F.T.M. QSAR-based virtual screening of natural products database for identification of potent antimalarial hits. *Biomolecules* **2021**, *11*, 459. [CrossRef] [PubMed]
14. González-Medina, M.; Prieto-Martínez, F.D.; Owen, J.R.; Medina-Franco, J.L. Consensus diversity plots: A global diversity analysis of chemical libraries. *J. Cheminform.* **2016**, *8*, 1–11. [CrossRef] [PubMed]
15. Hadipour, H.; Liu, C.; Davis, R.; Cardona, S.T.; Hu, P. Deep clustering of small molecules at large-scale via variational autoencoder embedding and K-means. *BMC Bioinform.* **2022**, *23*, 1–22. [CrossRef]
16. Monga, M.; Sausville, E.A. Developmental therapeutics program at the NCI: Molecular target and drug discovery process. *Leukemia* **2002**, *16*, 520–526. [CrossRef]

17. Hernández-Hernández, S.; Vishwakarma, S.; Ballester, P. Conformal prediction of small-molecule drug resistance in cancer cell lines. In Proceedings of the Conformal and Probabilistic Prediction with Applications, Brighton, UK, 24–26 August 2022; pp. 92–108.
18. Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754. [CrossRef]
19. Peón, A.; Naulaerts, S.; Ballester, P.J. Predicting the reliability of drug-target interaction predictions with maximum coverage of target space. *Sci. Rep.* **2017**, *7*, 1–11. [CrossRef]
20. O'Boyle, N.M.; Sayle, R.A. Comparing structural fingerprints using a literature-based similarity benchmark. *J. Cheminform.* **2016**, *8*, 1–14. [CrossRef]
21. Zagidullin, B.; Wang, Z.; Guan, Y.; Pitkänen, E.; Tang, J. Comparative analysis of molecular fingerprints in prediction of drug combination effects. *Brief. Bioinform.* **2021**, *22*, bbab291. [CrossRef]
22. Siramshetty, V.B.; Shah, P.; Kerns, E.; Nguyen, K.; Yu, K.R.; Kabir, M.; Williams, J.; Neyra, J.; Southall, N.; Nguyen, Đ.T.; et al. Retrospective assessment of rat liver microsomal stability at NCATS: Data and QSAR models. *Sci. Rep.* **2020**, *10*, 1–14. [CrossRef] [PubMed]
23. Cortés-Ciriano, I.; Škuta, C.; Bender, A.; Svozil, D. QSAR-derived affinity fingerprints (part 2): Modeling performance for potency prediction. *J. Cheminform.* **2020**, *12*, 1–17. [CrossRef] [PubMed]
24. Škuta, C.; Cortés-Ciriano, I.; Dehaen, W.; Kříž, P.; van Westen, G.J.; Tetko, I.V.; Bender, A.; Svozil, D. QSAR-derived affinity fingerprints (part 1): Fingerprint construction and modeling performance for similarity searching, bioactivity classification and scaffold hopping. *J. Cheminform.* **2020**, *12*, 1–16. [CrossRef]
25. Nielsen, F. Hierarchical clustering. In *Introduction to HPC with MPI for Data Science*; Springer: New York, NY, USA, 2016; pp. 195–211.
26. Murtagh, F.; Contreras, P. Algorithms for hierarchical clustering: An overview, II. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2017**, *7*, e1219. [CrossRef]
27. Landrum, G. RDKit: Open-Source Cheminformatics. Available online: https://www.rdkit.org (accessed on 1 November 2022). [CrossRef]
28. Butina, D. Unsupervised data base clustering based on daylight's fingerprint and Tanimoto similarity: A fast and automated way to cluster small and large data sets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 747–750. [CrossRef]
29. McInnes, L.; Healy, J.; Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv* **2018**, arXiv:1802.03426.
30. Le, H.M.; Kumar, S.; May, N.; Martinez-Baez, E.; Sundararaman, R.; Krishnamoorthy, B.; Clark, A.E. Behavior of Linear and Nonlinear Dimensionality Reduction for Collective Variable Identification of Small Molecule Solution-Phase Reactions. *J. Chem. Theory Comput.* **2022**, *18*, 1286–1296. [CrossRef]
31. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
32. Rousseeuw, P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65. [CrossRef]
33. Caliński, T.; Harabasz, J. A dendrite method for cluster analysis. *Commun. Stat.-Theory Methods* **1974**, *3*, 1–27. [CrossRef]
34. Davies, D.L.; Bouldin, D.W. A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell.* **1979**, *PAMI-1*, 224–227. [CrossRef]
35. Almeida, J.; Barbosa, L.; Pais, A.; Formosinho, S. Improving hierarchical cluster analysis: A new method with outlier detection and automatic clustering. *Chemom. Intell. Lab. Syst.* **2007**, *87*, 208–217. [CrossRef]
36. Bengfort, B.; Bilbro, R. Yellowbrick: Machine Learning Visualization. 2017. Available online: http://www.scikit-yb.org/ (accessed on 1 February 2023).
37. Zhao, Y.; Nasrullah, Z.; Li, Z. PyOD: A Python Toolbox for Scalable Outlier Detection. *J. Mach. Learn. Res.* **2019**, *20*, 1–7.
38. Ballester, P.J. Selecting machine-learning scoring functions for structure-based virtual screening. *Drug Discov. Today Technol.* **2019**, *32*, 81–87. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

C. LDMO-CCP paper

Saiveth Hernández-Hernández, Qianrong Guo, and Pedro J Ballester. "Conformal prediction of molecule-induced cancer cell growth inhibition challenged by strong distribution shifts". In: *Preprint from bioRxiv* (2024), pp. 1–16, https://doi.org/10.1101/2024.03.15.585269, *Under review.*

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.15.585269; this version posted March 17, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.15.585269; this version posted March 17, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

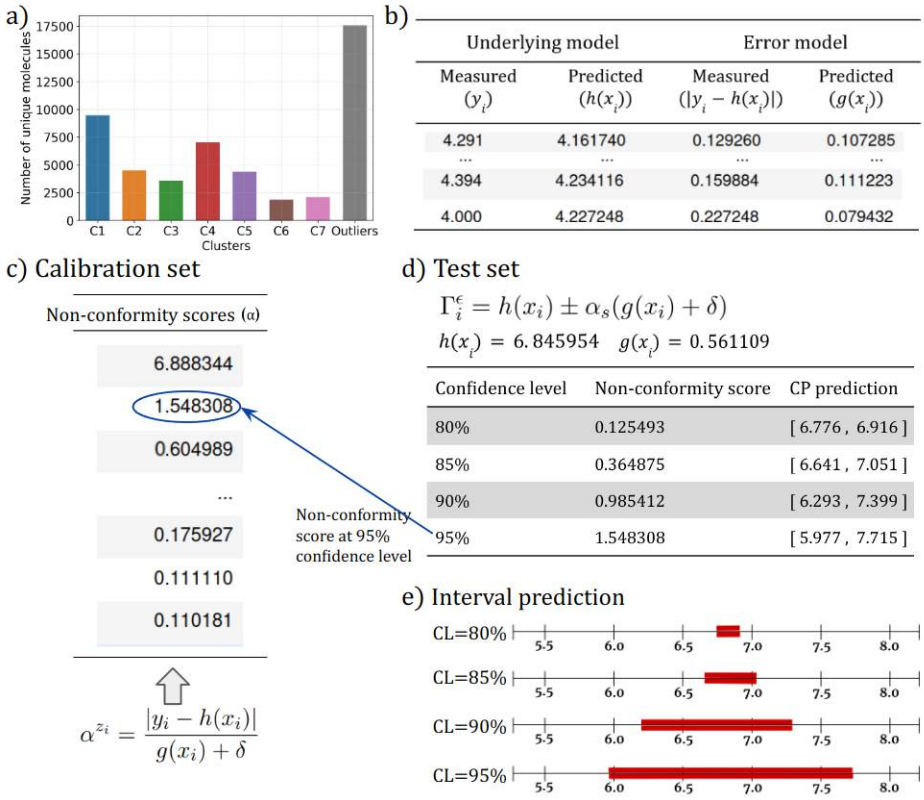


Fig. 1 Conformal Prediction enhances an underlying model with reliability estimates at a given confidence level. *A) Clusters used to build a CP model (LDMO-CV) per cell line. At each fold, 5 clusters are used as a proper training set, 1 cluster as a calibration set, and the remaining 1 cluster is used 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 $uncertainty = 2\alpha_s g(x_i)$*

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.15.585269; this version posted March 17, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.15.585269; this version posted March 17, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

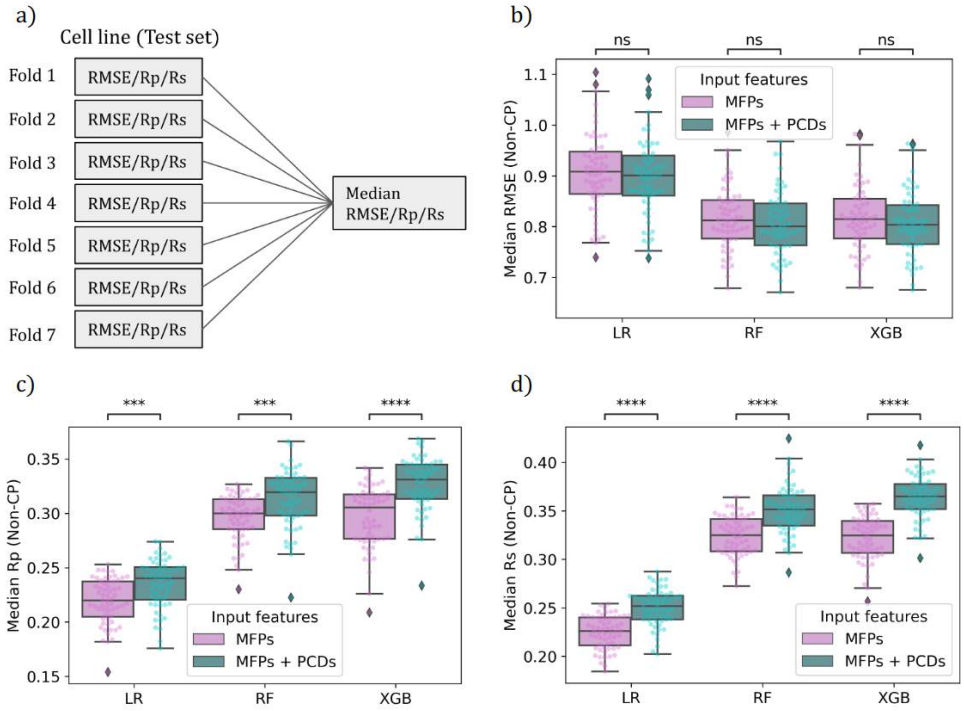


Fig. 2 The three underlying models improve when the PCDs are included as part of the input features (MFPs+PCDs). *Each molecule is encoded with an MFP of 1024 bits and 7 physicochemical descriptors. LR, RF, and XGB were used to evaluate the performance of each feature set (MFPs and MFPs+PCDs, identified 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

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.15.585269; this version posted March 17, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

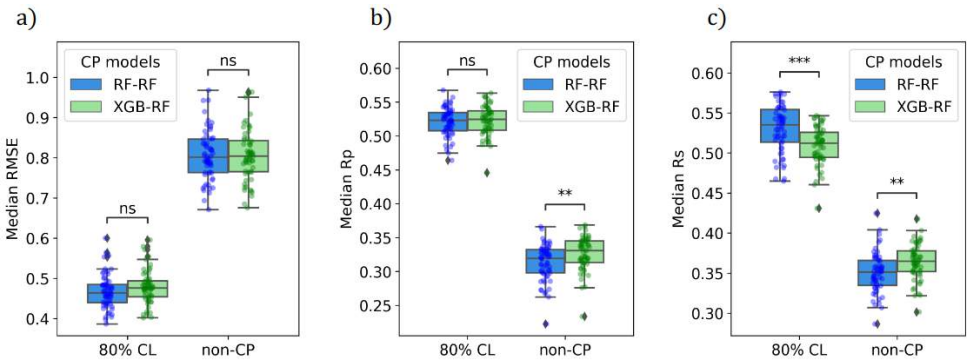


Fig. 3 CP-valid predictions exhibit lower errors and higher correlations than non-CP predictions. *RMSE, Rp, and Rs values were computed between the observed and predicted pGI_{50} values using either RF-RF or XGB-RF CP models. Each boxplot contains 60 points that represent the 7-fold-median a) RMSE, 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

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.15.585269; this version posted March 17, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.15.585269; this version posted March 17, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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).

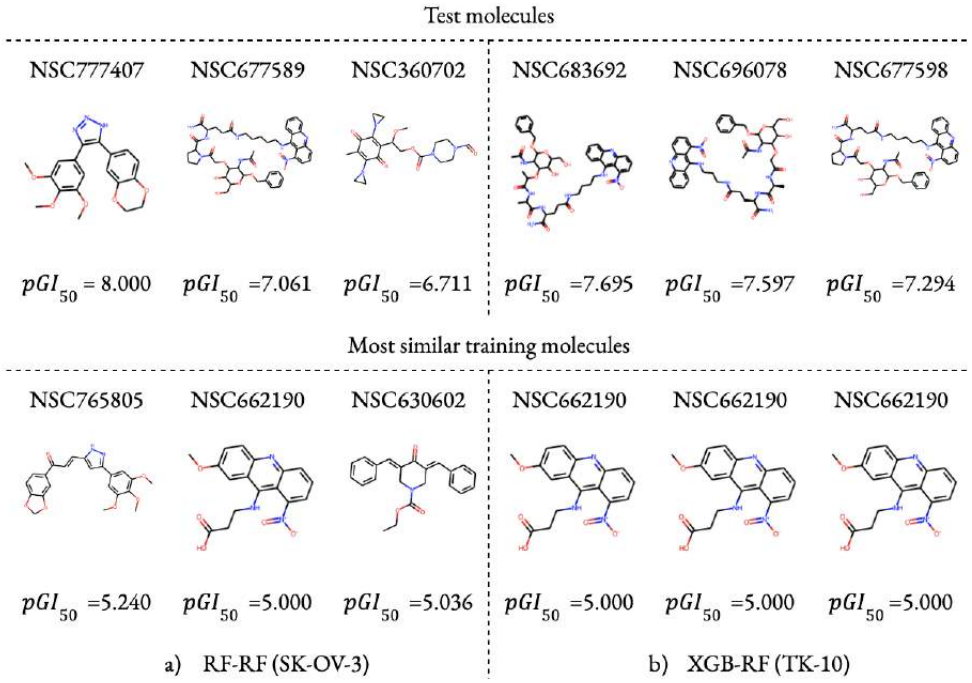


Fig. 4 ***Chemical structures of the three molecules with the highest measured pGI50s among the top 100 CP-valid predictions in the test set along with their most similar training molecules (non-outlier dataset).*** *Test molecules are shown at the top, while their most similar training molecules are 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

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.15.585269; this version posted March 17, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

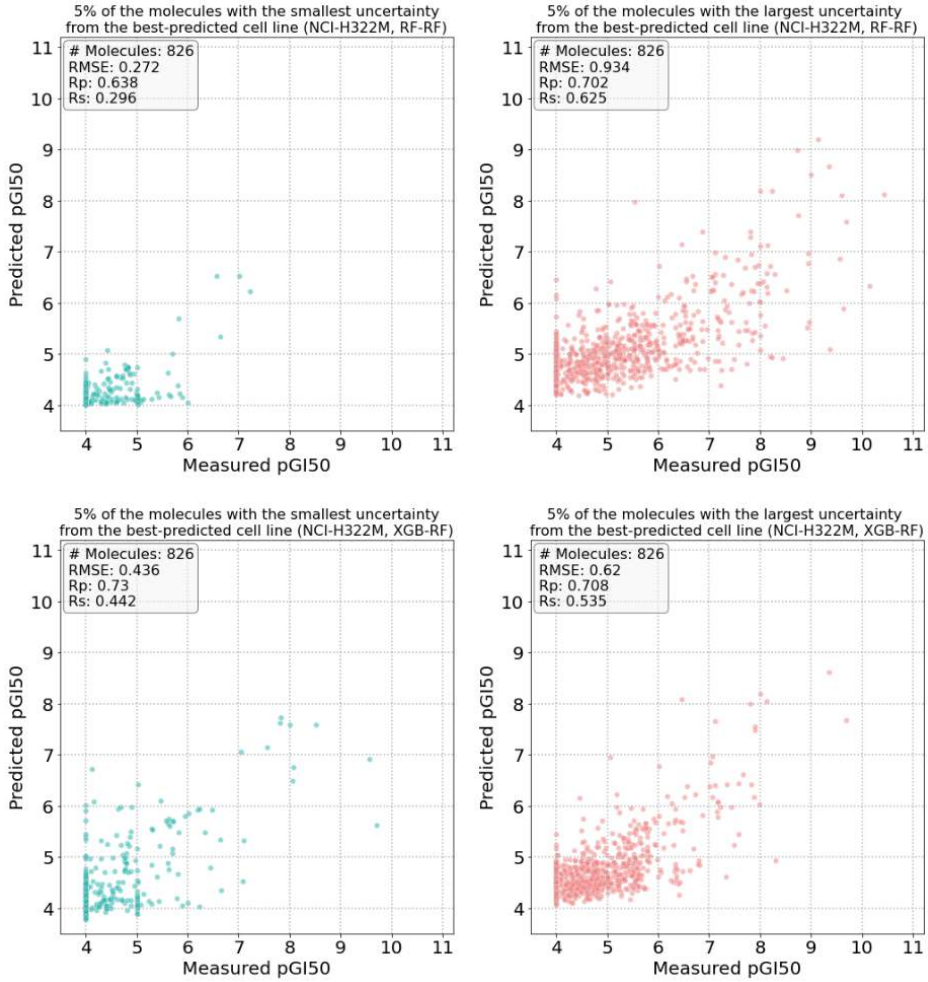


Fig. 5 The prediction of the pGI_{50} of molecule-cell line pairs improves when CP is used. *Observed and predicted pGI_{50} value in the best-predicted cell line, obtained by the RF-RF (top) and the XGB-RF (bottom) CP models. Predictions are obtained at an 80% confidence level. The uncertainty is calculated 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.*

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.15.585269; this version posted March 17, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.15.585269; this version posted March 17, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

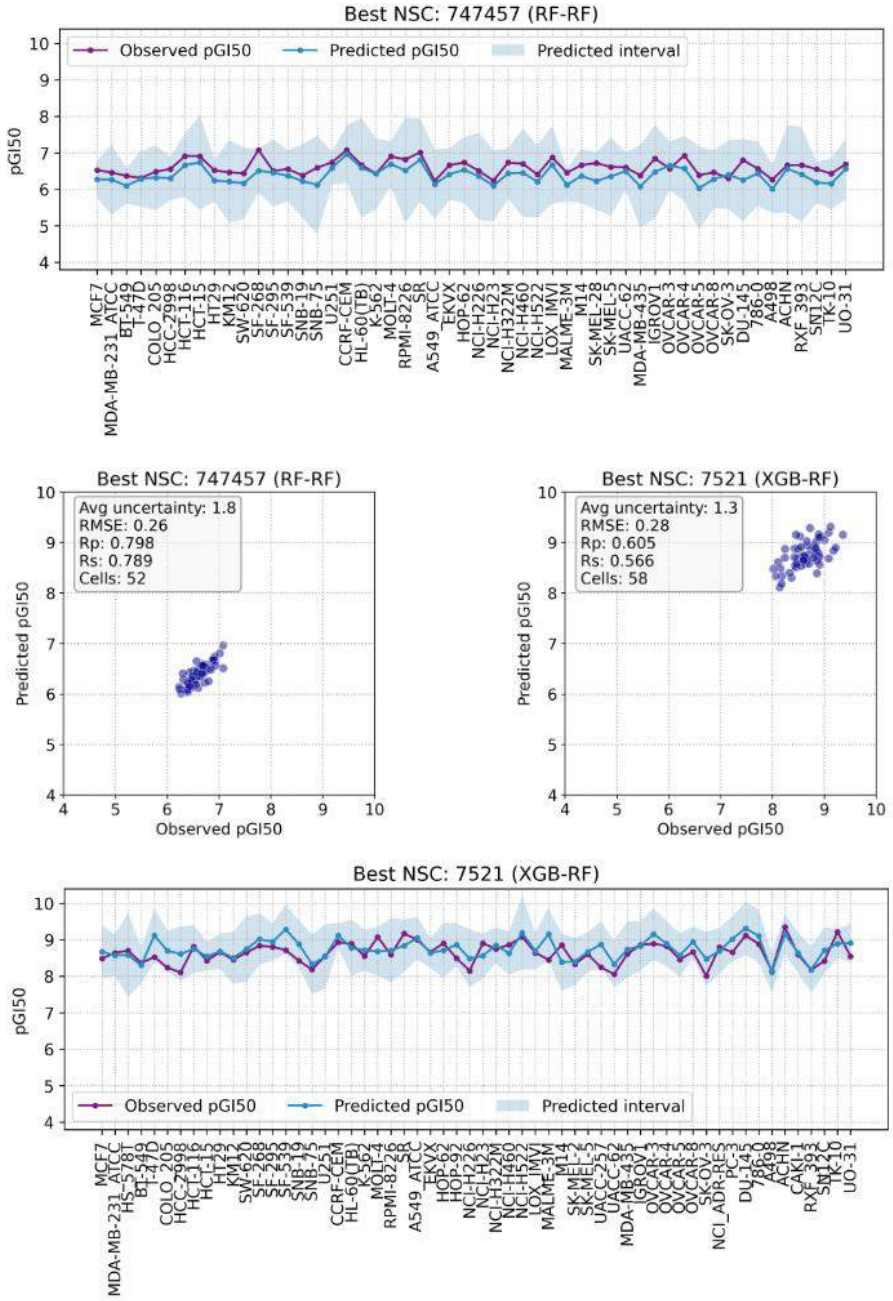


Fig. 6 The ensemble of CP models can predict the selectivity of a dissimilar molecule across NCI-60 cell lines. *Relationship between the measured and predicted pGI_{50} values across different cells, for the best-performing 80% CP-valid predictions. The analysis prioritizes the identification of molecules 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*

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.15.585269; this version posted March 17, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.15.585269; this version posted March 17, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.15.585269; this version posted March 17, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.15.585269; this version posted March 17, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.15.585269; this version posted March 17, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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).

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.15.585269; this version posted March 17, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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