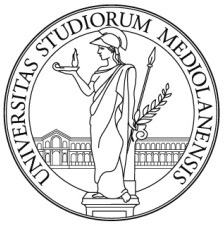

UNIVERSITÀ DEGLI STUDI DI MILANO

Doctoral School in Pharmaceutical Sciences - XXXVII Cycle
Department of Pharmaceutical Sciences

Artificial Intelligence (AI) Approaches in Drug Discovery: Development and Validation of New Strategies in Virtual Screening Campaigns

CHEM-07/A

Emanuela SABATO

Tutor: Prof. Alessandro PEDRETTI
Co-Tutor: Dott.ssa Angelica MAZZOLARI
Coordinator: Prof. Giulio VISTOLI

Academic year 2023/2024

Abstract

The field of drug discovery is increasingly benefitting from the advent of artificial intelligence (AI), which has proven to enhance both the efficiency and accuracy of Computer-Aided Drug Design (CADD), while significantly reducing the associated costs and timelines. This Work examines various case studies in which AI has been applied, with the general purpose of improving the accuracy and reliability of predictive studies. The first analysed area is the xenobiotics metabolism, assessed with the support of highly accurate input data, collected from the internally curated MetaQSAR database. The glutathione conjugation was chosen to investigate the enhancing role of structure-based studies in predicting the occurrence of a single metabolic reaction with Random Forest classification models. Later, a new tool (MetaSpot) able to predict the site of metabolism for most metabolic reactions was presented, as the natural extension of the previously published MetaClass, since they demonstrated synergistic and complementary roles. As a second application, XGBoost regression models were employed to investigate the polyspecificity of ABCB transporters, aiming to overcome common drawbacks in QSAR studies. Finally, given the proven relevance of descriptors in supervised predictive modelling, novel kinds of structure-based descriptors were developed and validated. FingerInt and HyperVolume are here proposed for enriching the pool of docking-derived descriptors, avoiding the multitude of ligand chemical descriptors, far away from the purpose of Explainable Artificial Intelligence (XAI).

Contents

Abstract **i**

1 Introduction **2**

1.1 The Role of Artificial Intelligence in Drug Discovery 3

1.1.1 Explainable Artificial Intelligence (XAI): A Brief Overview . 4

1.2 Virtual Screening and Artificial Intelligence-Aided Drug Discovery . . 4

1.3 Aim of the Thesis 5

1.4 Thesis Overview 6

1.4.1 Chapter 3: Xenobiotic Metabolism Prediction 7

1.4.2 Chapter 4: PCM Approach to ABCB Transporter Class . . . 7

1.4.3 Chapter 5: Novel Structure-Based Descriptors: FingerInt and HyperVolume 8

2 Theoretical Background **9**

2.1 Introduction 9

2.2 Artificial Intelligence and Machine Learning in Drug Discovery 10

2.3 Predictive Modelling 10

2.3.1 Classification Models 11

2.3.2 Regression Models 12

2.4 Algorithms 12

2.4.1 Random Forest (RF) 13

2.4.2 XG Boost 14

2.5 How to Build a Model 15
2.5.1 DataSelection and Dataset Building 15
2.5.2 Model Building 16
2.5.3 Model Evaluation 16
2.5.3.1 Classification Models 16
2.5.3.2 Regression Models 19
2.6 Chemical Descriptors 20
2.6.1 Fingerprints 21
2.6.2 Protein Descriptors 22
2.6.3 Scoring Functions as SB Descriptors 22
2.7 Proteochemometric Modelling (PCM) 24

3 Xenobiotic Metabolism Prediction 25

3.1 Xenobiotic Metabolism 25
3.1.1 Metabolism Prediction: State of the Art 26
3.1.1.1 MetaQSAR 27
3.1.2 Sections Overview 28
3.2 Conjugation with Glutathione 28
3.2.1 Glutathione Transferase (GST) 29
3.2.1.1 GST Structure 29
3.2.1.2 GSH Conjugation Mechanism and GST Role 31
3.2.2 Aim 33
3.2.3 Methods 34
3.2.3.1 Dataset Preparation 34
3.2.3.2 Ligand-Based (LB) Feature Calculation 35
3.2.3.3 Molecular Docking Studies 36
3.2.3.4 Rescoring Procedure and Structure-Based (SB) Approach 37

3.2.3.5 Model Building and Validation . . . 38
3.2.3.6 Hardness and Predictive Power . . . 38
3.2.4 Results and Discussion . . . 39
3.2.4.1 LB Predictive Models . . . 39
3.2.4.2 Molecular Docking Studies . . . 42
3.2.4.3 SB Predictive Models . . . 42
3.2.4.4 Hardness and Predictive Power . . . 45
3.2.5 Conclusions . . . 46
3.3 MetaSpot . . . 47
3.3.1 Introduction . . . 47
3.3.2 Methods . . . 48
3.3.2.1 Utilised Metabolic Data . . . 48
3.3.2.2 Generation of Models with Atom Typing (First Round) 48
3.3.2.3 Generation of Models Without Atom Typing (Second Round) . . . 49
3.3.3 Results . . . 50
3.3.3.1 MetaQSAR-Based Datasets . . . 50
3.3.3.2 Prediction of the Reactive Atoms for the Reaction Classes 50
3.3.3.3 Prediction of the Reactive Atoms for the Reaction Subclasses . . . 54
3.3.3.4 Discussion . . . 57
3.3.4 Conclusion . . . 60
3.3.5 Supplementary Information . . . 62

4 PCM Approach to ABCB Transporter Class 75

4.1 ABC Transporters . . . 75
4.1.1 ABC Transporters Structure . . . 76
4.1.2 Overview of the Diverse Subfamilies . . . 77
4.1.2.1 ABCB Subfamily Transporters . . . 78

4.2 Aim . . . 80
4.3 Methods . . . 80
4.3.1 Data Collection . . . 81
4.3.2 Target Selection . . . 84
4.3.3 Ligand Descriptors . . . 85
4.3.4 Target Descriptors . . . 85
4.3.5 Proteochemometric Modelling . . . 85
4.4 Results and Discussion . . . 86
4.4.1 Chemical Space Analysis . . . 86
4.4.2 Predictive Models . . . 88
4.4.2.1 10-fold Cross Validation on ChEMBL31 . . . 88
4.4.2.2 10-fold Cross Validation on ChEMBL33 . . . 90
4.4.2.3 External Validation: DeltaChEMBL33 Dataset . . . 92
4.4.3 Feature Importance Analysis . . . 94
4.5 Conclusions . . . 97

5 Novel Structure-Based Descriptors: FingerInt and HyperVolume 99

5.1 Interaction Fingerprints (IFP) . . . 99
5.2 Volume-Based Descriptors . . . 101
5.3 Aim . . . 103
5.4 Methods . . . 104
5.4.1 FingerInt (FI) . . . 104
5.4.2 HyperVolume (HV) . . . 105
5.4.3 Datasets Building . . . 106
5.4.4 Descriptors Calculation . . . 107
5.4.5 Model Building and Evaluation . . . 108
5.5 Results and Discussion . . . 108

5.5.1 Preliminary Benchmark (Experimental Data) 108
5.5.2 Extended Benchmark (Actives and Decoys) 114
5.6 Conclusion 115

6 Conclusions and perspectives **116**

Appendix **120**

Bibliography **175**

List of Figures

1.1 "AI +chem" literature collection 2000-2024. 2

2.1 Supervised learning workflow. 11

2.2 Random Forest scheme. 14

2.3 Confusion matrix for a binary classification. 17

2.4 Example of ROC curve. 18

3.1 Topology and structural representations of cytosolic GST. 30

3.2 Topology and structural representations of mitochondrial GST. 30

3.3 Structure of the glutathione (GSH). 31

3.4 Proposed mechanism of non-enzymatic GSH conjugation. 32

3.5 Glutathione conjugation scheme. 32

3.6 GST catalytic mechanism. 33

3.7 Scheme of the metabolic pathway of acetaminophen. 34

3.8 SB models: impact of FS on MCC value. 43

3.9 Cumulative average showing the correlation between ELUMO and predictive power. 45

3.10 Model performances for the metabolic classes in the two round of predictions. 51

3.11 Occurrences of the descriptors for metabolic classes in the two rounds of predictions. 52

3.12 Model performances for the metabolic subclasses in the two round of predictions. 55

3.13 Occurrences of the descriptors for metabolic subclasses in the two rounds of predictions. 56

3.14 MetaClass and MetaSpot: comparison of the performances. 58

4.1 Distribution of human ABC transporters. 76

4.2 ABC transporter domain core. 76

4.3 Heatmap of Activity value vs HUGO. 82

4.4 Heatmap of Organism vs HUGO. 82

4.5 Data reduction histograms. 84

4.6 PCA plot of compounds chemical space of ChEMBL33 dataset (matrix1). 86

4.7 t-SNE plot of compounds chemical space of ChEMBL33 dataset (matrix1). 87

4.8 PCA plot of ligand and protein information of ChEMBL33 dataset (matrix4). 88

4.9 Outlier analysis – CV ChEMBL31 model (matrix1). 90

4.10 Outlier analysis – CV ChEMBL33 model (matrix1). 91

4.11 Outlier analysis – DeltaChEMBL33 model (matrix1). 93

4.12 Scatter plot of outliers, with 2D structure of major ones. 94

4.13 Top 20 feature importance gain-based of the model tested on ChEMBL31 CV model (matrix4). 95

4.14 Top 20 feature importance gain-based of the model tested on ChEMBL33 CV model (matrix4). 95

4.15 Top 20 feature importance gain-based of the model tested on DeltaChEMBL33 (matrix4). 96

5.1 MCC values for each experimental receptor (BP approach). 110

5.2 MCC values for each experimental receptor (AV approach). 111

List of Tables

3.1 LB descriptor set. 35

3.2 Characterisation of the selected crystal structures. 36

3.3 LB model performance w/o FS step. 40

3.4 LB model performance with FS step. 40

3.5 LB model performance: MCC value comparison with and w/o FS. . . 41

3.6 LB models: Set size with and w/o FS step. 41

3.7 LB global model performance, with and w/o FS. 42

3.8 SB models: consensus voting results. 44

3.9 Performance metrics for the 17 considered classes of metabolic reactions in the first round of predictions. 63

3.10 Descriptors selected by the feature importance analyses for the 17 considered classes of metabolic reactions in the first round of predictions 64

3.11 Performance metrics for the 17 considered classes of metabolic reactions in the second round of predictions. 65

3.12 Descriptors selected by the feature importance analyses for the 17 considered classes of metabolic reactions in the second round of predictions 67

3.13 Performance metrics for the considered subclasses of metabolic reactions in the first round of predictions. 69

3.14 Descriptors selected by the feature importance analyses for the considered subclasses of metabolic reactions in the first round of predictions 70

3.15 Performance metrics for the considered subclasses of metabolic reactions in the second round of predictions. 72

3.16 Descriptors selected by the feature importance analyses for the considered subclasses of metabolic reactions in the second round of predictions 74

4.1 PCM models in 10-fold CV on ChEMBL31 dataset. 89

4.2 PCM models in 10-fold CV on ChEMBL33 dataset. 91

4.3 PCM models tested on DeltaChEMBL33. 92

5.1 Model performance (mean MCC values) of experimental set. 109

5.2 FI descriptors impact in 18 out of 25 receptors, in AV approach. . . . 109

5.3 HV descriptors impact in 12 out of 25 receptors in BP approach. . . . 110

5.4 Model performance of experimental set: ProLIF comparison. 112

5.5 Model performance of experimental set (22 out of 25 receptors): ProLIF comparison. 112

5.6 Model performance of experimental set: PLEC comparison. 113

5.7 Model performance (mean MCC values) of extended set. 114

Chapter 1

Introduction

The advent of Artificial Intelligence (AI) has precipitated a deep transformation across diverse spheres, including everyday life, professional environments, and scientific research. In particular, chemistry and drug discovery fields have experienced powerful advancements due to the integration of AI technologies with the already known and used techniques and expertise, as highlighted also by a recent work depicted in Figure 1.1. The application of AI algorithms and data-driven models has dramatically enhanced the efficiency and efficacy of drug development processes, leading to the accelerated discovery and optimisation of therapeutic compounds.[1]

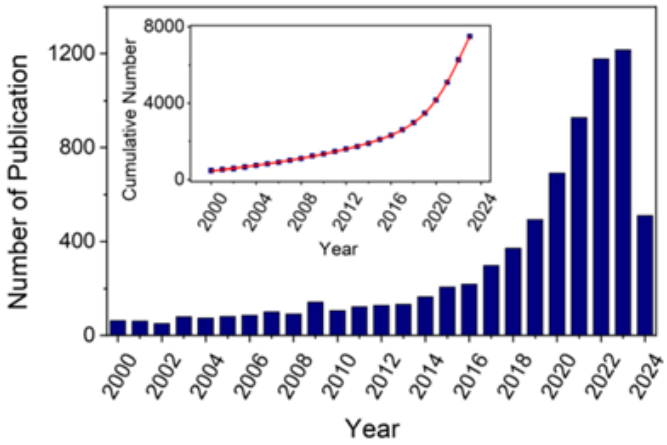

Figure 1.1: Yang *et al.* collected published literature from 2000 to 2024 on "AI + chem" on WoS platform. In the chart, the increasing number of publications on "AI + chem" per year is shown. Reproduced from [2]

Historically, the process of discovering and developing new drugs has been an arduous, time-consuming, and costly effort, often taking over a decade and costing

billions of dollars to bring a single drug to market. However, AI and ML are rapidly changing this scenario by providing tools that may significantly speed up the identification and optimisation of drug candidates. These technologies have not only accelerated various stages of drug discovery, but have also introduced new possibilities that were previously inconceivable.[3]

1.1 The Role of Artificial Intelligence in Drug Discovery

Artificial Intelligence has become an integral component of modern drug discovery, enabling the analysis of vast amounts of data and patterns identification, and making predictions with a high level of accuracy. The integration of AI into drug discovery is aided by several key factors, including the increasing availability of large-scale biological and chemical datasets, advancement in computational power, and the development of *ad-hoc* algorithms capable of processing complex information.[4]

One of the most significant contributions of AI in drug discovery is its ability to enhance the performance of diverse processes, such as target identification, hit discovery, lead optimisation, and preclinical trials. For instance, AI models can evaluate the drug-likeness of a molecule, predict if a molecule can undergo a certain reaction through the evaluation of their potential toxicity, predict molecular basis of protein-ligand interactions or search for novel potential drugs based on virtual screening campaigns. Moreover, AI can assist in the design of novel molecules with desired properties by generating new chemical structures for specific therapeutic targets.

The adoption of AI in drug discovery has grown exponentially over the past decade. According to a report by MarketandMarkets, the AI in drug discovery market was valued at approximately $0.9 billion in 2023 and is expected to reach $4.9 billion by 2028, growing at a compound annual growth rate (CAGR) of 40.2% from 2023 to 2028. [5] This rapid growth is due to the increasing application of AI technologies in the pharmaceutical industry.

1.1.1 Explainable Artificial Intelligence (XAI): A Brief Overview

Despite the undoubted benefits and advantages of AI in drug discovery, one of the main remaining challenges is the "black box" nature of many AI models. These models, particularly those based on Deep Learning (DL), are often highly complex and provide minimal, if any, insight into how their predictions are made. This lack of interpretability can be problematic, especially in a critical field as drug discovery, where understanding the underlying mechanisms of AI-driven decisions is a crucial step for ensuring the safety and effectiveness of new drugs.[6]

Explainable Artificial Intelligence (XAI) addresses this challenge by providing tools and techniques that make AI models clearer and more interpretable. XAI enables researchers to understand the decision-making processes of AI models, to identify potential biases, inaccuracies or limits, and to validate the scientific reliability of the predictions. This interpretability is essential for gaining trust in AI-driven insights and for discovering novel scientific knowledge.[7] Among the areas where AI and XAI have been successfully exploited, there are the design and optimisation of drug candidates, as well as the prediction of drug safety and toxicity.

1.2 Virtual Screening and Artificial Intelligence-Aided Drug Discovery

Virtual screening (VS) is a pivotal computational technique used in drug discovery to identify potential drug candidates from large chemical libraries. Traditionally, VS aims to overcome the high cost, the long time and the low success rate of experimental high-throughput screenings (HTS), in order to promote the development of new drugs. Indeed, VS acts as an enriching filter to extract from the considered database a focused subset which should comprise an increased number of active compounds thus rendering the following experimental screening faster and less expensive.[8]

Commonly, VS relies on methods, such as molecular docking and quantitative

structure-activity relationship (QSAR) models, to predict the feasibility of a compound binding to a target protein.[9] However, these methods have some weaknesses, especially in terms of their ability to capture the complex and nonlinear relationships, that often exist between molecular structures and biological activity.[10][11]

Artificial Intelligence has deeply enhanced VS studies by introducing more sophisticated and accurate prediction models, that can overcome the limitations of traditional methods. Employing large datasets of chemical compounds and biological targets, AI-aided virtual screenings are able to train models that can predict the activity, the binding affinity and the toxicity of compounds with high accuracy. Since these models are capable of identifying the relationships between molecular features and biological effects, they can lead to more reliable predictions and more efficient screening processes.[12]

AI models rely their predictive power on the extraction and exploitation of relevant molecular features that can accurately represent and distinguish between different chemical entities. By transforming complex molecular structures into quantifiable features, descriptors can encode the structural, topological and electronic characteristics of compounds.

One of the most significant advantages of AI-aided virtual screening is its ability to reduce the number of false positives and false negatives, thereby increasing the overall efficiency of the drug discovery process.

The impact of AI-aided virtual screening on drug discovery is evident from the growing number of successful applications and case studies. On this subject, a recent study [13] demonstrated that AI models can reduce the time and cost of pharmaceutical investigations by accelerating computational phases and improving hit identification accuracy by 10-15%, compared to classical methods.

1.3 Aim of the Thesis

The research project reported in this Thesis focuses on the development of different kinds of predicting models, analysing some case studies of biological targets, with their own peculiar interaction patterns.

The rapidly growing availability of chemical and biological data, added to the

increasing computational power, has made Artificial Intelligence an essential tool in modern pharmaceutical research. The primary aim of this Thesis is to explore and advance the application of AI techniques in the drug discovery field, with deep attention on developing predictive models that can accurately identify potential drug candidates. A critical aspect of this work is the selection of accurate and appropriate data, as the model's performance is directly related to the quality of the considered input data. By careful curation and pre-processing phases, this research seeks to minimise errors and biases that could compromise the predictive accuracy.

Indeed, there are significant criticisms associated with the accuracy of dataset building, particularly when non-experimental data, *i.e.* decoys, are used in the absence of experimentally validated ones. Decoys are synthetic or hypothetical compounds that are designed to resemble inactive compounds, but they are not experimentally validated yet. It is worth noticing that relying on decoys has many issues and challenges requiring meticulous and expert management, and the possibility of compromising the accuracy of the generated models must be carefully considered and handled.

Among the drawbacks of using decoys in AI models, there is the intrinsic uncertainty in the data, since they may not accurately describe the true physicochemical properties and biological behaviour of real inactive compounds. This can lead to the generation of confusing patterns during the training process, and, as a result, models might exhibit high-performance metrics during validation but fail to generalise to real-world scenarios. Another risk is represented by overfitting, which occurs because the model becomes extremely sensitive to the specific features of decoys and their applicability domain.[14]

One of the underlying reasons behind the project regards the inactive compound issue. Indeed, the main ambition has been a careful data selection step and the enhancement of the input data accuracy, using, whenever permitted, experimentally validated data to develop AI models as reliably as possible.

1.4 Thesis Overview

In order to finely investigate the contribution of AI techniques to diverse drug discovery phases, some case studies were examined, by means of different approaches,

enhancing the accuracy of considered input data and aspiring to more reliable predictive models.

Chapter 2 reports an overview of the main computational methods employed in this Thesis, providing particular insight into supervised predictive models and their principal features.

1.4.1 Chapter 3: Xenobiotic Metabolism Prediction

Xenobiotics are exogenous compounds, that the body metabolises into more hydrophilic substances to facilitate their excretion.

Chapter 3 provides new insights into the xenobiotic metabolism prediction. The work started with the investigation of a peculiar metabolism reaction, the conjugation with glutathione (GSH) (Section 3.2). This phase II reaction detoxifies electrophile compounds, and it is typically catalysed by glutathione S-transferase, while sometimes the spontaneous process occurs too. Traditionally, the study of drug-induced toxicity has relied on the identification of structural alerts (toxicophores), with related drawbacks, here prevented thanks to curated and appropriate input data (MetaQSAR-based). In order to improve drug safety assessment, the GSH conjugation was investigated, initially through ligand-based studies, and then the structural contribution of the enzyme was assessed.

With the aim to generalise and explore the whole plethora of metabolic reactions, in Section 3.3 a novel tool to recognise the specific sites of metabolism (SoM) is presented. Indeed, MetaSpot involves the development of SoM predictive models for all reaction classes and subclasses.

1.4.2 Chapter 4: PCM Approach to ABCB Transporter Class

The research activity described in Chapter 4 was performed during my seven-months Visiting period at the University of Vienna, under the scientific supervision

of Prof. Gerhard F. Ecker.

ABCB transporters constitute an important subfamily of transporter proteins because of their involvement in many homeostasis processes, detoxification and drug resistance phenomena. Given their crucial roles in human physiopathology, this work addresses the challenge of their multispecificity modelling, thanks to a proteochemometric (PCM) approach. It posed their main basis on the recent release of resolved structures, that represent a core aspect of the method, besides experimentally known ligands and protein-ligand descriptors.

1.4.3 Chapter 5: Novel Structure-Based Descriptors: FingerInt and HyperVolume

Molecular recognition, involving specific and stabilising interactions between small molecules and their biological targets, is crucial in living organisms. Understanding the mechanisms of protein-ligand recognition and binding is essential in drug discovery. Among various kinds of molecular descriptors and fingerprints to be used in predictive modelling, the protein-ligand interaction fingerprints provide valuable insights into interaction pattern studies.[15]

Based on these grounds and with the aim to propose new tools to better describe protein-ligand interactions in docking studies, Chapter 5 proposes two novel types of descriptors: some interaction fingerprints, called FingerInt and a volume-based set of descriptors, called HyperVolume descriptors.

Chapter 2

Theoretical Background

2.1 Introduction

This Chapter illustrates the main methodologies used in this Thesis work.

The focus is the development of new predictive models applied to some case studies, so the basis and the main features of machine learning prediction will be explained.

In this context, a detailed analysis is dedicated to chemical descriptors and fingerprints, which are crucial in cheminformatics for processing and analysing molecular structures.

At the end, an advanced computational technique that combines both chemical and biological data, Proteochemometric Modelling (PCM), will be presented.

2.2 Artificial Intelligence and Machine Learning in Drug Discovery

Artificial Intelligence is the ability of computer systems to act in a seemingly intelligent way, making decisions in response to new inputs to solve a given problem. In this context, machine and deep learning (ML and DL) methods allow the development of rational models by which the machine learns how to complete defined tasks, based on statistical analysis and pattern recognition of the raw data input. Generally speaking, AI-based algorithms can be applied for regression, classification, pattern recognition, similarity analyses, and clustering studies.[3]

According to Arthur Samuel (1959), Machine Learning (ML) is defined as a "field of study that gives computers the ability to learn without being explicitly programmed".

ML is a growing field of computational algorithms that aim to simulate human intelligence by learning from the given environment.

In the medicinal chemistry context, ML is a such suitable tool to predict the potential activity of a molecule, based on the known activity of a molecule set, *i.e.* thanks to the model ability to learn from the training set.[3]

2.3 Predictive Modelling

Predictive modelling is a subset of data mining that employs statistical techniques to predict the outcome of future events, based on input data. In the drug discovery area, this method is useful to rapidly screen large compound libraries to identify potential drug candidates, to understand the relationship between a molecule structure and its biological activity (QSAR modelling), or to optimise hit compounds through ADMET predictions. Thanks to the automatisation of many steps, predictive modelling can significantly reduce the time and costs, increasing the process efficiency, as well as results accuracy.

ML approaches for predictive modelling are divided into supervised and unsupervised learning.

Supervised machine learning considers labelled datasets aimed to train or supervise algorithms for accurate data classification and prediction (Figure 2.1). Since input data report a label, the model is able to evaluate its learning and predictive abilities and related accuracy.

On the other hand, unsupervised machine learning is capable of analysing and clustering unlabelled data, in order to find and highlight hidden associations or patterns, such as Principal Component Analysis (PCA) does.[16]

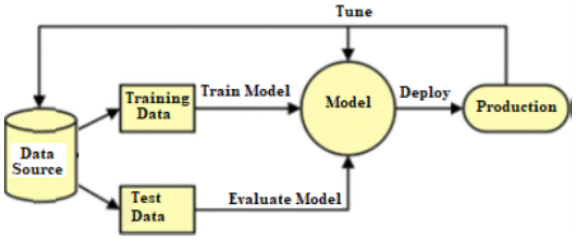

Figure 2.1: Supervised learning workflow. Reproduced from [16]

In supervised learning, since the algorithm learns from labelled data, it iteratively makes predictions and tunes for a more precise output. Although these models generally give more accurate results than unsupervised ones, they require human involvement that provides proper labels. On the other side, unsupervised models are able to detect patterns in unlabelled data in a shorter time, but human actions are required to evaluate and validate the results.

Supervised methods can rely on two different approaches, based on their main objective: classification and regression learning.

2.3.1 Classification Models

Classification models aim to predict categorical outcomes or labels based on input features. They are trained on labelled data, where each instance belongs to a class or category, usually determined through experimental evidence. Indeed, they are used to predict – or, better, classify - the discrete values in a multi-class outcome or, mostly, two-class outcome or binary mode (generically “1” and “0”). In computer-aided drug

design (CADD), they are employed to predict the biological activity of a compound (label: "active" or "inactive"), the metabolic reaction occurrence (label: "substrate" or "not substrate") or the molecule toxicity (label: "toxic" or "not toxic") or other ADMET properties.

The objective of classification models is to find a function that, learning from features of the training set, can divide the test set into the actual classes based on different parameters.[17]

2.3.2 Regression Models

Regression models aim to predict a continuous numerical outcome. They are trained on labelled data, in which every instance is characterised by a continuous variable, usually an experimental value too. In CADD, regression models are employed for quantitative prediction of some properties, such as activity (*e.g.* IC50, EC50, Ki), binding affinity, molecular features (*e.g.* logP, electrophilicity) or quantitative ADMET properties (*e.g.* solubility, permeability, clearance).

Regression model evaluation and validation differ from classification ones because the metrics are different, since the different problem faced. A detailed comparison is explained in Section 2.5.3.[17]

2.4 Algorithms

Predictive modelling relies on many different algorithms, which aim to extract useful information from input data, in order to use it to build the model. The algorithms should be able to learn and identify patterns and correlations that they will use to make predictions on new and unseen data.

Algorithms can be divided into linear and non-linear methods, based on the relationship they model between the input features and the output variable.

Linear methods assume that this relationship is represented by a straight line. These models are easily interpretable and well-performing for simple datasets, where the relationships are almost linear. Some common linear methods include linear

regression, which is used for predicting continuous numerical values, and logistic regression, used for predicting binary outcomes.[18]

Non-linear methods are able to model more complex relationships between the variables, where the output does not fluctuate linearly with the input change. They are more flexible, and they can extrapolate a wider range of relationships in the data, even if their enhanced ability increases with interpretability difficulty as well as considerable data requirement and computational power. Most employed non-linear methods include Decision Trees, Random Forest, Support Vector Machine (SVM), Gradient Boosting Machines (GBM) or XGBoost, Neural Networks (NN).[18]

2.4.1 Random Forest (RF)

Random Forest (RF) algorithm is an ensemble learning technique used in classification problems as well as regression ones. The ensemble methods combine multiple machine learning models to improve the overall predictive performance. In detail, RF employs bagging (bootstrap aggregating) to train multiple decision trees simultaneously (in parallel) on data random subsets and then it aggregates the resulting predictions through voting or averaging (Figure 2.2). For each bootstrap sample, RF builds a decision tree, creating a series of "if-else" questions to decide how to partition the data into smaller subsets. The peculiar feature of RF algorithm is that just a random subset of features is selected at each decision tree node, and the best feature is chosen for the split.

Each tree is grown without pruning, allowing the growth to its full extent. However, since the trees grow on different samples with different subsets of features, the trees will be different from each other and so the overfitting due to lack of pruning step is avoided.

In order to assess the importance of each feature in predicting the target variable, the RF algorithm uses the metric Mean Decrease Impurity (MDI). It provides a quantitative measure of how much a feature contributes to the overall model. In detail, MDI measures the reduction in impurity, the degree to which the target variable is distributed within a node. A pure node has a low impurity, while a mixed node has a high impurity level. This parameter is used to find the best split for each node (minimising the impurity of the child node) or to decide if a satisfactory

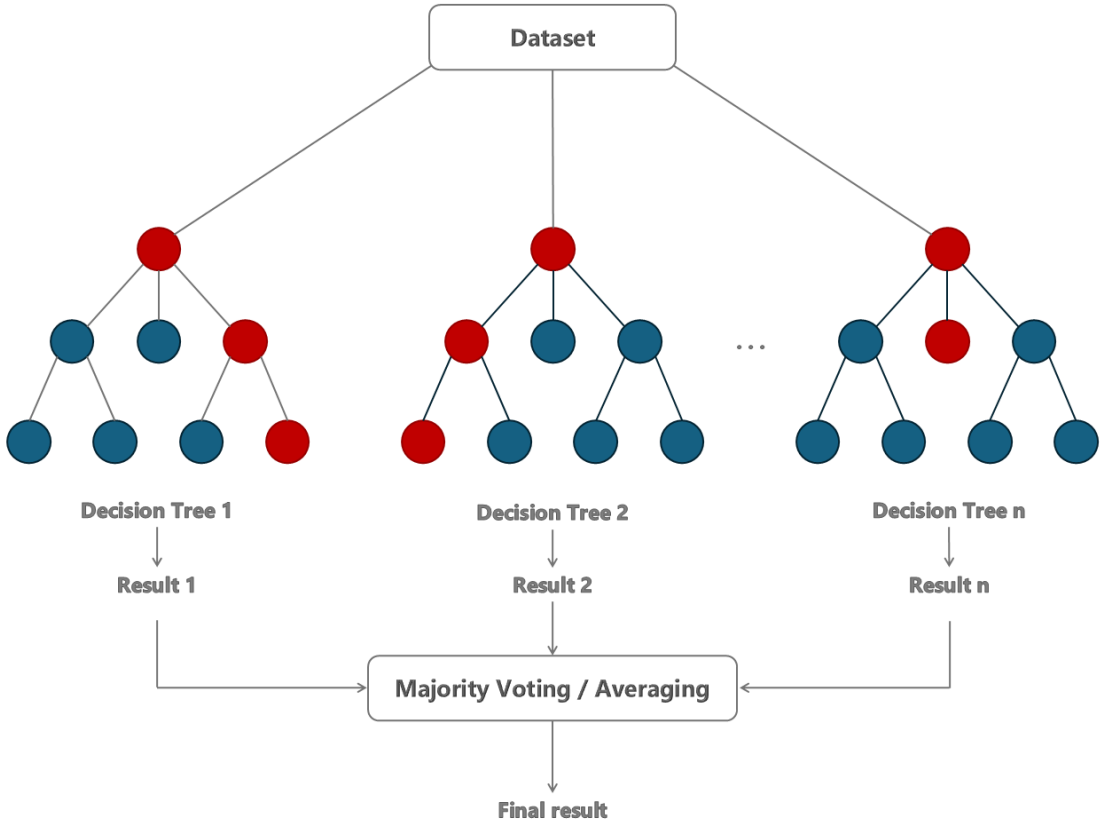

Figure 2.2: Random Forest scheme.

impurity level has been reached. The most used impurity measure is Gini impurity, which accounts for the probability of misclassifying a randomly chosen sample from the node.

RF algorithm is employed in classification models as well as in regression ones.

One of the main advantages of this algorithm is the ability to reduce the overfitting phenomenon due to its ensemble method. Furthermore, its inner feature selection at each node makes RF efficient in handling datasets with a large number of features and the diversity of trees in the forest increases the model robustness, reducing the noise (while, on the other hand, giving a low interpretability).

2.4.2 XG Boost

XGBoost (eXtreme Gradient Boosting) is an algorithm based on gradient boosting framework. It is used for both classification and regression problems, and it is known for its accuracy and ability to work with large datasets.

XGBoost is an ensemble learning algorithm, based on gradient boosting, a process where the algorithm builds a model iteratively, adding new decision trees to correct the errors of the previous one. Despite bagging, in boosting, models are trained sequentially, allowing the model to learn from its mistakes and to improve the performance iteratively.

After an initial prediction, the residual errors are computed based on the difference between the experimental value and the predicted one. Then, a new decision tree is trained to predict the residuals, extrapolating the patterns in the errors made by the previous model. The predictions are updated by adding a scaling factor (η) which monitors how much each tree contributes to the final prediction. This procedure is repeated until a convergence is reached.

Besides XGBoost common advantages such as the high accuracy, efficiency and ability to avoid overfitting phenomena, the resulting models could present a lower interpretability and a high computational cost, as all ensemble learning methods do.

2.5 How to Build a Model

2.5.1 DataSelection and Dataset Building

The first and most important step in building a predictive model is the data upon which it is based. Data selection is a key process that involves choosing relevant features that will contribute to the model's predictive power and relative accuracy, since the quality, the diversity and the volume of data have a direct impact on model performance.

The data collection can be performed by searching in available resources, such as online libraries or manually curated datasets based on literature.

After the collection step, the so obtained data should be processed under a cleaning step. It is a process aimed at detecting and deleting mistakes, duplicates or incomplete data within the dataset. Indeed, if the data is incorrect, the obtained outcomes will be unreliable. Most common operations include detecting counterions, inorganic compounds, molecules with rare elements, tautomers, unlabelled or

incongruent data, and missing values.

Concerning a global view of the feature set, in order to render homogeneous all features, when they present different scales or measure units, a standardisation step is required. In addition, correlated features should be taken into account and usually removed.

2.5.2 Model Building

Model building is the process of selecting and setting a mathematical or statistical model to learn from the dataset. This phase is highly dependent on the nature of the problem and the available data. In detail, the choice of the model is influenced by several factors, such as the complexity of the data, the problem type (classification or regression), the desired interpretability level and the available computational resources.

In this phase, hyperparameter tuning is performed. All the algorithm parameters (called hyperparameters) need to be optimised in order to achieve the best possible performance. Some common hyperparameters are the number of estimators, such as the number of trees in RF or XGBoost, the maximum depth of the decision trees or the criterion for splitting (*e.g.* minimum samples required to split a node).

2.5.3 Model Evaluation

Model evaluation is the final step in the predictive modelling process, and it is critical for assessing how the model generalises to unseen data. The evaluation metrics vary depending on the type of predictive model.

2.5.3.1 Classification Models

In a classification model, common metrics include: accuracy, precision, recall and the area under the receiver operating characteristic (ROC) curve.[19]

Figure 2.3 depicts the confusion matrix, composed of the amount of:

- True Positive (TP): instance where the model correctly predicts the positive class
- True Negative (TN): instance where the model correctly predicts the negative class
- False Positive (FP): instance where the model incorrectly predicts the positive class, when the actual class is negative
- False Negative (FN): instance where the model incorrectly predicts the negative class, when the actual class is positive.

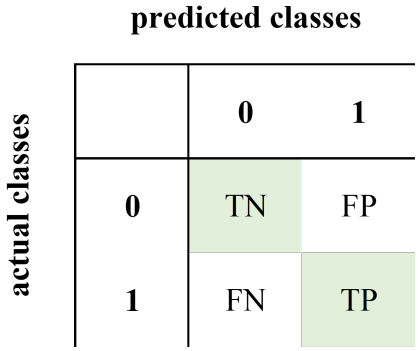

Figure 2.3: Confusion matrix for a binary classification.

Accuracy is the ratio of correctly predicted instances (sum of true positives TP and true negatives TN) to the total instances in the dataset (Equation 2.1).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.1}$$

Recall (also called Sensitivity) measures the ability of a model to find all relevant cases (TP) (Equation 2.2).

$$Recall = \frac{TP}{TP + FN} \tag{2.2}$$

Precision is the proportion of positive predictions that are actually correct (Equation 2.3).

$$Precision = \frac{TP}{TP + FP} \tag{2.3}$$

Specificity (True Negative Rate) is a measure of the TN that were correctly identified (Equation 2.4).

$$Specificity = \frac{TN}{TN + FP} \tag{2.4}$$

Area Under the Receiver Operating Characteristic Curve (AUC-ROC) measures the ability of the model to distinguish between classes. The ROC curve plots the TP rate (Recall) against the FP rate (Figure 2.4). An AUC of 0.5 indicates that there is no discrimination and the result is random, while an AUC of 1 indicates a perfect discrimination.

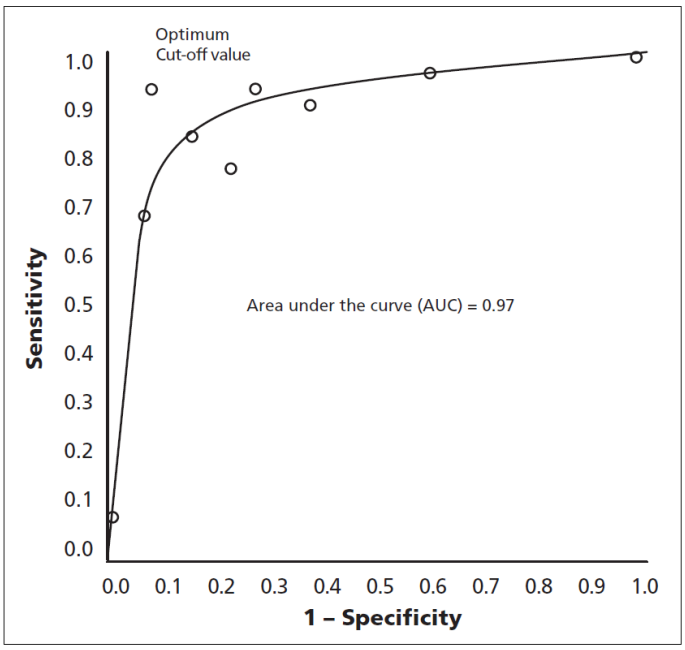

Figure 2.4: Example of ROC curve illustrating high discriminatory power. Reproduced from [20]

Matthews Correlation Coefficient (MCC) is a measure of the quality of binary classifications. Its range is from -1, where the prediction is perfectly negative, to 1, where the prediction is perfectly positive; 0 indicates the absence of correlation. It is a metric that considers all four confusion matrix categories, as depicted in Equation 2.5.

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{2.5}$$

F1-score is the harmonic mean of precision and recall, providing a balance between the metrics. It is useful when the class distribution is imbalanced (Equation 2.6).

$$F1_{score} = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{2.6}$$

2.5.3.2 Regression Models

In a regression model, commonly used metrics include: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R^2.[21][22]

Mean Absolute Error (MAE) measures the average absolute difference between predicted and actual values, without considering any outliers (Equation 2.7).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2.7}$$

Where:

- n is the number of instances
- y_i is the actual value
- $\hat{y}_i$is the predicted value.

Mean Squared Error (MSE) measures the average of the squares of the errors, giving higher weight to larger errors. Indeed, it gives an increasing exponential penalty to increasing prediction errors, meaning that the bigger the prediction error, the bigger the penalty (Equation 2.8).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2.8}$$

Root Mean Squared Error (RMSE) is the square root of the MSE and provides a measure in the same units as the predicted values (Equation 2.9). RMSE is useful for understanding how much error to expect in predictions. As for MSE, lower values indicate better performance.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2.9}$$

R^2 is also called the coefficient of determination of predicted values versus experimental values; it measures the proportion of variance in the dependent variable that can be explained by the independent variables in the model. It indicates a

goodness of fit (Equation 2.10).

$$R^2 = 1 - \frac{RSS}{TSS} \tag{2.10}$$

Where:

- RSS is the sum of squared residual
- TSS is the total sum of squares.

R^2 values range from 0 to 1: higher values indicate a better fit, indeed more variability is explained by the model.

Finally, Q^2 is a metric that is used in Cross-Validated predictions, indicating how well the model can predict new data and identifying overfitting phenomena, contrary to R^2. It is calculated in a similar way to R^2, but the focus is on predictions made on the validation set, instead of the training set (Equation 2.11).

$$Q^2 = 1 - \frac{PRESS}{TSS} \tag{2.11}$$

Where:

- PRESS is the sum of squares of predictive errors
- TSS (total sum of squares): variance in the predicted values.

It ranges from $-\infty$ to 1. A Q^2 value close to 1 means that the model has good predictive power and explains a high percentage of the variance in the validation data, without overfitting. A value of 0 suggests that the model does not perform better than a simple mean prediction, while a negative Q^2 value indicates that the model is likely overfitting the data.

2.6 Chemical Descriptors

Chemical descriptors are numerical values that represent molecule chemical properties and have a crucial role in predictive modelling.

They can be divided into different categories based on the molecular information

they encode. The first main two groups of descriptors refer to their dimensionality of the molecular representation, indeed they could be 2D or 3D type.

2D descriptors include features that are not dependent on the tridimensionality of the molecule. Indeed, they include:

- Constitutional descriptors: they derive from the molecular formula and 2D structure of the compound, *e.g.* number of atoms, count of various kinds of bonds
- Topological descriptors: they are derived from the molecular connectivity and are calculated based on the molecular graph; they include bond and atomic properties
- Electronic descriptors: they describe the distribution of electrons in the molecules, *e.g.* atomic charges, dipole moment, HOMO and LUMO
- Physicochemical descriptors, such as molecular weight or logP, which is a measure of lipophilicity.

3D descriptors derive from the three-dimensional representation of the molecule; they could be divided into:

- Geometric descriptors: they describe the spatial arrangement of atoms within a molecule, *e.g.* volume, surface area (solvent accessible surface area – SASA – and solvent accessible volume – SAV)
- Electronic descriptors: they derive from the electron density distribution in 3D space, such as electrostatic potential and molecular orbitals. They are usually calculated by semi-empirical methods.

2.6.1 Fingerprints

Fingerprints are numerical representations encoding for a molecule, where its structure is translated into a binary format (a series of "0" and "1"). Each bit corresponds to the presence (expressed with "1") or the absence (expressed with "0") of a substructure, a feature or a fragment.

Fingerprints are compact but still efficient representations of molecular structures and allow easy and fast comparison between two or more molecules, since similar structures will have similar fingerprints.

There are several kinds of fingerprints, each encoding for a different aspect of molecular structure or its properties. In substructure-based fingerprints, such as Daylight [23] or MACCS [24], each bit corresponds to a specific substructure (if it is present, the bit is "1", otherwise "0"). Circular fingerprints (*e.g.* ECFPs - Extended-connectivity fingerprints [25]) are able to generate atom-centred features based on neighbouring atoms and bonds. A last category worth mentioning is protein-ligand interaction fingerprints (IFP) which describe the interactions between a small molecule and its protein target (see Section 5.1).

2.6.2 Protein Descriptors

Protein descriptors are features that characterise the general properties of a protein or part of it. Due to the considerable dimension of this macromolecule, a selection of its structure (binding pocket, or an important site) is usually examined.

Protein descriptors are commonly used in PCM modelling, as explained in next paragraph.

Protein descriptors can be expressed as binary descriptors, where each feature refers to a residue, helix or chain; the label "0" or "1" indicates the absence or presence of that element. 3D protein descriptors require an adequate dataset of available protein 3D structures. Finally, sequential protein descriptors are based on amino acid sequence and encode for properties of each residue.

In the work described in Chapter 4, Z-scales descriptors were chosen. They are physicochemical descriptors derived by PCA, that has compressed down to three (Z-scales (3)) or five (Z-scales (5)) principal components. In detail, in Z-scales (5) set, Z1 is the lipophilicity, Z2 represents steric properties (such as steric bulk and polarisability) and Z3 describes electronic properties (polarity/charge); Z4 and Z5 relate with electronegativity, heat of formation, electrophilicity and hardness.[26][27]

2.6.3 Scoring Functions as SB Descriptors

Molecular docking studies are structure-based methods that aim to predict the binding mode and the site of interaction of a ligand within the considered target,

or to calculate the target-ligand absolute binding affinity. It is designed to be a simple and then fast approach, so that it can be easily applied to virtual screening campaigns.

Molecular docking assigns a score that indicates how good (or bad) is the predicted pose through specific descriptors, called scoring functions, that could be considered as surrogates of ligand bioactivity. For this reason, the accuracy and the predictive power of a scoring function represent one of the main critical aspects in determining the robustness of the docking simulation.

The most common scoring functions can be divided into four main groups: force field-based, empirical, knowledge-based and artificial intelligence-based.

Force field (ff)-based scoring functions (*e.g.* ChemPLP [28], DOCK [29]) rely on physical principles to calculate the interaction energy between the ligand and the target. They consider physical atomic interactions, such as van der Waals (VDW) and electrostatic interactions, bond stretching, bending and torsional forces.[30][31]

Empirical scoring functions estimate the binding affinity of a complex through the sum of several empirical energy terms, weighted by experimental coefficients. Some examples are X-Score [32], MLP_{InS} [33] or ChemScore [34]. These are based on regression analysis from experimental binding affinity data of resolved protein-ligand complexes.[30][35][31]

Knowledge-based scoring functions, such as PMF [36], rely on rules and principles statistically derived from protein-ligand complexes experimentally resolved. They use potentials derived from the more frequent atomic interactions-pairs.[30][31]

Finally, the newest AI-based (previously known as ML-based) scoring functions learn directly from experimental data, enabling them to capture non-linear and complex interactions that are difficult to represent explicitly. Some well-known examples include RF-Score [37] or NN-score [38].[39][40]

2.7 Proteochemometric Modelling (PCM)

Proteochemometric modelling (PCM) is a computational method that models the bioactivity of multiple ligands (usually small molecules) against multiple targets (usually proteins). It aims to provide a modelling of the molecular-protein interaction space; indeed, this model is trained on a dataset composed of many compounds tested on many targets, *e.g.* a protein family.

PCM models rely on various descriptors that represent both ligand and targets. Ligand descriptors can be binary flags indicating the presence of some functional groups, as well as more complex properties like PSA (Polar Surface Area) or polarisability. Target descriptors describe the protein structure or its sequence, often focusing on the binding site. The innovative PCM descriptors are cross-term ones, which can extrapolate non-linear interactions between the ligand and the target.[41][26][27]

Usually, PCM employs non-linear machine learning algorithms, such as RF, SVM or gradient boosting, but also linear methods, *e.g.* Partial Least Squares (PLS) or PCA analysis are applied.

Contrary to QSAR, which identifies and analyses interactions between a ligand and a single target, PCM applies multi-target modelling. This is very useful when exploring promiscuous drugs that interact with multiple proteins, or when studying target selectivity. PCM allows for more robust predictions by incorporating protein variations, enabling the model to generalise across related targets. While QSAR and PCM principles seem to be very similar, the incorporation of additional information in model training introduces the concept of congenericity: compounds with similar features share similar targets, and targets with similar compounds show shared features.

Chapter 3

Xenobiotic Metabolism Prediction

3.1 Xenobiotic Metabolism

Xenobiotics are defined as exogenous substances found within an organism that are not naturally produced or expected to be present within the organism. They are defined as "chemicals found but not produced in organisms or the environment. Some naturally occurring chemicals (endobiotics) become xenobiotics when present in the environment at excessive concentrations".[42]

The term xenobiotic derives from the Greek word "*xenos*" meaning foreigner and "*bios*" meaning life. They include pesticides, environmental pollutants, cosmetics, food additives, pharmaceutical compounds and drugs too.

Human metabolism has an inner defence strategy to face these foreign compounds; more detailed, xenobiotics are transformed into more hydrophilic metabolites in order to be less active and more easily excreted.

For what concerns drug metabolism, as it is the main topic of this Chapter, this step belongs to a wider biotransformation process which a drug undergoes since the administration step. In fact, the disposition of a drug in the body involves the ADMET process, providing absorption, distribution, metabolism, excretion and toxicity steps. As highlighted in a review of 2014 [43], metabolism represents the major clearance route of 75% of drugs.

Metabolism is a very complex process where drugs are structurally, enzymatically or chemically modified into different metabolites by various apposite enzymes. They can mediate the activation – as it happens with prodrugs –, the deactivation and degradation of a drug and, in some cases, they are also responsible for a conversion in toxic derivatives.[44]

3.1.1 Metabolism Prediction: State of the Art

As the last years have seen a significant change in the entire drug discovery and development pathway, also metabolism studies have benefited from *in silico* calculations and predictions.

Despite recent developments in medium- and high-throughput screenings for pharmacokinetic and metabolic studies, a complete experimental profile is available for a small subset of compounds only. This is the reason why the selection of these tested substances needs to be carefully carried out, often based on *in silico* methods.

Therefore, in this modern approach, the attractive strategy "fail early, fail cheap" is depictied as the most fruitful one since the ADMET profile should be determined as early as possible in the test cascade, in order to allow a timely assessment of proper profiles.[45]

Based on these grounds, the interest in computational approaches is dramatically increasing, since these tools consent to carry out metabolic studies in a reduced time and with a lower cost. Furthermore, the increasing number of commercial failures, drug withdrawals or other kinds of safety controversies lead to the urgent need to increase the accuracy of metabolic, toxicity and safety drug studies.

Searching by literature, metabolism prediction methods can be subdivided into two groups: local and global methods.

Local methods include three-dimensional quantitative structure-activity relationships (3D-QSAR), quantitative structure-metabolism relationships (QSMR), quantum-mechanical calculations (QM), molecular modelling and docking, and combinations of them. They provide structure- (SB) and ligand-based (LB) approaches suitable for single systems (single enzyme or single reaction). However, their precise requirements reduce their applicability, since it is essential to know the metabolic

reaction a certain substrate can undergo and the precise enzyme isoform, besides the enzyme 3D-structure in SB approaches.
Instead, global methods can be applied to various biological systems, and they are able to extrapolate general rules or metabolic networks. In addition, they can be applied to a large variety of compounds organised in metabolic databases, but their accuracy highly influences the reliability of the resulted predictions.[46]
Among the recently published global approaches, it is worth mentioning: Biotransformer [47], which integrates a rule-based system with ML techniques to predict small-molecule metabolism in humans; FAME [48], a tool designed to predict sites of metabolism (SoM) for phase I and II reactions; and XenoSite [49], which employs neural networks trained on quantum-mechanical descriptors. In detail, the FAME method was developed using datasets sourced from MetaQSAR.

3.1.1.1 MetaQSAR

As previously explained, the study of metabolism benefits from bioinformatic and chemoinformatic approaches, and there is the consequent need to work with accurate metabolic data to obtain accurate predictions as well. On these grounds and with the aim to overcome the inaccuracy drawback affecting other global methods, in 2018 Pedretti *et al.* [50] proposed a novel application, called MetaQSAR, specifically tailored to generate and manage metabolic databases. This is a relational database where all the metabolic reactions are manually collected by critically reviewing specialised literature from 2004 to 2016. The chosen classification for the metabolic reactions subdivides them into 3 main classes (*i.e.* redox reactions, hydrolyses and conjugations), 21 classes and 101 subclasses.

In Appendix I the full metabolic classification is listed.

All collected data from MetaQSAR demonstrated success in developing predictive models, in detail concerning the occurrence of conjugation reactions with glutathione and glucuronic acid. Furthermore, with the aim to enhance the MetaQSAR metabolic data accuracy, in 2021 MetaTREE was presented.[51] It is a new database built on the data extracted from publications reporting exhaustive metabolic trees only, so managing the problem of false negative instances that often influences the reliability of obtained models. Commonly, for a long time, the selection of negative samples has been affected by the absence of effective inactivity knowledge and data, going ahead with a huge approximation. Indeed, for the compilation of MetaTREE, in order to

reduce the number of false negative data, the authors focused on the papers, which report exhaustive metabolic trees, so determining true positive and true negative samples.
The so-developed database demonstrated enhanced predictive performances compared with the general tool MetaQSAR, due to a reliable and accurate selection of learning sets.

For all above stated reasons, in this research work, all metabolic data were extracted from MetaQSAR and MetaTREE, providing the primary source for predictive models generation as described in Sections 3.2 and 3.3.

3.1.2 Sections Overview

As mentioned, this Section is focused on the prediction of metabolism reactions, from different points of view. Firstly, in Section 3.2, predictive studies on glutathione conjugation reaction are proposed and discussed. Then, in Section 3.3, a new tool able to predict the site of metabolism is presented.

3.2 Conjugation with Glutathione

The glutathione (GSH) conjugation reaction is a metabolic reaction of phase II that results in the detoxification of some species, often potent electrophiles. It is usually catalysed by the enzyme glutathione-S-transferase (GST), but in the presence of high GSH intracellular concentrations – as observed in the liver or often in preneoplastic or neoplastic cells – the conjugation may occur spontaneously too.[52]

GSH plays many important roles in metabolic pathways.
In detail, it can lead to the formation of less toxic compounds than the parent xenobiotic, and more polar and hydrophilic GSH-S-conjugate.
Another role is in body clearance, since the conjugates are substrates for phase III transporters that guide the biliary and renal excretion.
In other cases, GSH conjugates can be more active than the parent compound, being involved in biotransformation reactions as in pro-drugs.[53][54][55]

Based on these grounds, knowledge of this reaction may be useful in designing

new therapeutic opportunities, both in the toxicity and safety fields, as well as in the discovery and development of new drugs.

3.2.1 Glutathione Transferase (GST)

GST (EC 2.5.1.18) isoenzymes have been found ubiquitously in nature, in different organisms and microorganisms, plants, fish, birds and mammalians. Among the detoxification enzymes, they play a crucial role in providing protection against electrophiles and products of oxidative stress.[56]

In mammalian, the first deepest studies of GST isoenzymes were carried out in rat liver, where many isoforms were initially characterised, and then studies were expanded to mouse and human tissues too.[57] Two supergene families encode for proteins with GST activity: the larger superfamily comprises cytosolic or soluble enzymes, while the other one is composed of microsomal proteins.
The human cytosolic/soluble superfamily comprehends at least 16 genes subdivided into eight classes: Alpha, Kappa, Mu, Pi, Sigma, Theta, Zeta and Omega, where only Kappa GST is mitochondrial and probably not located in the cytoplasm. The microsomal superfamily contains at least 6 genes expressed in membranes, called MAPEG (Membrane Associated Proteins in Eicosanoid and Glutathione metabolism).[58][59]
Through the millennials, the evolution has allowed a variety of functions, localisations and mechanisms. Indeed, the soluble GSTs seem to be involved mainly in the metabolism of xenobiotics, such as carcinogens, environmental pollutants and drugs, while MAPEG proteins are primarily involved in the biosynthesis of leukotrienes and prostanoids. Therefore, globally, all the GST isoenzymes contribute to cellular detoxification, autocrine and paracrine regulatory mechanisms.[56][59]

3.2.1.1 GST Structure

Among all isoenzymes mentioned above, the first characterised GST structure was the porcine pi-class enzyme (pGSTP1-1) and, in general, mammalian cytosolic GSTs were among the first to be structurally determined.[60]
GSTP isoenzyme presents the main features of cytosolic GST superfamily members: an N-terminal thioredoxin-like domain ($\beta\alpha\beta\alpha\beta\beta\alpha$) and a C-terminal domain (α-

helices), as illustrated in Figure 3.1.

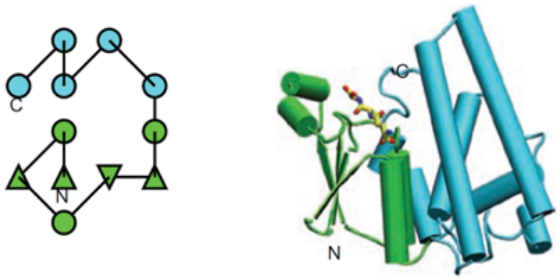

Figure 3.1: Topology and structural representations of cytosolic GST. The thioredoxin-like domain is represented in green, while the C-terminal domain in cyan. Reproduced from [61]

In the early 2000s, the characterisation of the mitochondrial GSTs (rGSTK1-1 and hGSTK1-1) highlighted some differences as well as some similarities as compared to the cytosolic GSTs.[62][63]
Also GSTK isoenzymes have a thioredoxin-like domain too, but with a different folding, indeed they have a DsbA-like α-helical domain located between helix-α2 and strand-β3, as Figure 3.2 depicts.

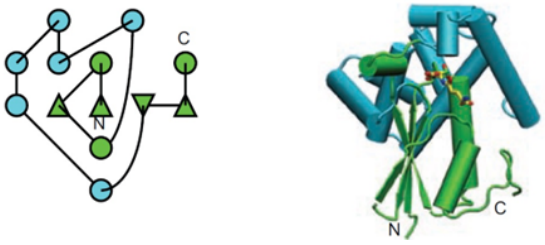

Figure 3.2: Topology and structural representations of mitochondrial GST. The thioredoxin-like domains are represented in green, while the DsbA-like insertion in cyan. Reproduced from [61]

In both cytosolic and mitochondrial GSTs, the binding site for glutathione (G-site) is positioned in the thioredoxin-like domain and a key conserved interaction is a cis-proline residue that establishes a hydrogen-bond with the backbone amine group of cysteinyl moiety of GSH. While GSH-binding mode results mostly conserved, xenobiotics one (H-site) seems to be more complex and class-specific. Indeed, as Ji *et al.* [64] highlighted in their studies, in the class μ GST, the xenobiotic substrate-binding site is a hydrophobic cavity, defined in rGSTM1-1 by the side chains of Tyr6, Trp7, Val9, Leu12, Ile111, Tyr115, Phe208 and Ser209. Differently, in the pi-class the H-site is half hydrophobic and half hydrophilic, as it happens in hGSTP1-1 where the cavity is defined by the side chains of Tyr7, Phe8, Val10, Arg13, Val104, Tyr108, Asn204 and Gly205 and five water molecules.

Mammalian cytosolic GSTs are all dimers. Throughout the years, many studies have been carried out to identify experimental conditions that promote the dissociation into monomers and to well characterise the isolated subunits. Lots of evidence supports the conclusion that the single monomers are inactive, as Abdalla *et al.* found [65], studying the dissociation of hGSTP1-1 after key residue mutations.[66]

Therefore, as previously described, although this enzyme is secreted as a monomer into plasma by platelets, the single subunit form is fully inactive.[67]

3.2.1.2 GSH Conjugation Mechanism and GST Role

GSH conjugation is catalysed by GST, but it can potentially occur in a non-enzymatic way too.

GSH is a tripeptide composed of L-γ-Glu, L-Cys, L-Gly and it is characterised by the acidity of the SH group; it has a pK_a of 9.0 in aqueous solution and it is poorly ionised at physiological pH (Figure 3.3).

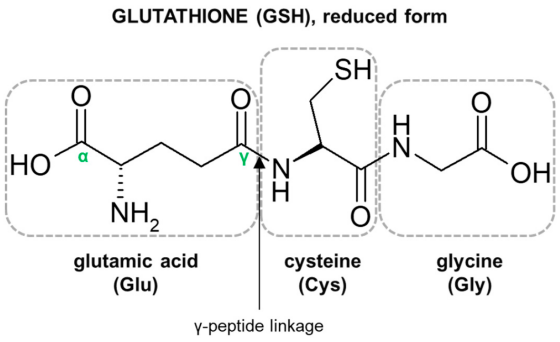

Figure 3.3: Structure of the glutathione (GSH). Glutamic acid is linked in a γ-peptide linkage to cysteine, which in turns forms an α-peptide linkage with glycine. Reproduced from [68]

GSH is a soft nucleophile in protonated and ionised form, which is translated into a certain reactivity towards soft electrophiles. When the xenobiotic is a hard electrophile, the spontaneous conjugation to GSH is not permitted, since it has a highly localised positive charge which is poorly polarisable, so it remains strongly localised in the presence of a soft nucleophile (Figure 3.4).[69][70]

Notably, SH acidity significantly increases when GSH is bound to GST ($pK_a \simeq$ 6-7), due to a stabilisation of the ionised thiol group.[63] Since the S^- anion represents the catalytically active group, its increased acidity determines an increase

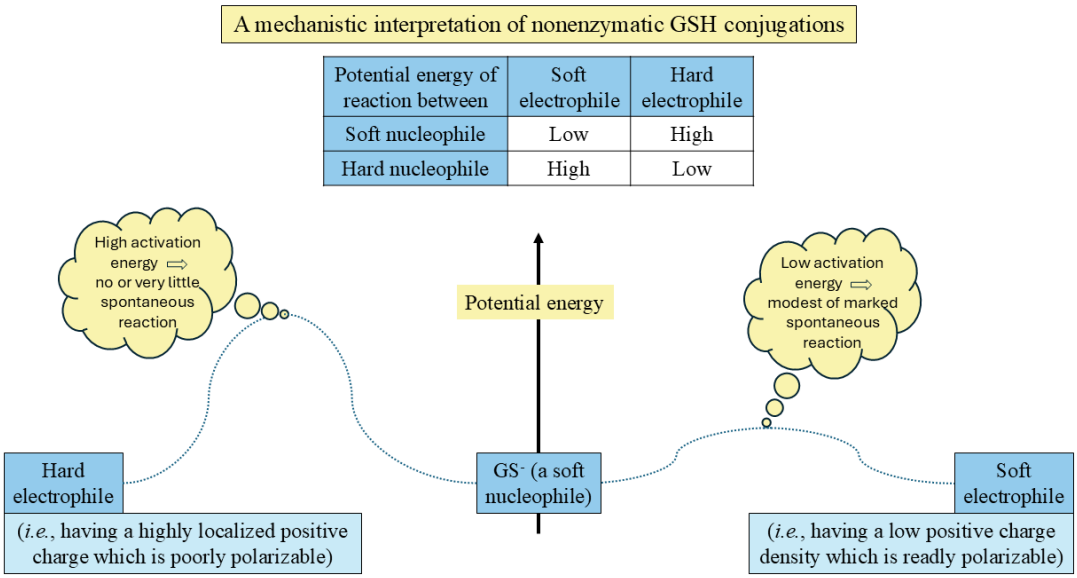

Figure 3.4: A proposed mechanism elucidating the non-enzymatic occurring of GSH conjugation. It is based on electrophilic and nucleophilic classification. [69][70] Reproduced from [71]

of nucleophilicity toward reactive electrophiles.

The GST enzymes catalyse the nucleophilic attack of the GSH on electrophilic substrates. In Figure 3.5 a general scheme of GSH conjugation with a generic electrophile R-X is reported.

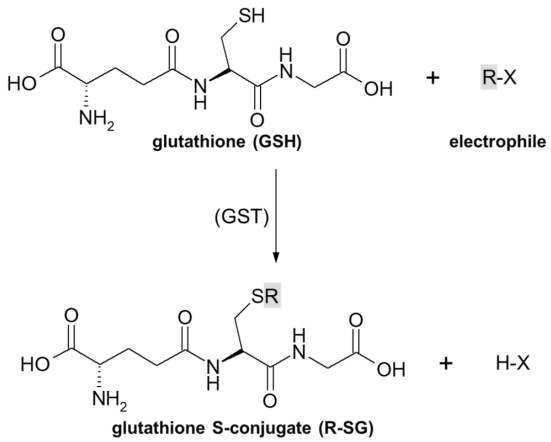

Figure 3.5: Glutathione conjugation to a generic electrophile R-X, resulting in the glutathione S-conjugate (R-SG). Reproduced from [68]

GST can catalyse many mechanisms of GSH conjugation reaction, such as nucleophilic aromatic substitutions, epoxide ring-opening reactions, or Michael additions to α,β-unsaturated ketones.[72]

About the catalytic mechanism, it seems that the GST first activates the GSH thiol group, generating the thiolate anion that increases the nucleophilicity, and obtaining the transition state compound. The reaction scheme is reported in Figure 3.6 and shows how Tyr6, Ser209 and Tyr115 have a key role.

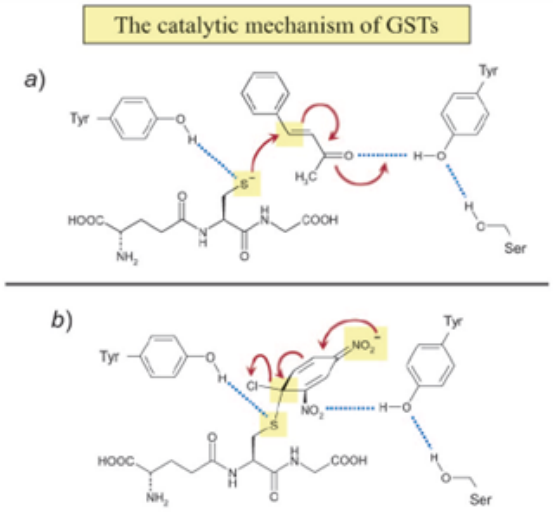

Figure 3.6: GST catalytic mechanism. (a) The substrate 4-phenylbut-3-en-2-one goes under a glutathione addition. (b) The represented transition state of 1-chloro-2,4-dinitrobenzene in a glutathione addition-elimination.[71]

3.2.2 Aim

Concerning metabolism and drug-induced toxicity prediction, over the years, the most used strategy has involved the presence of structural alerts, called toxicophores, such as aromatic nitro, aromatic amine, nitrosoamine, epoxide, α,β-unsaturated compounds, that can easily generate reactive metabolites.[73] Regarding GSH conjugation, the paracetamol case study is the most famous.[74] Indeed, acetaminophen is a huge cause of drug-related mortality in humans, inducing extensive hepatic necrosis after a single toxic dose, due to the generation of the N-acetyl-p-benzoquinone-imine (NAPQI) that induces GSH-depletion, leading to oxidative stress and, ultimately, liver damage (Figure 3.7).[75]

This work was born from the need to better investigate the GSH conjugation, in order to address drug safety and toxicity and other peculiar features of the reaction, such as GSH polarisability and ability to react spontaneously, as well as in a catalytic way. So far, the main approaches to predict reactivity with GSH have provided the exploitation of ligand-based descriptors, such as stereo-electronic ones that can

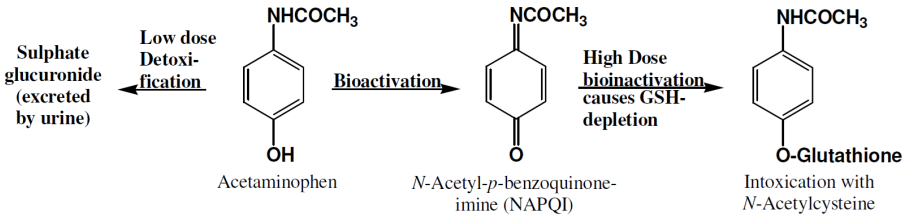

Figure 3.7: Scheme of the metabolic pathway of acetaminophen. Reproduced from [76]

parametrise both softness and electrophilicity, while structure-based approaches have often been left apart, maybe for the past deficiency of experimentally resolved GST structures. Luckily, the number of GST experimentally resolved is increasing and now a great number of them are available for structure-based studies.

On these grounds, the present study aims at investigating the enhancing role of docking simulations in predicting the reactivity with GSH. The calculations comprise three GST isozymes (*i.e.* hGSTA1, hGSTK1 and hGSTP1) chosen for their structural diversity and for their major role in xenobiotic metabolism.

3.2.3 Methods

3.2.3.1 Dataset Preparation

This work started from the creation of a proper dataset able to be applied to the glutathione reaction; in detail, a binary dataset was extracted. It was collected by searching in MetaTREE according to MetaQSAR metabolic reaction classification. Concerning the substrates, molecules with code 24.01 (nucleophilic additions of GSH) and 24.02 (addition-elimination reactions with GSH) were selected; the non-substrates were molecules giving metabolic reactions other than coded as 24.01/.02/.03/.04 (where 24.03 is the metabolic processing of GSH conjugates up to thiols, while 24.04 is the conjugation of GSH with Hg, As, Pt etc. compounds). A dataset of 981 molecules was obtained; then it was cleaned by excluding molecules with less than 13 atoms, with a molecular weight larger than 1000Da and molecules with counter-ions or rare elements. At the end, the resulting dataset was composed of 353 substrates (binary label: 1) and 523 non-substrates (binary label: 0).

Before further studies, the 3D structure of the collected molecules was optimised

by applying a standardisation protocol, including pH setting (physiological value: 7.4) plus a semi-empirical PM7- based structural refinement (as implemented in the MOPAC 2016 software [77]), with the keywords: CHARGES, PM7, EPS=80, PRECISE, GEO-OK, SUPER (reported in Appendix II with a brief explanation).

3.2.3.2 Ligand-Based (LB) Feature Calculation

Three different sets of molecular descriptors were calculated, as shown in Table 3.1.

	Descriptor	**Dimensionality level**	**N. of descriptors**	**Software**
A	Physicochemical	1D/2D/3D	27	VEGA ZZ
B	Semi-empirical	3D	20	MOPAC
C	Geometrical, topological, constitutional	1D/2D/3D	171	BlueDesc

Table 3.1: Main features of each set of descriptors: name, type, dimensionality level, size and generating software.

The first set of descriptors (Set A) is composed of molecular physicochemical properties calculated by VEGA ZZ [78], while Set B includes stereo-electronic molecular properties computed by semi-empirical simulations. In addition, a set of 171 more detailed 1D-, 2D-, 3D-descriptors (Set C) was investigated, including topological, constitutional and geometrical descriptors, which were calculated by the software BlueDesc [79]. The full list of these descriptors is provided in Appendix III.

Since the input data of these binary classification models include a large number of molecular descriptors, it was a fruitful strategy to filter such a wide variety of features to focus on the most informative ones. Two kinds of feature selection (FS) were carried out through KNIME [80] workflows: Forward Feature Selection (FFS) and Backward Feature Elimination (BFE) strategies. Both start with a Correlation Filter to remove redundant attributes, with a correlation coefficient threshold of 0.97 and, then, with a specific Feature Selection Loop the features that gave the best performant model (according to Matthew Correlation Coefficient – MCC) were selected.

3.2.3.3 Molecular Docking Studies

All docking studies performed in this work were carried out using crystal structures of human glutathione S-transferase (hGST). Three different isoforms of the enzyme were considered (GSTA1-1, GSTK1-1 and GSTP1-1), chosen as the most representative ones. GST-A is mostly found in the cytosol and it is related to oxidative stress mechanisms, GST-P is a cytosolic enzyme expressed in many tumoral tissues and responsible for neoplastic drug resistance phenomena [81], finally GST-K is a newfound mitochondrial GST, not present in the cytoplasm [82].

For each isoenzyme, the substrate binding site has been investigated, in the presence of the cofactor GSH. The selection of the PDB structures was done based on the experimental method (X-ray preferred), the resolution, the R-value free and the presence of the cofactor glutathione (GSH) and a substrate (or an adduct with GSH). For GST-A, the PDB ID 6YAW [83] was selected, for GST-K, PDB ID 1YZX [63] and for GST-Pi the PDB ID 3PGT.

These characteristics are reported in Table 3.2.

	GSTA1-1	**GSTK1-1**	**GSTP1-1**
PDB entry	6YAW	1YZX	3PGT
Method	X-ray diffraction	X-ray diffraction	X-ray diffraction
Resolution	2.19Å	1.93Å	2.14Å
R-value free	0.247	0.204	0.159
Presence of a co-crystallised ligand	Adduct of cinnamaldehyde	GSH sulfinate	Adduct of BDPE

Table 3.2: Characterisation of the selected crystal structures, from PDB.

All structures were prepared in the same workflow: after removing crystallographic water molecules, ions and other crystallographic reagents, the selected structures were completed by adding the hydrogen atoms to amino acid residues and ionising the protein considering physiologic pH at 7.4. H^{++} server was chosen to predict the protonation state of ionisable protein groups.[84] A general check of the protein was

performed, including chiral atoms, cis bonds, ring intersections and bond lengths. When an adduct was present, it was divided into GSH and substrate, and the hydrogen atoms were fixed. The protein was refined with two minimisations of 10000 steps each, using NAMD software [85], implemented in VEGA ZZ suite. The first one was carried out with all the protein atoms fixed except for hydrogens; the second one was performed with the backbone atoms under constraints, to preserve the resolved folding. In both of them, the GSH and the substrate were unrestrained. Before the following docking studies, all substrates were removed.

Few settings of docking studies were considered: with PLANTS v.1.2 software [86] and GOLD v.5.8.1 software [87], for each of the three isoenzymes. Docking studies were carried out, selecting as binding pocket a 7Å radius sphere around the centre of mass of co-crystallised GSH, generating 10 docked poses for each ligand.

For the PLANTS docking, the calculation was set with the ChemPLP scoring function and the search exhaustiveness speed1. For the GOLD docking, "very flexible" docking was set as efficiency and ChemPLP as scoring function.

3.2.3.4 Rescoring Procedure and Structure-Based (SB) Approach

All the generated docking poses were rescored by using Rescore+ tool [88] implemented in Vega ZZ. From the resulting poses, 24 additional scoring functions have been calculated and used as descriptors (for the full list, see Appendix IV rescore descriptors):

- The different components of PLANTS and X-Score scoring functions
- The MLP scores describing the hydrophilic/hydrophobic complementarity [33]
- The Elect and ElectDD scores representing the electrostatic interactions
- CHARMM Lennard-Jones components for the evaluation of the van der Waals interactions
- The Contacts scores which account for the number of residues surrounding the bound ligand [89].

The resulting scores were analysed in two different ways to train ML models. In the first case, for each score, the value of the top-ranked poses was considered for each ligand (BP approach). In the second case, all the docked poses were considered by

computing the mean values of each scoring function for each ligand (AV approach).

3.2.3.5 Model Building and Validation

All the models described in this work were built by means of tailored KNIME workflows by using Random Forest algorithm with these settings: 10-fold cross-validation with stratified sampling, information gain ratio as quality measure, 100 decision trees and maximum depth of trees as 10.

Few scenarios were considered:

- Ligand-based models:
  - Performance of Set A (with and without FS)
  - Performance of Set B (with and without FS)
  - Performance of Set C (with and without FS)
  - Global performance of LB descriptors (with and without FS)
- Structure-based models:
  - Performance of each isoenzyme with PLANTS algorithm (BP and AV approaches, with and without FS)
  - Performance of each isoenzyme with GOLD algorithm (BP and AV approaches, with and without FS)

Regarding SB studies, a supplementary analysis was carried out, called consensus voting.[90] It allows the aggregation of predictions from multiple decision trees to form the final output of the model. The class that gets the majority of votes becomes the final prediction. By averaging or taking the majority vote across multiple trees, the variance of the model is reduced, increasing the robustness and leading to more accurate and reliable predictions.

3.2.3.6 Hardness and Predictive Power

To assess the impact of the electrophilicity on the GSH conjugation occurrence, hence on its spontaneous or catalytic manner, an overall evaluation and correlation between stereo-electronic descriptors and model predictive power were considered. For each isoenzyme, considering PLANTS docking results analysed by BP and AV approaches, the predictive power was expressed by binary labels 0 and 1. Then, after

choosing a proper descriptor on which all data was ranked, for each of these six cases, the cumulative average was calculated with Equation 3.1:

$$C_n = \frac{1}{n} \sum_{i=1}^{n} B_i \tag{3.1}$$

where:

- C_n is the cumulative average at row n
- B_i is the value of column B at row i
- n is the number of rows up to the current row.

This cumulative average expresses the average predictive power of the model up to that point, providing insight into how the overall performance of the model evolves as more instances are considered, so along the specific considered property.

3.2.4 Results and Discussion

3.2.4.1 LB Predictive Models

In order to evaluate the performance of the so generated predictive models, few parameters were taken into account, as reported in Table 3.3. The preliminary LB models demonstrated good ability in discriminating between substrates and non-substrates, with an MCC value of 0.648 for the physicochemical descriptors set, 0.530 for the semi-empirical ones and 0.607 for the geometrical, topological and constitutional descriptors set.

Then, a Feature Selection step was carried out and the performances of resulting models are shown in Table 3.4.

It is worth noticing that the evaluation metrics are enhanced, and the MCC value for each set is improved too, as well as for all parameters, compared with the previous study performed without FS. The Backward Feature Elimination strategy approach showed a greater improvement of metrics such as Accuracy or MCC concerning Set A and Set B, while the Set C model demonstrated a better improvement with Forward Feature Selection (Table 3.5).

Evaluation matrix	Set A VEGA ZZ	Set B MOPAC	Set C BlueDesc
Accuracy	0.832	0.777	0.813
Sensitivity	0.745	0.649	0.717
Specificity	0.891	0.864	0.878
F-Measure	0.782	0.701	0.755
MCC	**0.648**	**0.530**	**0.607**

Table 3.3: Model performance on single sets of descriptors, without FS step.

Evaluation metrics	Set A		Set B		Set C	
	FFS	**BFE**	**FFS**	**BFE**	**FFS**	**BFE**
Accuracy	0.854	0.856	0.784	0.787	0.866	0.852
Sensitivity	0.748	0.773	0.657	0.646	0.776	0.771
Specificity	0.925	0.912	0.870	0.881	0.927	0.906
F-Measure	0.805	0.813	0.711	0.709	0.824	0.807
MCC	**0.694**	**0.699**	**0.545**	**0.550**	**0.721**	**0.689**

Table 3.4: Model performance on single sets of descriptors, with FS step, with the two described methods in 3.2.3.

Even if in Set A and Set B the difference between the two approaches is not so relevant (less than 1%), in Set C FFS proved to be significantly more effective than BFE, with a clear improvement of 7%. Globally, Set C benefited much more from FS than the other two sets did. This different behaviour and FS influence may be dependent on the composition and, in detail, on the size of Set C. BlueDesc set is composed of 171 descriptors, compared to 27 and 20 attributes of Set A and B, respectively. A selection step surely impacts much more on a great quantity and variety of features, rather than on a small dataset. Indeed, for each Set, the FS reduced the number of descriptors, as depicted in Table 3.6.

Evaluation metrics	Set A		Set B		Set C	
	FFS	**BFE**	**FFS**	**BFE**	**FFS**	**BFE**
MCC w/o FS	0.648		0.530		0.607	
MCC w FS	0.694	0.699	0.545	0.550	0.721	0.689
Improvement %	7.16%	7.83%	2.77%	3.66%	18.79%	13.58%

Table 3.5: MCC value comparison, without and FS strategies, and improvement expressed as %.

Evaluation metrics	Set A		Set B		Set C	
	FFS	**BFE**	**FFS**	**BFE**	**FFS**	**BFE**
Entire dataset w/o FS	27		20		171	
Entire dataset w FS	7	7	8	10	25	10
Decrement %	-74%	-74%	-60%	-50%	-85%	-94%

Table 3.6: Number of features w/o and with FS step, and percentage of decrement.

Concerning Set C, the decrement is about 85% and 94% for FFS and BFE, respectively, showing a broader impact than happened in Set A and B (74% and 50-60%).

In order to assess the global descriptive power of the considered set of features, an overall model was considered. As first step, all descriptors were taken into account, resulting in 218 attributes; then, as similarly done with single sets, the FS step was carried out. The performances are reported in Table 3.7.

Firstly, it is important to note that the global model gave an improved performance, even without FS step. This behaviour could be explained since this kind of model can describe the chemical space of the considered dataset more meticulously, considering at the same time physicochemical descriptors, semi-empirical quantum-mechanical and topological ones. FS step improves the performance, and the more effective strategy resulted FFS in this case too.

Evaluation metrics	Global LB model		
	w/o FFS	**FFS**	**BFE**
Accuracy	0.839	0.889	0.881
Sensitivity	0.751	0.824	0.807
Specificity	0.899	0.933	0.931
F-Measure	0.790	0.857	0.846
MCC	**0.662**	**0.769**	**0.752**

Table 3.7: Performance of the global model, considering the entire set of descriptors (Set A, Set B and Set C), w/o and with FS step.

3.2.4.2 Molecular Docking Studies

Regarding molecular docking studies, GOLD program was not able to dock the entire datasets within the considered binding pocket, resulting in incomplete docked poses. In detail, 13.6% of the molecules were missing in GST-A docking, while 32.6% in GST-K docking. Notwithstanding this, the balancing between two binary classes remains almost unaltered and so the results were equally analysed.

3.2.4.3 SB Predictive Models

Thanks to the availability of different GST experimental structures, the following predictive models were generated, enabling the exploration of protein-ligand interactions.

Considering all scenarios – three isoenzymes, PLANTS and GOLD docking programs, BP and AV approaches – globally, PLANTS software performed better than GOLD one, with a mean MCC value of 0.440 and 0.386, respectively. This result may be ascribed to an evident GOLD difficulty in identifying the requested stable protein-ligand complexes – reason why many ligands were not docked within the binding pocket. GST-K demonstrated a worse performance than GST-A and GST-P, maybe due to its scarce structural knowledge and its different cellular location.

Finally, comparing BP and AV approaches, the results highlighted that AV one better described the binding mode, probably due to a wider consideration of the site.

As previously done in LB studies, FS pre-processing was carried out, applying both already mentioned strategies, FFS and BFE. In Figure 3.8, MCC values of all scenarios are depicted in histograms, while the broken line refers to its improvement expressed as a percentual difference between the considered FS approach and the model built on the entire set of descriptors.

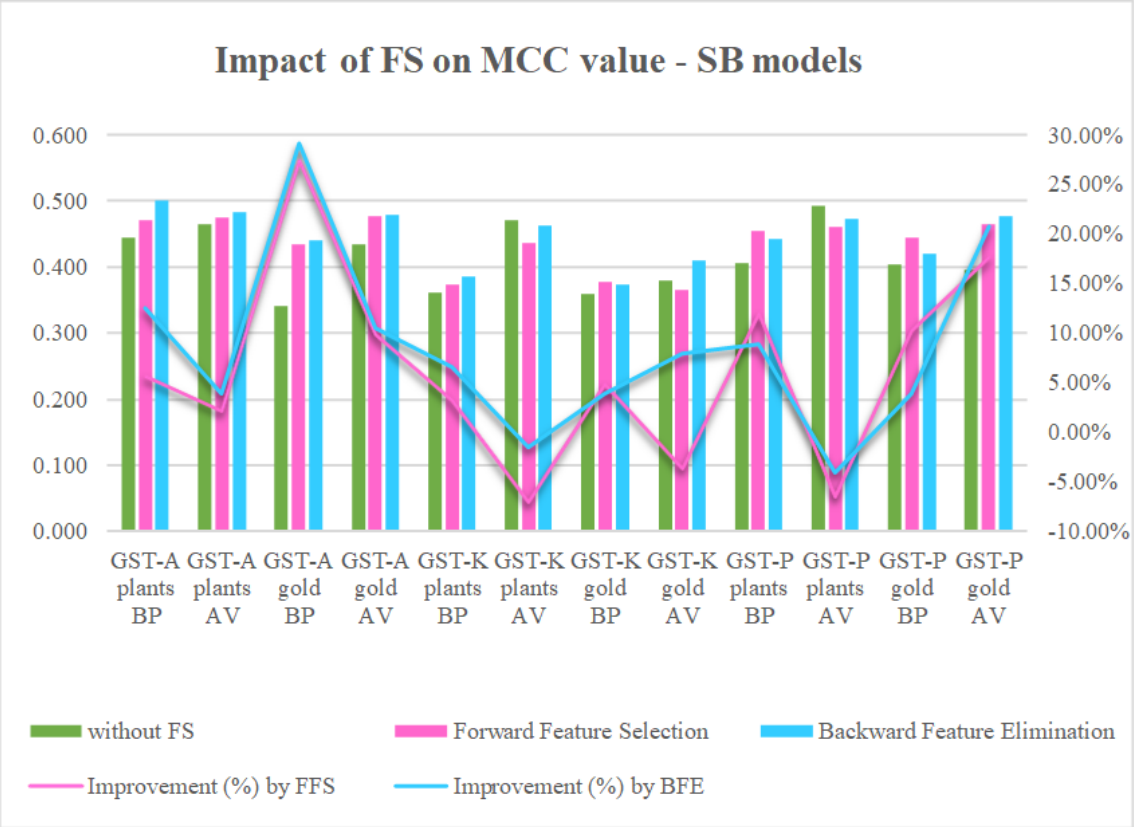

Figure 3.8: Trend of MCC value through all isoenzymes, without and with FS. On the left vertical axis, the MCC value is reported for each scenario investigated, while on the right axis the percentage of improvement due to FS step. The histograms refer to the MCC value, without (in green) and with FS step (in magenta with FFS strategy and in cyan with BFE one). The two broken lines refer to the respective percentual improvement.

Generally, in most cases, the FS step improved the model performance, except for the case of AV approach with PLANTS in isoenzymes GST-K and GST-P. From a binding pocket (BP) analysis, carried out with Fpocket [91], GST-K and GST-P BP volumes resulted very similar ($\simeq$5666 Å^3 and 6180 Å^3, respectively), while the volume of the GST-A BP has been calculated at $\simeq$8391 Å^3. The smaller BPs of GST-K and GST-P could be the reason why these isoenzymes underperformed in these scenarios.

Comparing the two FS approaches with the results obtained without FS, in 9 out of 12 scenarios the FFS demonstrated a mean improvement of 10.34%, while in 10 out of 12 scenarios the BFE improved the model performance with a mean improvement of 10.78%. In conclusion, regarding SB studies, BFE showed a slightly better improvement ability compared to FFS strategy and in 6 out of 12 scenarios works better than FFS.

Similarly to the global model generated for LB studies, in order to combine the multiple SB predictions, a consensus voting approach was chosen.

In A consensus, class prediction 1 is composed of all instances predicted as 1 by at least one isoenzyme, while all instances predicted as 0 simultaneously by all isoenzymes are considered in prediction class 0.
In B consensus, class prediction 1 is composed of all instances predicted as 1 by at least two isoenzymes, while all the others are put in prediction class 0.
In C consensus, class prediction 1 is composed of all instances predicted as 1 simultaneously by all three isoenzymes.

As shown in Table 3.8, the results highlighted a better performance of PLANTS docking once again. In detail, A and B were the two enhancing approaches, where at least one or two predictions are included, respectively; the higher increment was reached by A consensus with AV approach by PLANTS docking – with a 15.44% improvement.

Consensus	**PLANTS**		**GOLD**	
	BP	**AV**	**BP**	**AV**
Voting ≥ 1	0.447	0.537	0.417	0.382
Voting ≥ 2	0.426	0.499	0.091	0.429
Voting ≥ 3	0.380	0.425	0.471	0.365

Table 3.8: Consensus voting results, for PLANTS and GOLD docking, considering BP and AV approaches.

3.2.4.4 Hardness and Predictive Power

The stereo-electronic ELUMO descriptor showed a peculiar correlation with the predictive power of the generated model. Indeed, when the dataset is composed of only poor electrophiles (high ELUMO), the predictive power proved to be low, supporting the theory where these molecules cannot react with GSH due to their poor electrophilicity and so neither the structural approach can predict the reaction occurrence.
When the considered dataset was populated with medium electrophile molecules (decreased ELUMO) too, the predictive power showed its maximum potential; with a dataset including hard electrophiles (low ELUMO) too, the model starts to be less performant since the spontaneous ability of these molecules to react with GSH in a non-enzymatic way.

This trend is highlighted in Figure 3.9: the first molecules on the left have high ELUMO, so poor electrophiles, which decreases proceeding to the right.

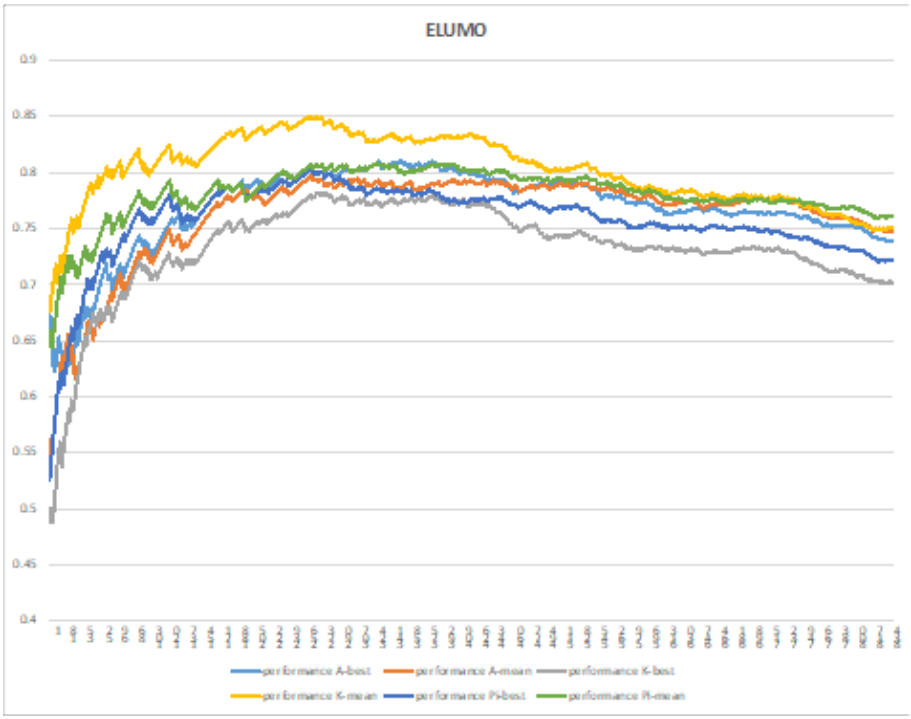

Figure 3.9: Cumulative average showing the correlation between ELUMO and predictive power.

3.2.5 Conclusions

The aim of this study was to evaluate the contribution of structure-based information involving the enzymatic structure as well as the metabolite structure in the metabolism prediction.

Globally, the study highlighted that LB studies reported better predictive power rather than SB ones. Among LB models, the best performance was obtained with the VEGA ZZ set of descriptors, physicochemical ones. All models improved their predictive ability with the variable selection, thanks to the statistical noise reduction.

The reason why SB studies resulted worse than LB ones could be attributed to many reasons. The traditional molecular docking simulation involved a static treatment of the protein-ligand interactions without considering the solvation and entropic effects of the system. One more thing to consider is the promiscuity of the metabolic enzymes. Since they must metabolise so structurally diverse molecules, their binding site is not well described by a static docking study. A last reason to be mentioned is that docking calculations can simulate only the recognition process while the following (and crucial) enzymatic step is not considered. Stated differently, a given molecule can yield a good complex with the enzyme while failing to undergo the following metabolic reaction. It is important to highlight that this situation cannot be accounted for by canonical docking simulations.

Going deeper into the molecular properties analysis and investigating the enhancing role of molecular docking studies, it emerged that electrophilicity plays a crucial role in determining the occurrence of the GSH reaction and its spontaneous or catalytic-mediated way.
Regarding GST activity, electrophiles can be subdivided into good or poor enzymatic substrates. When combining molecule softness and GST activity, electrophiles can be categorised into four groups. Specifically, they may react with GSH solely through spontaneous processes (soft and poor substrates), through both spontaneous and enzymatic mechanisms (soft and good substrates), or may be completely incapable of reacting with GSH (hard and poor substrates).
In this work, the correlation between electrophilicity and predictive power of generated models was highlighted, proving and confirming the satisfactory performance of the chosen approach. The results were obtained thanks to curated and accurate metabolic data, which certainly indicates a high reliability reached level of the

prediction.

Globally, this study represents a starting point for future investigations about the role of GST in GSH conjugation and future insights combining docking plus quantum mechanical calculations should provide better performances by considering both recognition and catalytic events.

3.3 MetaSpot

The content of this Section has been object of one publication (from which it was entirely taken) and one communication:

- Angelica Mazzolari, Pietro Perazzoni, **Emanuela Sabato**, Filippo Lunghini, Andrea R. Beccari, Giulio Vistoli, & Alessandro Pedretti. (2023). MetaSpot: A General Approach for Recognising the Reactive Atoms Undergoing Metabolic Reactions Based on the MetaQSAR Database. International Journal of Molecular Sciences, 24(13). https://doi.org/10.3390/ijms241311064
- **Emanuela Sabato**, Pietro Perazzoni, Giulio Vistoli, Angelica Mazzolari, Alessandro Pedretti. MetaClass and MetaSpot: Machine Learning-based tools for metabolism prediction. 23rd European Symposium on Quantitative Structure-Activity Relationship (EuroQSAR), Heidelberg, September 26-30, 2022, pag. 160 (Poster) [p̂resenting author].

3.3.1 Introduction

The global applicability of MetaQSAR was further confirmed by a recent study, in which properly extracted datasets allowed the development of models to predict the occurrence of almost all the reaction classes plus some suitably populated subclasses by using the Random Forest (RF) classification algorithm. These predictive models were implemented in a freely released tool, MetaClass, able to predict which metabolic reactions undergo an input molecule.[92]

The here-reported study can be seen as a step further since it is based on the same MetaQSAR-based datasets, and its aim involves the development of RF-based

models to recognise the specific sites of metabolism (SoM) for all the reaction classes and subclasses already considered by the MetaClass approach. Similarly to what was already done for MetaClass, the here-developed predictive models are released into a suite of freely available scripts, called MetaSpot, to predict which atoms undergo specific metabolic reactions for a given input substrate.

3.3.2 Methods

3.3.2.1 Utilised Metabolic Data

The study utilised the same first-generation metabolic reactions extracted from the MetaQSAR database [93] and already utilised in the previous MetaClass study [94]. In detail, they include 3788 metabolic reactions, which involve 2787 different substrates. As done in the previous study, all the collected metabolic reactions were utilised during the model generation to cover the widest possible chemical space of the substrates. Based on the MetaClass results, the substrates are here simulated in their neutral form. A detailed description of the set-up of the collected substrates and the calculation of the utilised descriptors can be found elsewhere.[94] Briefly, the compounds were optimised by the PM7 semi-empirical method which also allowed the calculation of an extended set of stereo-electronic descriptors. For each molecule, a set of physicochemical and geometrical descriptors were calculated by the VEGA ZZ suite of programs. Atom typing involved the Kier-Hall E-states [92] as well as the lipophilic atomic increments of Broto and Moreau [95].

3.3.2.2 Generation of Models with Atom Typing (First Round)

As described above, the predictive analyses involved all the classes and subclasses, including at least 50 instances. For each case, two predictive analyses were performed including or excluding the atom types among the utilised descriptors. The analyses including atom types were performed by using an *ad-hoc* script for the VEGA environment that:

a) extracts from MetaQSAR the substrates for a given class or subclass
b) identifies within the substrates the reactive and non-reactive atoms and

c) calculates for the included substrates a set of atom- and ligand-based descriptors.

For each considered class and subclass, the analyses comprised all collected atoms and involved the following tasks carried out by the above-mentioned script:

d) performing the feature selection to reduce the number of considered descriptors
e) randomly under-sampling the NR atoms to obtain balanced datasets
f) developing the corresponding predictive model by using the RF algorithm.

Considering the very high number of NR atoms in all performed analyses, tasks (e) and (f) were repeated 100 times to minimise the randomness of the obtained results. Here, and in the analyses without atom types, the models were developed by using the RF algorithm, as implemented in the WEKA 3.8.6 software [96], by applying the default parameters since, in the previous MetaClass study, these conditions afforded the best results.[94] The feature selection was also carried out by WEKA based on both the BestFirst and the WrapperSubsetEval algorithms, as previously described.[94] All the models generated in the first round of predictions were transformed in C-based scripts by using the Tree2C program [97]. Briefly, Tree2C converts a tree model as generated by WEKA into a C-based script for the VEGA program. The so-generated script predicts the reactive centres for the molecule loaded within the VEGA workspace by applying the tree model and by calculating on the fly all the required descriptors.

3.3.2.3 Generation of Models Without Atom Typing (Second Round)

To assess the capacity to predict the reactivity of the sites of metabolism regardless of their atom types, the second set of predictions was developed by considering the NR atoms with the same atom type(s) of the reactive centres. For each analyses class and subclass, a specific dataset was manually collected by combining the reactive atoms with an equal number of randomly selected non-reactive atoms having the same atom type(s) of the reactive ones. The so-collected datasets undergo model generation and feature selection by adopting the same procedures described above for the first round of predictions.

3.3.3 Results

3.3.3.1 MetaQSAR-Based Datasets

The study was based on the same first-generation metabolic reactions utilised by the previous study.[94] Specifically, the analyses involved 3788 metabolic reactions, which are classified into 3 major classes (*i.e.* redox reactions, hydrolyses and conjugations), 21 classes and 101 subclasses. Predictive models were generated for all classes and subclasses which include at least 50 metabolic reactions. For each simulated dataset, attention was focused only on the molecules which undergo the corresponding metabolic reaction and considering as reactive atoms (R), the atoms involved in the biotransformation and as non-reactive atoms (NR) all the remaining atoms without exceptions. The choice of focusing on the substrates to collect the non-reactive atoms had two major reasons: (a) avoiding extremely unbalanced datasets (and however the number of NR atoms is still markedly higher); (b) minimising the number of false negative NR atoms.

For each class and subclass, the performed predictive analyses can be subdivided into two steps. In the first step, the utilised datasets comprised all collected NR and R atoms. Each dataset was balanced by randomly undersampling the majority NR class and this task was repeated 100 times to minimise the randomness. The second step was based on specifically collected datasets comprising only atoms of the same atom type(s) (according to Kier-Hall E-states [95]) of the reactive centres to assess the capacity of the developed models to discriminate between reactive and the non-reactive centres of the same atom type. The models developed by the two steps can be sequentially utilised since the models of the first step are intended to perform a coarse filter to discard those atoms which cannot undergo a given biotransformation, while the models of the second round are more finely tuned to recognise the truly reactive centres.

3.3.3.2 Prediction of the Reactive Atoms for the Reaction Classes

Figure 3.10 and Table 3.9 (see 3.3.5) report the performances for the considered classes of metabolic reactions as reached by the first round of predictions in which

the NR atoms are randomly extracted from all the possible NR atoms belonging to the collected substrates. Table 3.9 reports the overall performance as computed by considering both classes. For each metabolic reaction, Table 3.9 comprises the mean and range values of the performance metrics as obtained by repeating the model generation 100 times. For the three major classes of reactions, Figure 3.10 and Table 3.9 also include the corresponding mean values as well as the overall means.

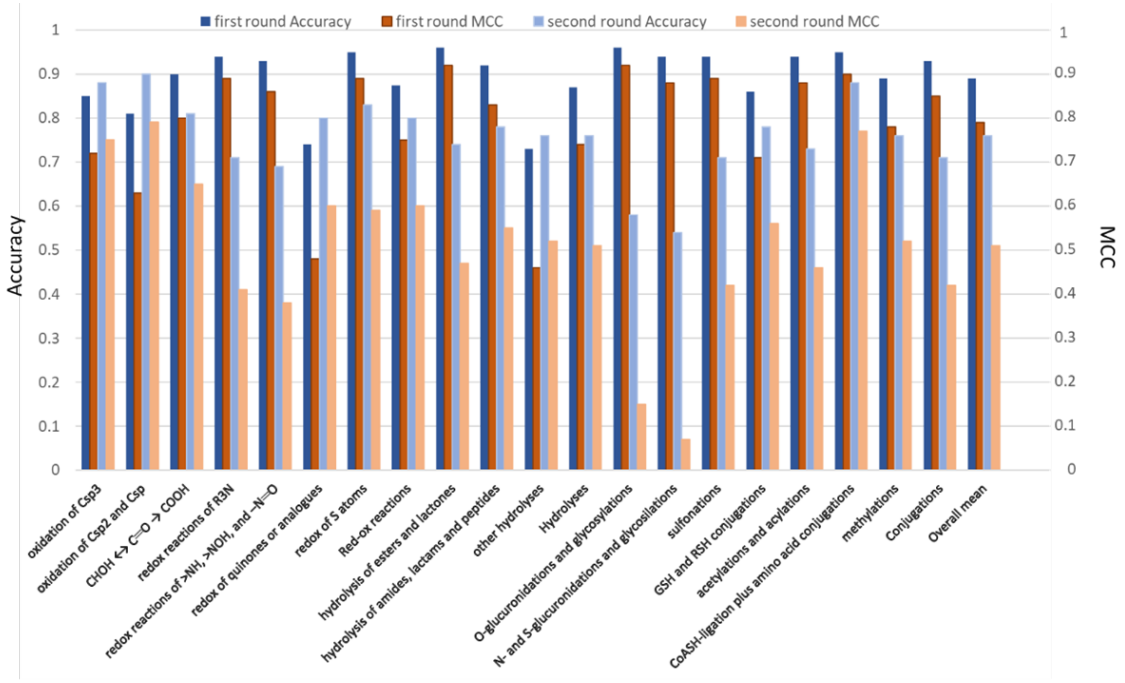

Figure 3.10: Performances (as described by Accuracy and MCC values) reached by the classification models for the classes of metabolic reactions in the two rounds of predictions. The average values for the three major classes (*i.e.* redox, hydrolyses and conjugations) as well as the overall means are also reported.

Figure 3.10 reveals satisfactory performances for almost all considered metabolic reactions with the conjugations which afford, on average, the best results, while redox reactions and hydrolyses yield slightly worse performances. The reported accuracy values indicate that, on average, 90% of reactive atoms are correctly predicted. Only two reactions classes show accuracy values comprised between 0.7 and 0.8 with 11 out of 17 accuracy values being greater than 0.9. The MCC values also show similar trends: 11 out of 17 metabolic reactions show MCC ≥ 0.8 and only 2 classes show MCC < 0.5.

Overall, the models appear to be quite stable as confirmed by the accuracy ranges, which are lower than 0.02 in 9 cases out of 17, as well as by the MCC ranges, which are lower than 0.15 for 10 classes. Only two classes, *i.e.* redox of quinones and other hydrolyses, show modest performances in terms of both mean and range values. The

obtained outcomes do not reveal significant correlations between performance and number of instances in each class. This suggests that all the considered classes are populated enough to develop reliable models. The observed performance differences are thus ascribable to the intrinsic complexity of each metabolic reaction rather than to the chemical space covered by its substrates.

Figure 3.11 and Table 3.10 (see 3.3.5) report the analysis of the feature importance for the 17 predicted biotransformations:Table 3.10 compiles the selected descriptors for each model, while Figure 3.11 focuses on their frequencies. The first consideration involves the number of selected features which ranges from 1 to 11 with an average value equal to 4.8. The rather limited number of included variables should assure a reliable robustness for the generated models avoiding overfitting biases. Figure 3.11 emphasises the remarkable role played by atom typing as evidenced by the Kier and Hall E-states [95], which appear in 12 models out of 17. In fact, the Broto's and Moreau's atomic increments for lipophilicity [98] can also be seen as a sort of atom typing in which the atom types are indirectly encoded by their atomic increments. Indeed, replacing the Broto's and Moreau's atomic constants with the corresponding atom types provides truly superimposable overall performances. As a matter of fact, all models include at least one atom type and 7 models out of 17 include both atom types.

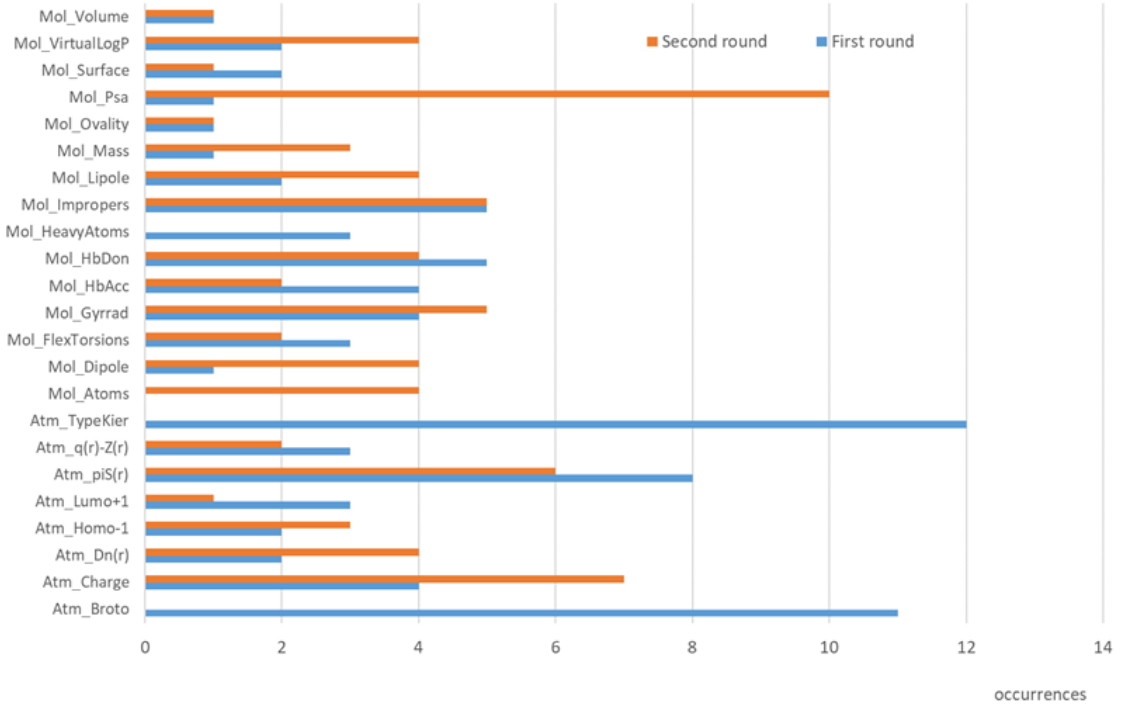

Figure 3.11: Occurrences of the descriptors as derived by analysis of feature importance for the predictive models for the classes of reactions in the two rounds of predictions.

Among the stereo-electronic parameters, atomic self-polarisability (πS) and atomic charge are the most frequently included descriptors. The role of atomic self-polarisability agrees with previous studies which exploited such a descriptor to predict the sites of metabolism [99] and, more in general, to rationalise the chemical reactivity. Concerning the physicochemical descriptors, the number of H-bonding groups plays a key role reasonably as it encodes for the substrate's capacity to interact with the enzymes. In eight models, there is at least one parameter related to molecular size, which may describe the capacity of a given molecule to be accommodated within the enzymatic cavity. The highlighted marginal role of the virtual log P descriptor (see Table 3.10) can be explained by considering that lipophilicity is already encoded by Broto's and Moreau's atomic increments.

Taken together, the obtained performances suggest that satisfactory predictions can be achieved by correctly recognising the atom types which can undergo a given metabolic reaction. Even though almost all models also include stereo-electronic and/or physicochemical descriptors which should encode for the intrinsic reactivity and accessibility of each atom, the primary role played by atom typing might result in a high number of false positives since there is the risk that all atoms belonging to a given atom type are predicted as reactive regardless of their actual reactivity. Hence, the second round of predictions was performed by excluding atom typing to assess whether the stereo-electronic and physicochemical descriptors alone can recognise the reactive atoms regardless of their type. To do this and for each considered class, a balanced dataset was generated by randomly collecting only NR atoms having the same type(s) of the reactive sites.

Figure 3.10 and Table 3.11 (see 3.3.5) report the performances obtained by this second round of predictions and compare them with those reached during the first round of predictions. While the performances of the first round were similarly satisfactory with limited differences between the considered classes, the second round reveals significant differences, and some models prove to be unsatisfactory. In detail, Figure 3.10 and Table 3.11 show that: (a) the sites of metabolism of the redox reactions (especially those involving carbon atoms) can successfully be recognised, (b) the reactive atoms of hydrolyses are recognised although with slightly worse performances compared to Table 3.9 especially for the hydrolysis of esters, while (c) the sites of conjugations are predicted with difficulty and the developed models show markedly worse performances compared to the first round of predictions. Overall, only 3 reaction classes show enhanced performances in this second round and the drop in the MCC value is greater than 0.4 in 6 cases. The accuracy values indicate

that there are 3 classes for which the correctly identified reactive atoms are lower than 70%. In detail, the reported performances evidence that the sites of glucoronidation cannot be conveniently predicted and the redox reactions on nitrogen atoms are predicted with marked difficulty.

Figure 3.11 and Table 3.12 (see 3.3.5) highlight the role of the selected features for this second round of predictions. In addition, the developed models include a limited number of variables which range from 3 to 9 with an average value equal to 4.9. Concerning the stereo-electronic parameters, these models confirm the major role of self-polarisability and atomic charges, while greater differences are seen among the physicochemical descriptors. Figure 3.11 emphasises the key role of descriptors related to H-bonding as confirmed by the PSA parameter which is included into 10 models. Descriptors related to the molecular size also play a remarkable role since they are included in 15 models out of 17.

Taken together, the results of this second round of predictions suggest that the reactive atoms of redox reactions largely depend on stereo-electronic factors, which are conveniently captured by the considered descriptors. In contrast, reactive atoms involved in hydrolyses and conjugations seem to be influenced by different factors not completely encoded by the considered descriptors. Future studies involving additional descriptors and/or docking simulations could enhance the predictive models for these classes.

3.3.3.3 Prediction of the Reactive Atoms for the Reaction Subclasses

Figure 3.12 and Table 3.13 (see 3.3.5) show the performances for the considered subclasses as reached by considering all possible NR atoms in the first round of predictions. A bird's eye view of the obtained results reveals the subclasses provide satisfactory performances in agreement with those already reported for the metabolic classes. The remarkable results are confirmed by the accuracy values which indicate that the correctly predicted centres are $\geq 90\%$ for 16 subclasses out of 23, while the accuracy is less than 0.8 only in two cases. MCC values provide similar trends with 15 out of 23 MCC values greater than 0.8. Only the oxidation of aryls and phenols and the addition-elimination reactions of glutathione yield less than satisfactory models. The comparison between Figure 3.10 and Figure 3.12 shows that the finer

classification by subclasses has a beneficial role for the hydrolyses, while redox reactions and conjugations show roughly comparable performances. The models for subclasses also reveal an appreciable stability as evidenced by the accuracy range which is < 0.1 for 15 cases out of 19 (see Table 3.13).

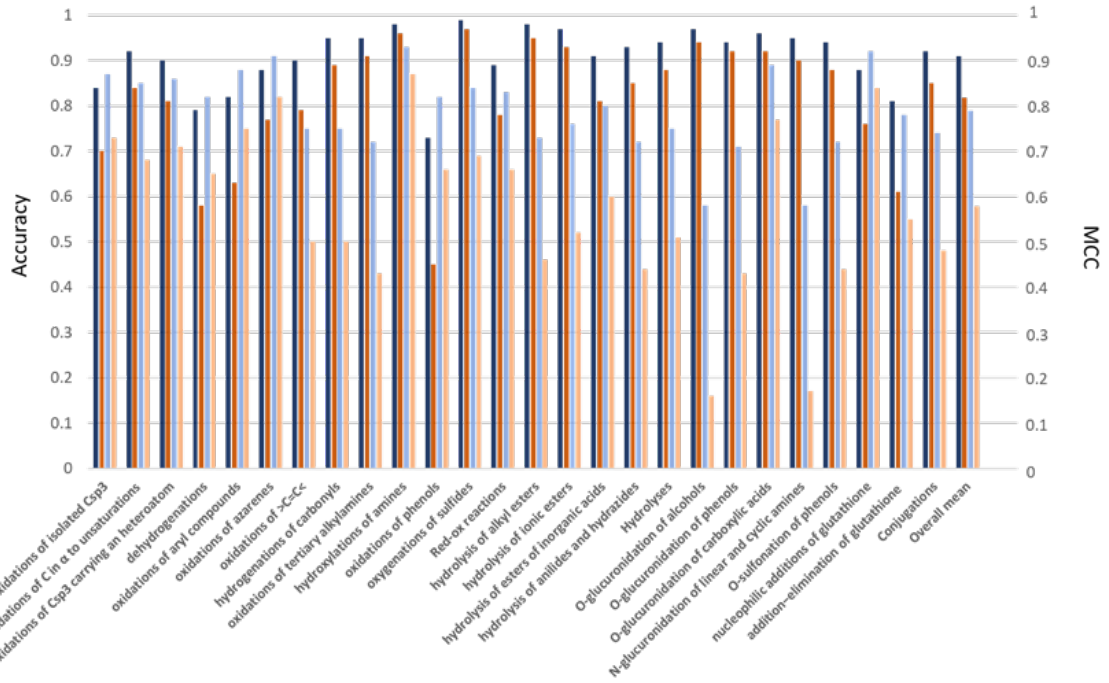

Figure 3.12: Performances (as described by Accuracy and MCC values) reached by the classification models for the subclasses of metabolic reactions in the two rounds of predictions. The color code of the columns is the same as Figure 3.11 (*i.e.* blue and brown for accuracy and MCC of the first round; azure and light brown for accuracy and MCC of the second round). The average values for the three major classes (*i.e.* redox, hydrolyses and conjugations) as well as the overall means are also reported.

Table 3.14 (see 3.3.5) and Figure 3.13 report the results of the feature importance for the 19 considered subclasses: Table 3.14 compiles the selected descriptors, and Figure 3.13 shows their frequencies. Here also, the number of involved descriptors is rather limited, a fact that should avoid overfitting conditions. In detail, this ranges from 1 to 9 variables with an average equal to 5.0. The analysis of the selected features further confirms the remarkable role of the atom typing, since all models include either Kier-Hall E-states [95] or Broto's and Moreau's atomic increments [98], and 12 models out of 19 include both atom types. On average, these results emphasise a greater role of the atom typing in subclasses compared to classes. This finding can be explained by considering that the finer classification of the metabolic reactions enhances the capacity of the atom types to detect the reactive centres.

Concerning the stereo-electronic descriptors, the developed models confirm the

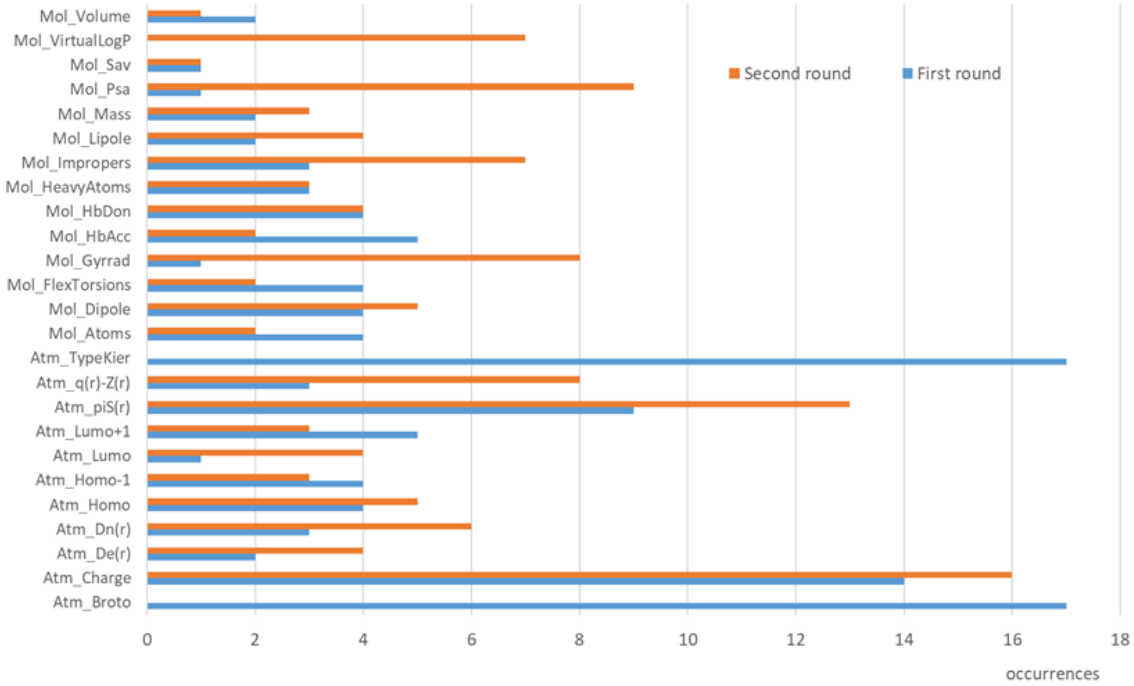

Figure 3.13: Occurrences of the descriptors as derived by analysis of feature importance for the predictive models for the subclasses of reactions in the two rounds of predictions.

key role of atomic charges and self-polarisability (πS) in describing the chemical reactivity. In addition, the HOMO/LUMO energies appear 13 times, thus underlining the relevant role played by nucleophilicity and electrophilicity in drug metabolism. The frequencies of the physicochemical descriptors agree with what was observed in the previous models and emphasise the marked role of H-bonds and molecular size (the related descriptors appear 10 times in both cases) which encode for interacting capacity and accessibility, respectively.

Figure 3.12 and Table 3.15 (see 3.3.5) detail the performances reached by the subclasses in the second round of model generation in which the utilised datasets include only NR atoms of the same type(s) of the reactive centres. As already observed above, the second round of predictions reveals marked differences in the performances of the considered subclasses. On average, the reactive atoms for the redox reactions are conveniently predicted with almost all accuracy values greater than 0.8 and 4 subclasses (out of 12) show even better performances compared to the first round. Table 3.15 confirms that the atoms undergoing redox reactions can be suitably recognised based on their stereo-electronic features and the finer classification of these reactions (as done in subclasses) improves the recognition of the reactive atoms.

The finer classifications of hydrolyses do not improve the resulting performances,

which remain for all considered subclasses around 0.75 (as already seen in Table 3.11). This output suggests that the detection of the labile functional groups (as made by atom types) plays a key role in determining the performances for these metabolic reactions, and the stereo-electronic properties are not completely effective in recognising the labile moieties. Regarding glucuronidations, Figure 3.12 reveals that a finer classification improves the performances for predicting metabolic reactions of phenols and carboxylic acids, while the reactions of alcohols and the N-glucuronidations are still unsuitably predicted. Finally, Figure 3.12 shows that the poor performance observed for the conjugations with GSH in Figure 3.10 is ascribable to the reactions which occur via GSH addition and elimination, while the nucleophilic GSH additions can be conveniently predicted.

As previously seen, Table 3.16 (see 3.3.5) and Figure 3.13 detail the feature importance for this second round of predictions on the 19 considered subclasses. This analysis confirms the pivotal role of atomic charges and self-polarisability (πS) and evidences the enhanced relevance of the atomic charge density the role of which in predicting chemical reactivity is well-documented in literature.[100] Among the physicochemical descriptors, Figure 3.13 emphasises the relevance of H-bonds (as especially encoded by PSA) and molecular size with 15 occurrences in both cases.

3.3.3.4 Discussion

Since both studies are based on the same MetaQSAR-based training sets and roughly involve the same set of descriptors, Figure 3.14 compares the performances (as expressed by the MCC values) of the here-reported models to recognise the reactive atoms with those reported in the previous MetaClass study to predict the occurrence of a given metabolic reaction. For simplicity, the analysis is focused on the predictive models as generated for the classes of metabolic reactions and compares the MetaClass performances with those here reported for both rounds of predictions.

Figure 3.14 clearly emphasises that the reactive atoms involved in most metabolic reactions can be conveniently recognised by RF-based classifiers in which atom typing plays a key role (first round). For almost all classes these models outperform the other predictions suggesting that atom typing has a markedly greater role in the recognition of reactive atoms rather than in predicting the occurrence of a given biotransformation. The relevance of atom typing comes as no surprise when

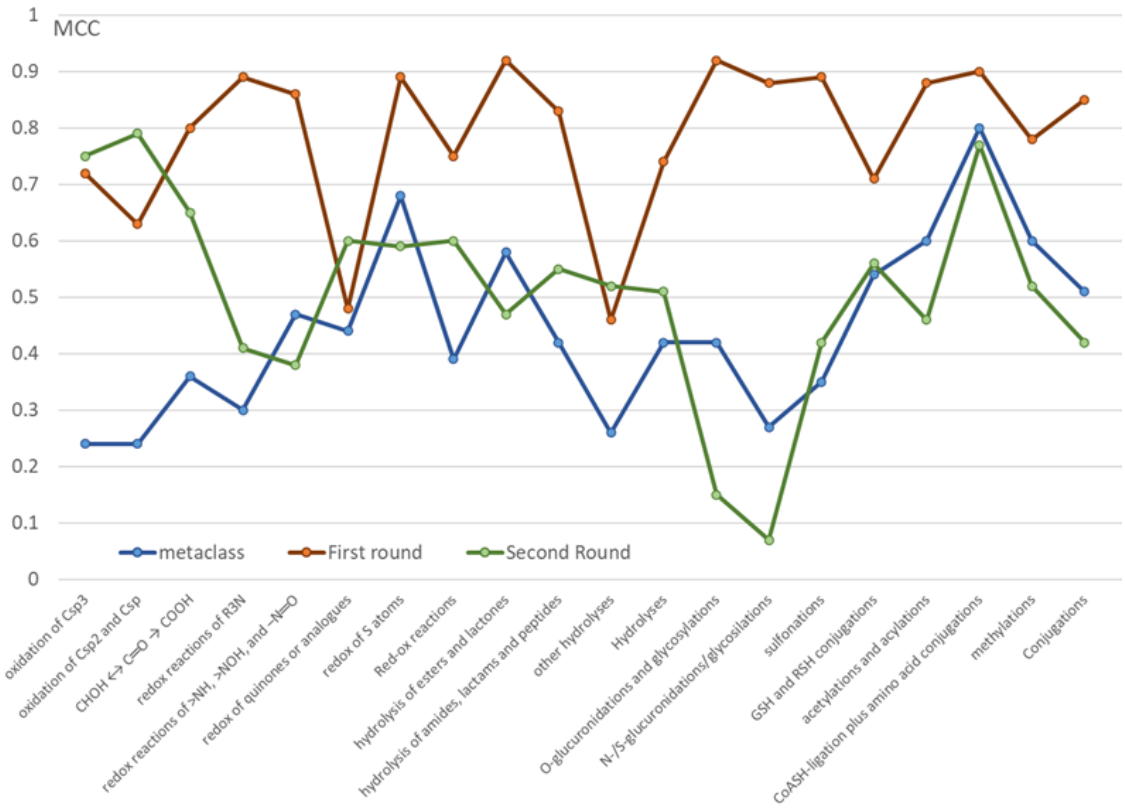

Figure 3.14: Comparison of the performances (expressed as MCC values) of the models generated by the previous MetaClass project with those here developed by the two rounds of predictions.

considering that most available global methods were based on knowledge-based rules and atom typing is probably the most effective way to transform these qualitative rules into computationally tractable molecular descriptors. Even though all the proposed models also include stereo-electronic descriptors to account for the chemical reactivity of each atom, the key role played by atom types can lead to a significant number of false positive since the models tend to predict that all atoms of a given type can yield the corresponding metabolic reaction regardless of their reactivity. Stated differently, the models based on atom typing prove successful in discarding the atoms which cannot undergo a given biotransformation, while being less performing in recognising the truly reactive atoms among those which can undergo a given metabolic reaction.

Interestingly, the second round of prediction which do not include the atom typing, provide, on average, worst performances which shows a certain degree of similarity with those reached by the MetaClass models. This confirms that atom typing has a very modest role in predicting the occurrence of a given metabolic reaction and both sets of models similarly depend on the included physicochemical

and stereo-electronic descriptors. While showing lower MCC values, many developed models of the second round exhibit satisfactory performances thus indicating that the recognition of the sites of metabolism based on reactivity descriptors is yet possible and deserves further efforts to enhance the resulting performances, especially for those classes and subclasses which provided here unsatisfactory predictive models (see above).

Indeed, and while considering the generally satisfactory results reported here, this study can be seen as an encouraging starting point which can be surely enhanced by extending the number and the informative richness of the considered descriptors. The descriptors utilised in this study were indeed chosen to cover a significant space of the (structural, physicochemical and stereo-electronic) molecular properties while being computationally simple and not time-expensive. This choice was made by considering the high number of substrates included in the MetaQSAR-based datasets and to develop simple predictive models to be quickly applied to new molecules. Notwithstanding this, one may figure out that a wider and tailored set of descriptors could yield improved predictive models. For example, the observed relevance of various stereo-electronic descriptors (here computed by semi-empirical calculations) suggests that the inclusion of similar parameters computed by DFT methods, although computationally markedly more demanding, might have a beneficial effect. The role played in many models by molecular descriptors encoding for the interaction capacities (*e.g.*, H-bonding, molecular size and accessibility) suggests that targeted docking simulations could describe the binding process in a more direct and informative way. The enrichment of the considered descriptors may involve also the atom typing by including extended atom types and/or fingerprints. Nevertheless, the enrichment of this kind of descriptors should be cautiously pursued since an increased role of atom typing could increase the number of false positives.

In more detail, the results of this study (especially concerning the second round of predictions) can be summarised as follows:

- Redox reactions on carbon atoms are conveniently predicted even without atom typing; this outcome indicates that their reactive atoms mostly depend on the considered stereo-electronic descriptors (*e.g.*, atomic charges and self-polarisability)
- Redox reactions involving nitrogen atoms (and to minor extent Sulphur atoms) are satisfactorily predicted only when using atom types, thus suggesting that their reactivity depends on different factors compared to carbon atoms

- Hydrolysis reactions yield markedly more accurate predictions when including atom types which reasonably allow an easy detection of the labile groups
- Conjugations with glucuronic acid are substantially unpredictable without atom types; this finding suggests that the reactivity of the involved centres depends on stereo-electronic factors not included in the considered descriptors
- Most reactions with glutathione can be successfully predicted even without atom types, a result which can be explained by considering that these conjugations depend on the here parameterised electrophilicity of the reactive centres.

As mentioned above, the two sets of predictions (*i.e.* with and without atom types) can be synergistically combined since the models including atom types can be used to exclude the atoms which cannot undergo the considered biotransformation, while the models without atom types should recognise the truly reactive centres.

3.3.4 Conclusion

The present study describes the MetaSpot approach, which is based on the metabolic data extracted from the MetaQSAR database and allows the predictions of the reactive centres for almost all metabolic reactions. Apart from a few documented exceptions, all the models developed within the MetaSpot project provided satisfactory performances (even without atom typing), emphasising the possibility of conveniently recognising the reactive atoms. The MetaSpot project can be seen as the natural extension of the already published MetaClass study [94] since MetaClass predicts which metabolic reactions undergo a given molecule, while MetaSpot predicts which are the involved sites of metabolism. The two approaches can be combined and can have a mutually validating role since the reactive centres predicted by MetaSpot for a given biotransformation can be confirmed if MetaClass predicts that the substrate can undergo the biotransformation. Similarly, a metabolic reaction predicted by MetaClass can be confirmed if MetaSpot finds at least one reactive atom for this reaction.

As detailed in the discussion, this study can be seen as an encouraging starting point, and the reported models could be improved both by enriching the arsenal of the considered descriptors and by including structure-based approaches (such as docking simulations), which can simulate the recognition between the substrate and the involved enzymes. The predictive models could also be improved by extending

and refining the collected metabolic data. Enriched metabolic data should maximise the chemical space covered by the collected reactions, thus extending the predictive power of the developed models. In addition, enriched metabolic data could be utilised to generate suitable external sets for a more precise validation and tuning of the selected predictive models.

3.3.5 Supplementary Information

			Accuracy		MCC		AUC	
Class ID	Cases	Description	Mean	Range	Mean	Range	Mean	Range
01	3050	Oxidation of Csp3	0.85	0.005	0.72	0.04	0.87	0.03
02	1798	Oxidation of Csp2 and Csp	0.81	0.008	0.63	0.07	0.83	0.06
03	278	CHOH $\longleftrightarrow$ C=O $\longrightarrow$ COOH	0.90	0.02	0.80	0.19	0.95	0.06
05	252	Redox reactions of R3N	0.94	0.02	0.89	0.14	0.98	0.05
06	330	Redox reactions of >NH, >NOH, and –N=O	0.93	0.01	0.86	0.12	0.97	0.03
07	198	Redox of quinones or analogues	0.74	0.03	0.48	0.30	0.81	0.16
08	274	Redox of S atoms	0.95	0.01	0.89	0.09	0.98	0.04
Main class 1		**Redox reactions**	**0.87**	**0.01**	**0.75**	**0.14**	**0.91**	**0.06**
11	534	Hydrolysis of esters and lactones	0.96	0.01	0.92	0.09	0.98	0.02
12	258	Hydrolysis of amides, lactams and peptides	0.92	0.02	0.83	0.21	0.97	0.06
14	168	Other hydrolyses	0.73	0.03	0.46	0.33	0.80	0.18
Main class 2		**Hydrolyses**	**0.87**	**0.02**	**0.74**	**0.21**	**0.92**	**0.09**
21	800	O-glucuronidations and glycosylations	0.96	0.05	0.92	0.05	0.96	0.04
22	298	N- and S-glucuronidations and glycosilations	0.94	0.01	0.88	0.15	0.95	0.06

23	244	Sulfonations	0.94	0.02	0.89	0.18	0.97	0.07
24	320	GSH and RSH conjugations	0.86	0.02	0.71	0.14	0.92	0.05
25	140	Acetylations and acylations	0.94	0.02	0.88	0.24	0.98	0.05
26	130	CoASH-ligation plus amino acid conjugations	0.95	0.02	0.90	0.15	0.98	0.04
27	110	Methylations	0.89	0.04	0.78	0.33	0.93	0.13
Main class 3		**Conjugations**	**0.93**	**0.02**	**0.85**	**0.18**	**0.96**	**0.06**
Overall means			**0.89**	**0.02**	**0.79**	**0.17**	**0.93**	**0.07**

Table 3.9: Performance metrics (Accuracy, MCC and AUC values) for the 17 considered classes of metabolic reactions in the first round of predictions. For each metric, the reported mean and range values are derived by repeating 100 times the random undersampling of the NR atoms.

Class ID	Description	Selected Features
01	Oxidation of Csp3	Atm_Broto, Atm_Kier, Mol_HbDon
02	Oxidation of Csp2 and Csp	Atm_Kier, Mol_HbDon
03	CHOH $\longleftrightarrow$ C=O $\longrightarrow$ COOH	Atm_Broto, Atm_Kier, Atm_Charge, Atm_q(r)-Z(r), Atm_Lumo+1, Mol_Lipole
05	Redox reactions of R3N	Atm_Broto, Atm_Dn(r), Atm_Homo-1, Mol_Rotors
06	Redox reactions of >NH, >NOH, and –N=O	Atm_Kier, Atm_piS(r), Atm_Homo-1, Atm_Lumo+1, Mol_Atoms, Mol_HeavyAtoms, Mol_Ovality, Mol_PSA
07	Redox of quinones or analogues	Atm_Broto, Atm_piS(r), Mol_HbAcc, Mol_HbDon, Mol_Impropers, Mol_Surface
08	Redox of S atoms	Atm_Broto, Atm_Kier, Atm_piS(r), Mol_Gyrrad, Mol_Lipole
11	Hydrolysis of esters and lactones	Atm_Broto, Atm_Kier, Atm_Charge, Atm_Dn(r), Mol_Dipole, Mol_Impropers, Mol_Sas
12	Hydrolysis of amides, lactams and peptides	Atm_Broto, Atm_Kier, Atm_piS(r), Mol_Rotors, Mol_Gyrrad, Mol_LogP

14	Other hydrolyses	Atm_Kier, Atm_q(r)-Z(r), Atm_piS(r), Mol_Lumo, Mol_Gyrrad, Mol_HbDon
21	O-glucuronidations and glycosylations	Atm_Kier
22	N- and S-glucuronidations and glycosilations	Atm_Broto
23	Sulfonations	Atm_Kier, Mol_HbAcc
24	GSH and RSH conjugations	Atm_Broto, Atm_Charge, Atm_piS(r), Mol_Impropers
25	Acetylations and acylations	Atm_Kier, Atm_piS(r), Mol_Impropers
26	CoASH-ligation plus amino acid conjugations	Atm_Broto, Atm_Kier, Atm_Charge, Atm_piS(r), Atm_Lumo+1, Mol_Gyrrad, Mol_HbAcc, Mol_HbDon, Mol_Impropers, Mol_LogP, Mol_Volume
27	Methylations	Atm_Broto, Atm_q(r)-Z(r), Mol_Rotors, Mol_HbAcc, Mol_HeavyAtoms, Mol_Mass

Table 3.10: Descriptors selected by the feature importance analyses for the models developed for the 17 considered classes of metabolic reactions in the first round of predictions (here and in the following Tables the prefixes Atm and Mol stand for atom- and ligand-based descriptors).

Class ID	Description	Accuracy	MCC	AUC	ΔMCC	ΔAccuracy
01	Oxidation of Csp3	0.87	0.75	0.93	+0.03	+0.03
02	Oxidation of Csp2 and Csp	0.90	0.79	0.93	+0.16	+0.09
03	CHOH $\longleftrightarrow$ C=O $\longrightarrow$ COOH	0.83	0.65	0.89	-0.15	-0.04
05	Redox reactions of R3N	0.71	0.41	0.78	-0.48	-0.19
06	Redox reactions of >NH, >NOH, and –N=O	0.69	0.38	0.75	-0.48	-0.25
07	Redox of quinones or analogues	0.80	0.60	0.91	+0.12	+0.06

08	Redox of S atoms	0.83	0.59	0.87	-0.30	-0.09
Main class 1	**Redox reactions**	**0.80**	**0.60**	**0.86**	**-0.16**	**-0.06**
11	Hydrolysis of esters and lactones	0.74	0.47	0.83	-0.45	-0.22
12	Hydrolysis of amides, lactams and peptides	0.78	0.55	0.83	-0.28	-0.14
14	Other hydrolyses	0.76	0.52	0.85	+0.06	+0.03
Main class 2	**Hydrolyses**	**0.76**	**0.51**	**0.84**	**-0.22**	**-0.11**
21	O-glucuronidations and glycosylations	0.58	0.15	0.61	-0.77	-0.38
22	N- and S-glucuronidations and glycosilations	0.54	0.07	0.59	-0.81	-0.40
23	Sulfonations	0.71	0.42	0.75	-0.47	-0.23
24	GSH and RSH conjugations	0.78	0.56	0.89	-0.15	-0.08
25	Acetylations and acylations	0.73	0.46	0.79	-0.42	-0.22
26	CoASH-ligation plus amino acid conjugations	0.88	0.77	0.96	-0.13	-0.07
27	Methylations	0.76	0.52	0.81	-0.26	-0.13
Main class 3	**Conjugations**	**0.71**	**0.42**	**0.77**	**-0.43**	**-0.22**
Overall means		**0.76**	**0.51**	**0.82**	**-0.27**	**-0.13**

Table 3.11: Performance metrics (Accuracy, MCC and AUC values) for the 17 considered classes of metabolic reactions as derived by the second round of predictions in which the NR atoms are purposely selected to have the same atom type(s) of the reactive centers. The differences compared to the first round in accuracy and MCC values are also reported.

Class ID	Description	Selected Features
01	Oxidation of Csp3	Atm_piS(r), Mol_Dipole, Mol_Lipole, Mol_Mass, Mol_Ovality, Mol_PSA, Mol_Volume
02	Oxidation of Csp2 and Csp	Atm_Charge, Atm_piS(r), Mol_Impropers, Mol_PSA, Mol_LogP
03	CHOH $\longleftrightarrow$ C=O $\longrightarrow$ COOH	Atm_Charge, Atm_q(r)-Z(r), Mol_Lipole, Mol_Mass, Mol_PSA
05	Redox reactions of R3N	Atm_Charge, Atm_piS(r), Mol_Dipole, Mol_Surface
06	Redox reactions of >NH, >NOH, and –N=O	Atm_Charge, Atm_Homo-1, Mol_HbAcc, Mol_Impropers, Mol_PSA
07	Redox of quinones or analogues	Atm_Homo-1, Atm_Lumo, Mol_Atoms, Mol_Dipole, Mol_HbDon, Mol_Sav
08	Redox of S atoms	Atm_Dn(r), Atm_q(r)-Z(r), Atm_Homo, Mol_HbDon, Mol_PSA
11	Hydrolysis of esters and lactones	Atm_Charge, Atm_piS(r), Atm_Homo-1, Mol_Impropers, Mol_Mass, Mol_PSA
12	Hydrolysis of amides, lactams and peptides	Atm_piS(r), Mol_Gyrrad, Mol_Impropers, Mol_PSA
14	Other hydrolyses	Mol_Gyrrad, Mol_Impropers, Mol_Lipole, Mol_PSA, Mol_LogP
21	O-glucuronidations and glycosylations	Atm_De(r), Mol_HbDon
22	N- and S-glucuronidations and glycosilations	Mol_Atoms, Mol_Dipole, Mol_LogP
23	Sulfonations	Atm_Dn(r), Atm_Homo, Mol_Rotors, Mol_Gyrrad, Mol_HbDon, Mol_PSA
24	GSH and RSH conjugations	Atm_Dn(r), Atm_Lumo, Mol_Atoms, Mol_Gyrrad
25	Acetylations and acylations	Atm_Charge, Atm_Dn(r), Atm_Lumo, Mol_PSA

26	CoASH-ligation plus amino acid conjugations	Atm_Charge, Atm_piS(r), Atm_Lumo+1
27	Methylations	Mol_Atoms, Mol_Rotors, Mol_Gyrrad, Mol_HbAcc

Table 3.12: Descriptors selected by the feature importance analyses for the models developed for the 17 considered classes of metabolic reactions in the second round of predictions.

			Accuracy		MCC		AUC	
Subclass ID	Cases	Description	Mean	Range	Mean	Range	Mean	Range
01.01	193	Oxidations of isolated Csp3	0.84	0.08	0.70	0.03	0.90	0.06
01.02	233	Oxidations of C in α to an unsaturated system	0.92	0.05	0.84	0.10	0.96	0.04
01.03	492	Oxidations of isolated Csp3 carrying an heteroatom	0.90	0.04	0.81	0.09	0.94	0.04
01.04	81	Dehydrogenations	0.79	0.14	0.58	0.27	0.83	0.14
02.01	440	Oxidations of aryl compounds	0.82	0.04	0.63	0.08	0.82	0.07
02.02	94	Oxidations of azarenes	0.88	0.13	0.77	0.26	0.93	0.10
02.03	55	Oxidations of >C=C<	0.90	0.13	0.79	0.26	0.94	0.09
03.02	71	Hydrogenations of cabonyls	0.95	0.09	0.89	0.18	0.97	0.05
05.01	65	Oxidations of tertiary alkylamines	0.95	0.08	0.91	0.16	0.99	0.04
06.01	62	Hydroxylations of amines	0.98	0.05	0.96	0.11	0.99	0.01
07.04	51	Oxidations of phenols	0.73	0.25	0.45	0.50	0.79	0.26
08.03	83	Oxygenations of sulfides	0.99	0.05	0.97	0.08	0.99	0.02

Main class 1		**Average**	**0.89**	**0.09**	**0.78**	**0.18**	**0.92**	**0.08**
11.01	103	Hydrolysis of alkyl esters	0.98	0.06	0.95	0.11	0.99	0.05
11.03	100	Hydrolysis of anionic and cationic esters	0.97	0.07	0.93	0.13	0.98	0.06
11.08	51	Hydrolysis of esters of inorganic acids	0.91	0.18	0.81	0.35	0.97	0.08
12.02	55	Hydrolysis of anilides and hydrazidas	0.93	0.13	0.85	0.26	0.97	0.08
Main class 2		**Average**	**0.94**	**0.11**	**0.88**	**0.21**	**0.98**	**0.07**
21.01	85	O-glucuronidation of alcohols	0.97	0.05	0.94	0.11	0.97	0.07
21.02	152	O-glucuronidation of phenols	0.94	0.05	0.92	0.10	0.98	0.04
21.03	99	O-glucuronidation of caboxylic acids	0.96	0.08	0.92	0.16	0.99	0.03
22.01	97	N-glucuronidation of linear and cyclic amines	0.95	0.09	0.90	0.19	0.97	0.07
23.01	70	O-sulfonation of phenols	0.94	0.17	0.88	0.34	0.97	0.08
24.01	89	Nucleophilic additions of glutathione	0.88	0.11	0.76	0.23	0.94	0.07
24.02	68	Reactions of glutathione addition-elimination	0.81	0.13	0.61	0.29	0.77	0.19
Main class 3		**Average**	**0.92**	**0.10**	**0.85**	**0.20**	**0.93**	**0.08**

Table 3.13: Performance metrics (Accuracy, MCC and AUC values) for the considered subclasses of metabolic reactions as derived by the first round of predictions. For each metric, the mean and range values are derived by repeating 100 times the random undersampling of the NR atoms.

Subclass ID	Description	Accuracy
01.01	Oxidations of isolated Csp3	Atm_Broto, Atm_Kier, Atm_Charge, Atm_Dn(r), Atm_piS(r), Atm_Homo, Atm_Lumo+1, Mol_Atoms, Mol_Impropers
01.02	Oxidations of C in α to an unsaturated system	Atm_Broto, Atm_Charge, Atm_piS(r), Mol_Rotors, Mol_HbDon
01.03	Oxidations of isolated Csp3 carrying an heteroatom	Atm_Kier, Atm_Charge, Atm_piS(r), Atm_Homo-1, Mol_Atoms, Mol_HbAcc, Mol_HeavyAtoms, Mol_Mass
01.04	Dehydrogenations	Atm_Broto, Atm_Charge, Atm_q(r)-Z(r), Atm_piS(r), Atm_Lumo+1, Mol_HbDon, Mol_Mass, Mol_Volume
02.01	Oxidations of aryl compounds	Atm_Broto, Mol_HbDon
02.02	Oxidations of azarenes	Atm_Broto, Atm_Kier, Atm_Charge, Atm_Dn(r), Atm_De(r), Atm_piS(r), Atm_Homo-1, Mol_Dipole, Mol_Sav
02.03	Oxidations of >C=C<	Atm_Broto, Atm_Kier, Atm_Charge, Atm_De(r), Mol_Lipole
03.02	Hydrogenations of cabonyls	Atm_Broto, Atm_Kier, Atm_Charge, Atm_De(r), Mol_Lipole
05.01	Oxidations of tertiary alkylamines	Atm_Kier, Atm_Homo-1, Mol_Dipole, Mol_Rotors, Mol_HbAcc, Mol_PSA
06.01	Hydroxylations of amines	Atm_Broto, Atm_Kier, Atm_Charge

07.04	Oxidations of phenols	Atm_Broto, Atm_Kier, Atm_Charge, Mol_HeavyAtoms, Mol_Volume
08.03	Oxygenations of sulfides	Atm_Broto, Atm_Kier, Atm_Charge, Atm_piS(r), Atm_Homo, Mol_Atoms, Mol_HbAcc
11.01	Hydrolysis of alkyl esters	Atm_Kier, Atm_Charge, Mol_HbAcc
11.03	Hydrolysis of anionic and cationic esters	Atm_Broto, Atm_Charge, Mol_Dipole
11.08	Hydrolysis of esters of inorganic acids	Atm_Kier, Atm_Lumo, Atm_Lumo+1, Mol_Atoms, Mol_Lipole
12.02	Hydrolysis of anilides and hydrazidas	Atm_Broto, Atm_Charge, Atm_q(r)-Z(r)
21.01	O-glucuronidation of alcohols	Atm_Broto, Atm_Kier
21.02	O-glucuronidation of phenols	Atm_Broto, Atm_Kier, Mol_HbAcc, Mol_HbDon
21.03	O-glucuronidation of caboxylic acids	Atm_Broto, Atm_Kier, Atm_Homo
22.01	N-glucuronidation of linear and cyclic amines	Atm_Kier, Atm_Charge, Atm_piS(r), Atm_Homo, Atm_Lumo+1, Mol_Dipole, Mol_Rotors, Mol_HeavyAtoms, Mol_Impropers
23.01	O-sulfonation of phenols	Atm_Broto, Atm_Kier, Atm_q(r)-Z(r), Mol_Gyrrad
24.01	Nucleophilic additions of glutathione	Atm_Broto, Atm_Kier, Atm_Charge, Atm_piS(r), Atm_Lumo+1, Mol_Impropers
24.02	Reactions of glutathione addition-elimination	Atm_Broto

Table 3.14: Descriptors selected by the feature importance analyses for the models developed for the considered subclasses of metabolic reactions in the first round of predictions.

Subclass ID	Description	Accuracy	MCC	AUC	ΔAccuracy	ΔMCC
01.01	Oxidations of isolated Csp3	0.87	0.73	0.93	+0.03	+0.02
01.02	Oxidations of C in α to an unsaturated system	0.85	0.68	0.92	-0.07	-0.16
01.03	Oxidations of isolated Csp3 carrying an heteroatom	0.86	0.71	0.90	-0.04	-0.10
01.04	Dehydrogenations	0.82	0.65	0.88	+0.03	+0.07
02.01	Oxidations of aryl compounds	0.88	0.75	0.92	+0.06	+0.12
02.02	Oxidations of azarenes	0.91	0.82	0.97	+0.03	+0.07
02.03	Oxidations of >C=C<	0.75	0.50	0.84	-0.15	-0.29
03.02	Hydrogenations of cabonyls	0.75	0.50	0.82	-0.20	-0.39
05.01	Oxidations of tertiary alkylamines	0.72	0.43	0.82	-0.15	-0.49
06.01	Hydroxylations of amines	0.93	0.87	0.98	-0.05	-0.09
07.04	Oxidations of phenols	0.82	0.66	0.85	+0.09	+0.19
08.03	Oxygenations of sulfides	0.84	0.69	0.84	-0.15	-0.27
Main class1	**Average**	**0.83**	**0.66**	**0.89**	**-0.06**	**-0.11**
11.01	Hydrolysis of alkyl esters	0.73	0.46	0.84	-0.25	-0.49
11.03	Hydrolysis of anionic and cationic esters	0.76	0.52	0.89	-0.21	-0.41
11.08	Hydrolysis of esters of inorganic acids	0.80	0.60	0.83	-0.11	-0.21
12.02	Hydrolysis of anilides and hydrazidas	0.72	0.44	0.79	-0.21	-0.41

Main class 2	**Average**	**0.75**	**0.51**	**0.84**	**-0.19**	**-0.38**
21.01	O-glucuronidation of alcohols	0.58	0.16	0.63	-0.39	-0.77
21.02	O-glucuronidation of phenols	0.71	0.43	0.77	-0.23	-0.49
21.03	O-glucuronidation of caboxylic acids	0.89	0.77	0.94	-0.07	-0.15
22.01	N-glucuronidation of linear and cyclic amines	0.58	0.17	0.65	-0.37	-0.73
23.01	O-sulfonation of phenols	0.72	0.44	0.80	-0.22	-0.44
24.01	Nucleophilic additions of glutathione	0.92	0.84	0.97	+0.04	+0.08
24.02	Reactions of glutathione addition-elimination	0.78	0.55	0.84	-0.03	-0.06
Main class 3	**Average**	**0.74**	**0.48**	**0.80**	**-0.18**	**-0.36**

Table 3.15: Performance metrics (Accuracy, MCC and AUC values) for the considered subclasses of metabolic reactions as derived by the second round of predictions in which the NR atoms are purposely selected to have the same atom type(s) of the reactive centers. The differences with the first round in accuracy and MCC values are also reported.

Subclass ID	Description	Accuracy
01.01	Oxidations of isolated Csp3	Atm_piS(r), Mol_Dipole, Mol_Mass, Mol_PSA, Mol_LogP
01.02	Oxidations of C in α to an unsaturated system	Atm_Charge, Atm_piS(r), Mol_Mass
01.03	Oxidations of isolated Csp3 carrying an heteroatom	Atm_Charge, Atm_q(r)-Z(r), Mol_PSA, Mol_LogP, Mol_Volume
01.04	Dehydrogenations	Atm_piS(r), Atm_q(r)-Z(r), Mol_Dipole, Mol_Rotors, Mol_HeavyAtoms, Mol_Impropers, Mol_Lipole, Mol_PSA

02.01	Oxidations of aryl compounds	Atm_Charge, Atm_Dn(r), Atm_Homo, Mol_Dipole, Mol_Gyrrad, Mol_Lipole, Mol_Mass, Mol_PSA, Mol_LogP, Mol_Volume
02.02	Oxidations of azarenes	Atm_Charge, Atm_q(r)-Z(r), Atm_piS(r), Mol_Gyrrad, Mol_HbAcc, Mol_LogP
02.03	Oxidations of >C=C<	Atm_Dn(r), Atm_Homo-1, Atm_Homo, Atm_Lumo, Mol_Gyrrad, Mol_LogP
03.02	Hydrogenations of cabonyls	Atm_Charge, Atm_q(r)-Z(r), Atm_Lumo+1, Mol_HbAcc, Mol_PSA
05.01	Oxidations of tertiary alkylamines	Atm_Charge, Atm_De(r), Atm_piS(r), Mol_Gyrrad
06.01	Hydroxylations of amines	Atm_De(r), Atm_piS(r), Mol_PSA
07.04	Oxidations of phenols	Atm_q(r)-Z(r), Atm_piS(r), Mol_Atoms, Mol_Dipole, Mol_LogP
08.03	Oxygenations of sulfides	Atm_Charge, Atm_q(r)-Z(r), Atm_Homo, Atm_Lumo, Mol_HbDon
11.01	Hydrolysis of alkyl esters	Atm_Charge, Atm_piS(r), Atm_Homo-1, Mol_Impropers, Mol_Sas
11.03	Hydrolysis of anionic and cationic esters	Atm_piS(r), Mol_Impropers
11.08	Hydrolysis of esters of inorganic acids	Atm_Charge, Atm_Dn(r), Atm_q(r)-Z(r), Atm_Lumo+1
12.02	Hydrolysis of anilides and hydrazidas	Atm_Charge, Atm_piS(r), Mol_HeavyAtoms, Mol_Impropers
21.01	O-glucuronidation of alcohols	Atm_Charge, Atm_Dn(r), Mol_HbDon, Mol_Impropers, Mol_Sav

21.02	O-glucuronidation of phenols	Atm_Charge, Atm_Dn(r), Atm_De(r), Atm_q(r)-Z(r), Atm_Homo, Mol_Gyrrad, Mol_HbDon, Mol_Volume, Mol_HeavyAtoms, Mol_PSA, Mol_LogP
21.03	O-glucuronidation of caboxylic acids	Atm_Charge, Atm_piS(r), Atm_Homo-1
22.01	N-glucuronidation of linear and cyclic amines	Atm_Charge, Atm_piS(r), Mol_Gyrrad, Mol_Impropers, Mol_PSA
23.01	O-sulfonation of phenols	Atm_Charge, Atm_De(r), Atm_Homo, Mol_Gyrrad, Mol_HbDon, Mol_Lipole, Mol_PSA, Mol_Sas
24.01	Nucleophilic additions of glutathione	Atm_Charge, Atm_Dn(r), Atm_piS(r), Atm_Lumo, Mol_Rotors, Mol_Impropers, Mol_Ovality
24.02	Reactions of glutathione addition-elimination	Atm_Lumo, Atm_Lumo+1, Mol_Atoms, Mol_Dipole, Mol_Gyrrad, Mol_LogP

Table 3.16: Descriptors selected by the feature importance analyses for the models developed for the considered subclasses of metabolic reactions in the second round of predictions.

Chapter 4

PCM Approach to ABCB Transporter Class

4.1 ABC Transporters

Adenosine triphosphate (ATP) binding cassette (ABC) transporters are a large family of proteins that catalyse the translocation of various substrates across cellular membranes, utilising ATP hydrolysis for energy. Transported molecules include ions, sugars, polysaccharides, vitamins, amino acids, peptides, lipids, hormones and xenobiotics. The diversity in substrate specificity reflects the diversity of physiological roles of ABC transporters, since their involvement in lipid and osmotic homeostasis, antigen processing, cell division, drug resistance, detoxification and many others.[101][102]

Their fundamental roles explain their ubiquitous distribution as Figure 4.1 depicts, and it has been highlighted that ABC transporters have been conserved across *archaea*, *eubacteria* and *eukarya* kingdoms.[103][101]

This family represents one of the four major human gene families.[105] Humans have 48 ABC genes, with 44 encoding membrane transporters organised into five families (A, B, C, D, and G).[106][101]

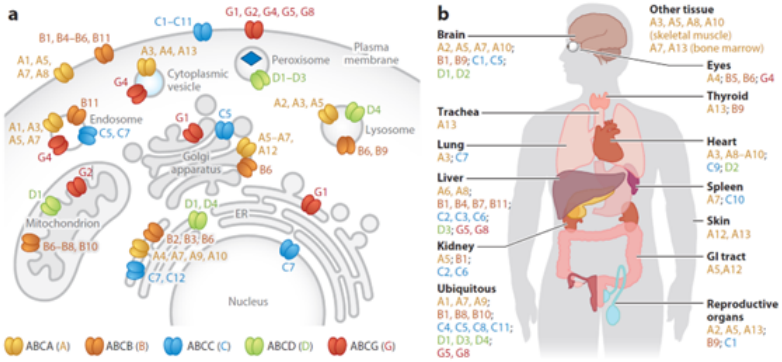

Figure 4.1: Cellular (a) and organ/tissue (b) distribution of human ABC transporters. Abbreviations: ER endoplasmic reticulum, GI gastrointestinal. Reproduced from [104]

4.1.1 ABC Transporters Structure

Structural biology studies have been crucial in deepening the knowledge of ABC transporter structure and in understanding their functions. These transporters are composed of two core domains (Figure 4.2): the nucleotide-binding domain (NBD) and the transmembrane domain (TMD). The NBD is responsible for ATP binding and hydrolysis, while the TMD facilitates substrate translocation – export or import. Human ABC transporters are classified into "half" or "full" transporters, depending on how the domains are organised. Half-transporters require dimerisation to function, whereas full transporters contain both functional halves within a single polypeptide.

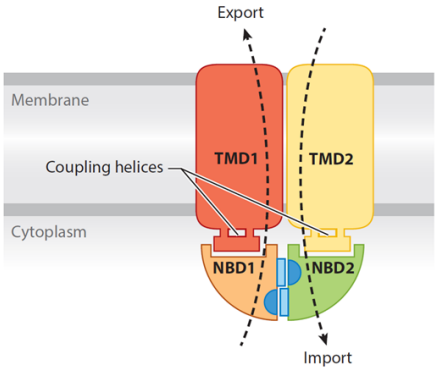

Figure 4.2: General architecture of ABC transporter domain core, arrows indicate the transport direction. Reproduced from [104]

NBD domain is highly conserved across all ABC transporters, in both sequence and structure. Structurally it consists of a RecA-like fold made of an α-helical domain and a β-sheet domain. The key motifs that are conserved across all families include: Walker A and B motifs – involved in ATP binding (α and β phosphates) and coordination of the Mg^{2+} ion; Q-loop and H-loop – involved in hydrolysis of

ATP; signature motif (LSGGQ) – involved in the orienting of ATP.[107]

The TMD is less conserved compared to the NBD, and this is due to the diverse substrate specificities seen among different ABC transporters. The TMD consists of many transmembrane helices that form a channel through which substrates are transported. In ABC exporters, the TMD undergoes conformational changes, to allow substrates to move across the membrane.[108]

Their combination of conserved and variable features enables ABC transporters to carry out their diverse physiological roles effectively.

4.1.2 Overview of the Diverse Subfamilies

Among the hABC transporters involved in membrane transport, the ABCA family is crucial in lipid transport and cholesterol homeostasis, with ABCA1 playing a role in the formation of high-density lipoproteins (HDL). Mutations in ABCA1 are linked to Tangier disease, a disorder characterised by low HDL levels and cardiovascular complications.[109][110][111]

The ABCB subfamily represents the most widely studied one. In detail, the well-known ABCB1 (P-glycoprotein) is involved in drug resistance by expelling drugs from cells, particularly in cancer treatment, where they contribute to multidrug resistance (MDR). More details are reported in Section 4.1.2.1.

The ABCC subfamily contains 12 genes, most of which encode MDR-associated proteins (MRPs), primarily involved in drug resistance. This subfamily also includes three unique transporters linked to ion transport.

Additionally, ABCC8 (SUR1) and ABCC9 (SUR2) regulate potassium ion channels in β-cells and cardiac muscle [112], with mutations linked to hypoglycemia, hyperinsulinism [113] and cardiomyopathies [114]. ABCC7 (CFTR) acts as a chloride ion channel essential for fluid homeostasis. Mutations in CFTR lead to cystic fibrosis, a serious genetic disorder causing respiratory failure.[115]

Additionally, ABCC8 (SUR1) and ABCC9 (SUR2) regulate potassium ion channels in β-cells and cardiac muscle [112], with mutations linked to hypoglycemia, hyperinsulinism [113] and cardiomyopathies [114].

The ABCD subfamily features four half-transporters, mainly involved in the transport of long-chain fatty acids into peroxisomes. ABCD1, for instance, transports very long-chain fatty acids and it is associated with X-linked adrenoleukodystrophy (X-ALD), a severe neurological disorder [116]. ABCD4 is involved in vitamin B12 trafficking from lysosomes, and mutations in this gene cause vitamin B12 deficiency.[117]

The ABCG subfamily consists of five half-transporters, which form homodimers or heterodimers. This subfamily members are involved in sterol and lipid transport, and drug resistance. ABCG1 and ABCG4 transport cholesterol [118][111], while ABCG2 is a significant multidrug exporter that plays a crucial role in the blood-brain barrier, contributing to multidrug resistance in cancer [119]. ABCG5 and ABCG8 form a heterodimer that is critical for cholesterol transport in hepatocytes.[120]

4.1.2.1 ABCB Subfamily Transporters

The ABCB subfamily (also known as MDR/TAP family) includes 10 proteins with a various substrate specificity.

This subfamily plays critical roles in various physiological processes, such as multidrug resistance (MDR) and antigen processing. ABCB transporters are found in various organisms, from bacteria to humans, and are involved in the transport of lipids, bile acids, peptides, and drugs across membranes. Their activity is catalysed by ATP hydrolysis, which provides the energy needed for the translocation process.

The most widely studied hABC transporter is ABCB1, also known as P(permeability)-glycoprotein (or MDR protein).[121] ABCB1 acts as a polyspecific multidrug exporter that is involved in cellular detoxification, a process exploited by cancer cells to develop chemotherapy resistance. This transporter is located in the main blood-organ barriers, but also in the liver, kidneys, intestines, and genetic variants of ABCB1 have been associated to diseases such as inflammatory bowel disease [122]. However, targeting ABCB1 in cancer treatments has been a hard challenge due to a systemic toxicity and off-target effects. The crucial role of ABCB1 in drug interactions and pharmacokinetics has led to its inclusion on the FDA's list of essential transporters for screening in drug development.[123]

ABCB2 (TAP1) and ABCB3 (TAP2) form a heterodimer that plays a critical role

in antigen processing by transporting peptides from the cytosol to the endoplasmic reticulum (ER) lumen.[104]

ABCB4 (also called MDR3 since its sequence similarity with ABCB1) is expressed in liver cells, and it protects them from bile toxicity, translocating phosphatidylcholine into bile canaliculi.[124] Mutations in this gene can cause progressive familial intrahepatic cholestasis.[125]

The plasma membrane protein ABCB5 is involved in drug resistance, and it can regulate corneal epithelium development.[126]

The mitochondrial transporter ABCB6 is associated with heme biosynthesis, with mutations linked to skin disorders (dyschromatosis)[127] and ocular abnormalities.[128]

ABCB7 is important for iron-sulfur cluster maturation in cells, and mutations can cause X-linked sideroblastic anemia.[129]

ABCB8 plays a role in mitochondrial iron homeostasis and is crucial for heart function, indeed the deletion of *Abcb8* in mice led to spontaneous cardiomyopathies.[130]

ABCB9 (also known as TAPL – TAP-like) transports peptides into lysosomes.[131]

Although the precise biological function of ABCB10 is not completely known, its overexpression is linked to increased hemoglobin production and it may have a role in type 2 diabetes.[132][133]

Also known as the bile salt export pump, ABCB11 has a critical role for bile salt transport in the liver, and its mutations cause PFIC2, a severe liver disorder in children.[134]

While these proteins play essential roles in normal physiological processes, their overexpression or dysregulation is associated with some diseases or complex clinical status, particularly in cancer and multidrug resistance. Since hP-gp mainly acts as a pump, expelling anticancer drugs from tumoral cells, the inhibition of it or other ABCB transporters could increase the intracellular concentration of drugs and hopefully enhance the therapeutic effect. Others clinical applications of ABCB inhibitors include the attempt to enhance Alzheimer drugs effects, inhibiting P-gp block of them across the blood-brain barrier.

ABCB transporters cover wide applications in therapeutic area and the need for further research on interactions between transporters and pathological processes is becoming more and more relevant.

4.2 Aim

Since the understanding of drug-transporter interactions can mitigate adverse drug reactions, improving drug safety, ML predictive models – capable of identifying the propensity of a compound to interact with some transporters – are vital in early-stage drug development.

The polyspecific nature of ABC transporters, added to the limited knowledge on the molecular basis of their multispecificity, makes traditional molecular modelling and QSAR approaches challenging. Indeed, although QSAR is the most common ligand-based (LB) approach, the main drawback is its consideration of a single target per time. For this reason, a technique able to consider simultaneously multiple targets is more suitable, as proteochemometric (PCM) modelling does. Combining ligand information as well as target information, PCM can predict the variable of interest, such as the activity of a compound, of a various datasets; furthermore, it presents the big advantage to shedding light on hypothetical binding modes too.

Recent studies highlighted that PCM has been successfully employed to some SLC (solute carriers), ABC transporters, OATP (organic anion transporting peptides) and MDR proteins [135] and some GABA transporters [136].

Given its relevance and aiming to include a structure-based approach thanks to the recent release of resolved structures, the present work proposes a PCM approach to better investigate ABCB subfamily transporters, through regression models capable of predicting the bioactivity value of compounds against the proteins.

4.3 Methods

The main program used in this work was KNIME platform since its user-friendly interface and the possibility to integrate built-in nodes with different programming

languages in the workflow, such as Python. Due to this features, regression predictive models were built.

4.3.1 Data Collection

In order to automate, a KNIME workflow developed by Mr. Fabian Kaiser, BSc [137] was modified and embedded in the PCM project pipeline. Bioactivity data for the training set were retrieved from ChEMBL version 31.

Training Set
In total, 15941 datapoints were retrieved from ChEMBL31 for the ABCB transporters. Given that the quality of the data significantly impacts the model quality and predictive capability, a rigorous curation of the data was conducted in the following steps:

Step 1: Standard relation: "="

Datapoints with a bioactivity value of type "=" were kept, while other types, such as "<" or ">" were discarded because exact bioactivity values were needed to build regression models. This resulted in 11461 datapoints.

Step 2: Standard units "nM"

Only datapoints with the unit of nanomolar were kept, resulting in 4361 points.

Step 3: Relationship type: "D" for Direct

Only datapoints with an experimental bioactivity value were kept, discarding evaluation on homologous protein. After this step, there were 3259 datapoints left.

Step 4: Standard type: "IC50"

In order to choose which kind of value for building the model (IC_{50}, EC_{50}, K_d or K_i), a heatmap was generated (Figure 4.3). Datapoints of type IC_{50} covered a wider range of proteins.

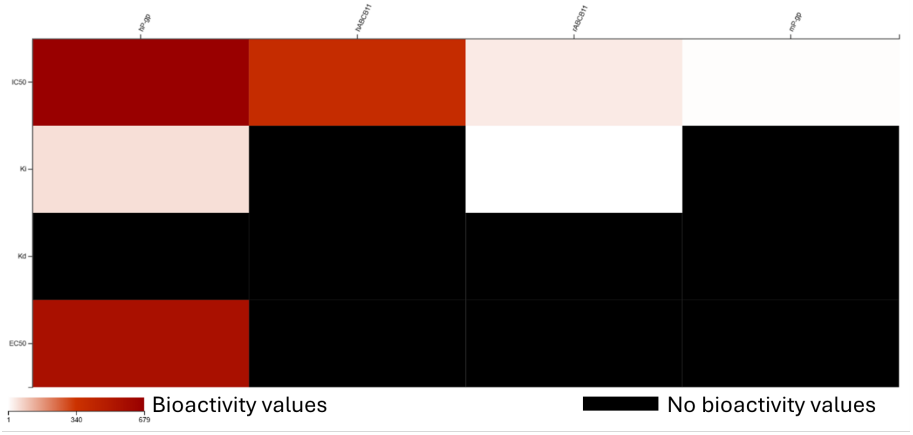

Figure 4.3: Heatmap of Activity value vs HUGO nomenclature (human protein-coding genes unique name).

Step 5: Organism: "Homo sapiens" and others The IC_{50} values need to be standardised before using them as end points to train a model. ChEMBL does that by calculating the negative logarithm of the molar concentration, *i.e.* "pIC_{50}" which is equal to -log(IC_{50}).

Thanks to the generation of the heatmap depicted in Figure 4.4, it was highlighted that the targets with enough data were: hP-gp, mP-gp, hABCB11 and rABCB11.

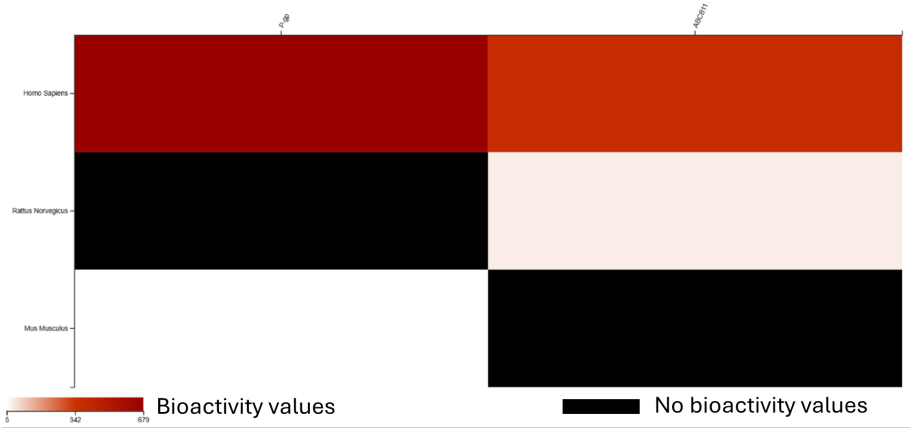

Figure 4.4: Heatmap of Organism vs HUGO nomenclature (human protein-coding genes unique name).

Step 6: Molecular standardisation

Molecular structures found in databases in their raw form cannot be used directly for modelling. A study showed that molecules in databases have an error rate in their structure representation ranging from 0.1 to 3.4% [138]. A standardisation process of molecular structures is fundamental to remove molecular structure inconsistencies

and any molecular duplicates that commonly occur in databases.

The following step were performed:

- removing missing molecules
- removing radio labelled molecules
- removing any charged compounds
- producing a standardised InChIKey for each molecule
- removing molecules with rare elements, such as Te and Se.

In this way, all molecular structures are standardised and uniformly represented by their 2D structure, so it is easier to find out duplicates.

Step 7: Grouping datapoints

All molecules with the same InChIKey were grouped into one datapoint and the mean value of all bioactivity values was calculated and considered as the final bioactivity value. So, the following group of bioactivities were grouped together in a single datapoint:

- bioactivities of identical molecules measured in different publications;
- bioactivities of identical molecules measured in different bioassays;
- bioactivities of stereoisomers;
- bioactivities of molecules and their salt forms.

At the end, the calculation of the Standard Error of the Mean (SEM) was calculated and datapoints with a SEM>0.2 were excluded.

The same curation steps were applied to ChEMBL32 and ChEMBL33 too, in order to assess their data and to choose which version to use.

The size reduction of the datasets is described by the following histograms (Figure 4.5).

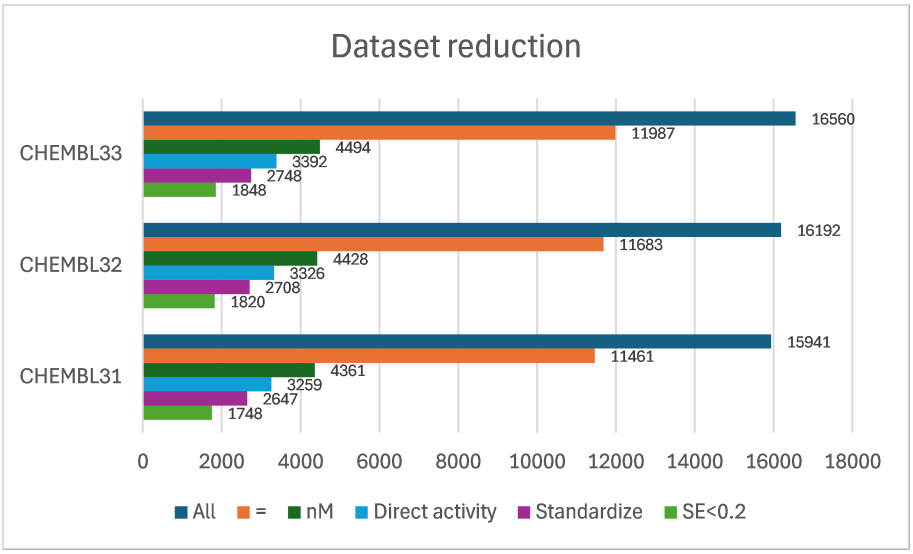

Figure 4.5: Data reduction histograms.

Test Set

To test the built models, two internal validations were obtained by cross-validation on ChEMBL31 and ChEMBL33. Then, the external test set was retrieved from ChEMBL33, using only the molecules of this version not present in ChEMBL31 (DeltaChEMBL33).

4.3.2 Target Selection

In this study, proteins are represented by the residues that constitute their binding pocket. To acquire them, sequence alignments were performed: (1) ABCB family member sequences were downloaded from UniProt [139]; (2) MOE program was used to align sequences of the ABCB family with a co-crystallised structure (PDB ID: 6QEX); (3) a radius of 4.5Å around the co-crystallised ligand was chosen; (4) all amino acids within this neighbourhood were selected.

All known isoforms – human and non-human – were considered, but only four of them – hABCB1, also known as hP-gp, mABCB1 (mP-gp), hABCB11 and rABCB11 – were studied because the only ones with enough data (inclusion in ChEMBL and known sequence).

4.3.3 Ligand Descriptors

In this work RDKit [140] has been used to calculate the descriptors of molecules, in detail constitutional and physicochemical descriptors were calculated (full list in Appendix V).

4.3.4 Target Descriptors

In this work the protein was described basing on its sequence, in detail individual amino acids that form the binding pocket. For them, Z-scales descriptors were calculated [141]. They are based on 26 physicochemical properties that have been compressed using Principal Component Analysis (PCA) down to five principal components, with the first three components describing lipophilicity, volume/polarisability and polarity, respectively [41]. Z-scales descriptors are called in the following pattern: Z1_2 standing for the first Z-scale descriptor of the second residue, and so on.

4.3.5 Proteochemometric Modelling

Four matrices were built considering, for each compound, the binary classification for each isoform (label 0: not active on the isoform; label 1: active on the isoform), the pChEMBL value, the RDKit molecular descriptors and, in addition, the following features were taken into consideration:

- Matrix1: no more descriptors
- Matrix2: mean value of amino acids descriptors (Z-scale-based physicochemical properties) for each helix
- Matrix3: amino acids descriptors of residues contained in almost 45% of binding pockets (considering all co-crystalised structures of all resolved isoforms)
- Matrix4: all amino acids descriptors.

In this work, all models were built with XG Boost algorithm, doing a regression analysis, using tailored KNIME workflows. Some PCM models evaluation and validation parameters were calculated, in detail the predictive power of the models has been evaluated in term of R^2 (ranging from $-\infty$ to 1, where the higher the R^2

metric, the higher predictivity), considering the training phase, and the validations have been evaluated in terms of Q^2, considering the testing phase. Other assessment parameters were considered, such as Mean Squared Error (MSE) and Mean Absolute Error (MAE).[142] Those two parameters are residual methods to understand the error in predictions. While MAE indicates the absolute error without considering any outliers, MSE gives an increasing exponential penalty to increasing prediction errors, indeed the bigger the prediction error, the bigger the penalty.

4.4 Results and Discussion

4.4.1 Chemical Space Analysis

The chemical space was analysed through a Principal Component Analysis (PCA) that compressed data into a low number of dimensions.

For all considered datasets, 119 descriptors from RDKit were calculated and then, compressed in two components, as Figure 4.6 shows for ChEMBL33 dataset.

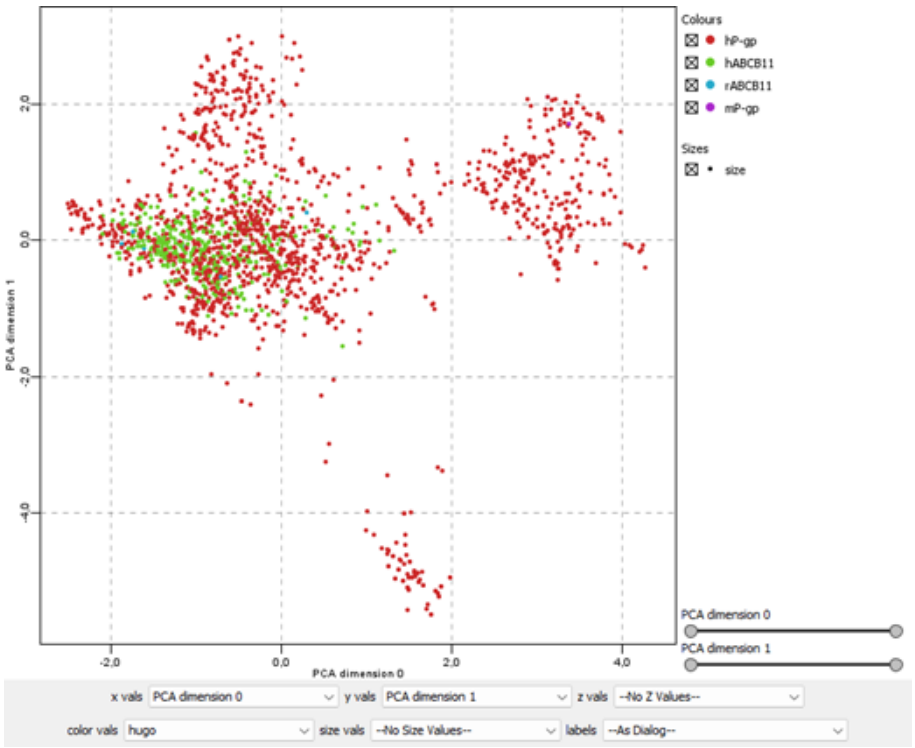

Figure 4.6: PCA plot of compounds chemical space of ChEMBL33 dataset (matrix1).

From this analysis, the huge overlap between red and green dots was evident, indicating that hP-gp molecules and hABCB11 ones are almost equally distributed, even if the hABCB11 compounds resulted in a smaller area. Purple and blue dots are less evident and very sparse, due to lower quantity of mP-gp and rABCB11 molecules, compared to the other targets.

In addition to PCA, the machine learning algorithm t-SNE (t-distributed Stochastic Neighbor Embedding) was employed for the dimensionality reduction and to visualise the considered data in two dimensions, to better investigate the hypothetical presence of hidden clusters. As Figure 4.7 shows, the central cluster of green and red dots indicates a certain chemical similarity between hP-gp and hABCB11 molecules. Some red dots are located on the borders, suggesting some molecules with different features from the majority. hABCB11 molecules, instead, appear to be less variable.

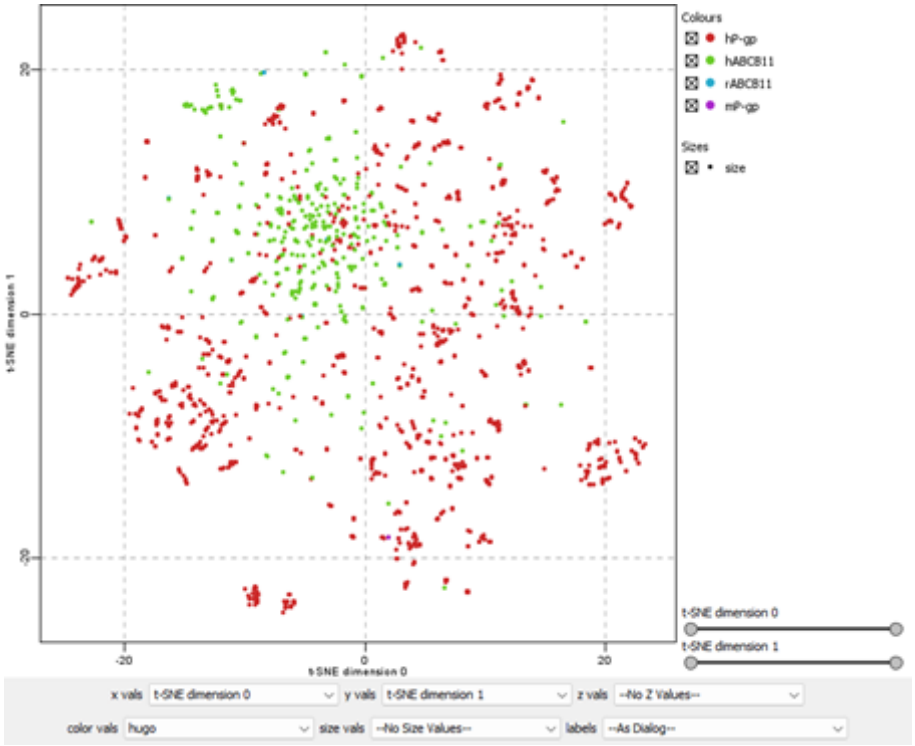

Figure 4.7: t-SNE plot of compounds chemical space of ChEMBL33 dataset (matrix1).

Then, protein information was added to PCA plot (Figure 4.8) using matrix4, in order to evaluate their impact.

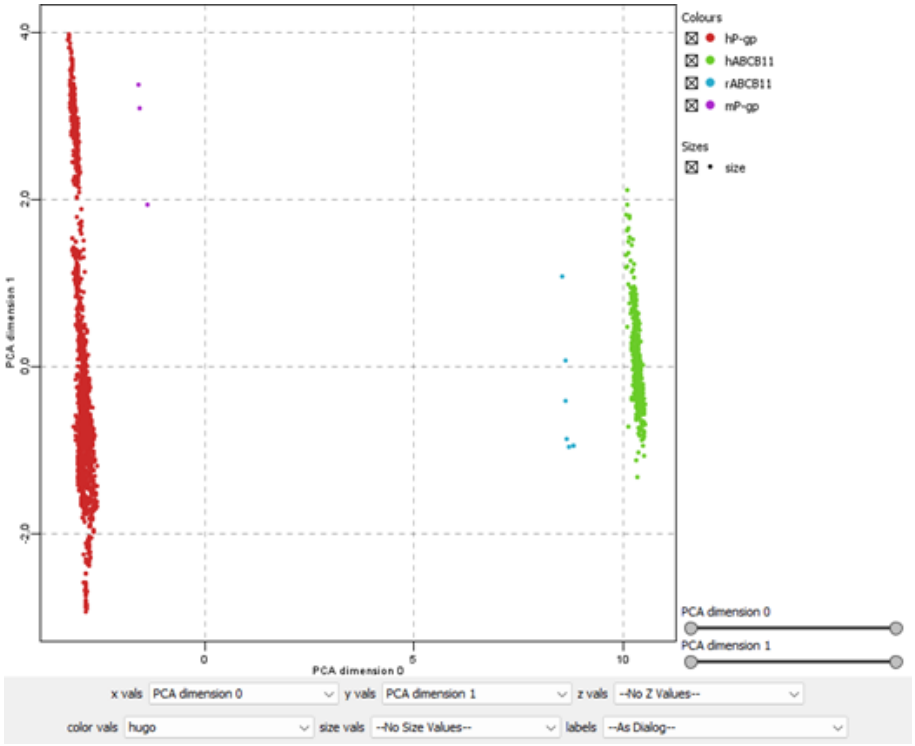

Figure 4.8: PCA plot of ligand and protein information of ChEMBL33 dataset (matrix4).

It appears evident that protein information provides additional discriminatory power since the clusters become more distinct in the PCA plot. This suggests that combining hP-gp and hABCB11 in one dataset could provide an enhanced predictive ability of the PCM model, due to enrichment into chemical space.

4.4.2 Predictive Models

The final considered datasets were composed in this way:

- ChEMBL31: 1703 compounds
- ChEMBL33: 1798 compounds
- DeltaChEMBL33: 98 compounds

4.4.2.1 10-fold Cross Validation on ChEMBL31

The first predictive model was assessed by splitting the ChEMBL31 dataset 70/30, applying a 10-fold Cross Validation as training, then a hyperparametrisation

tuning was applied in order to select the best hyperparameters for the test set. The hyperparameter optimisation was carried out with a Grid search method called "brute force" on these parameters: number of trees (n_estimators), learning rate (eta) and depth of trees (tree_depth). Table 4.1 shows the R^2 for the CV and the Q^2 of test phase. It showed a satisfactory performance in predicting the bioactivity value of considered molecules, as proved by R^2 and Q^2. All four matrices demonstrated very similar results across most metrics. The slightly higher Q^2 in matrix2 may suggest that mean protein descriptor values better describe the system, although the differences are so small that they can be not relevant.

XG Boost model - ChEMBL31 CV

	only activity and binary classification	**mean value of helices descriptors**	**BP with 45% presence in all BPs**	**BP with all residues**
	Matrix 1	**Matrix 2**	**Matrix 3**	**Matrix 4**
R^2	**0,691**	**0,689**	**0,689**	**0,689**
CV_MAE	0,418	0,425	0,424	0,424
CV_MSE	0,352	0,354	0,354	0,354
CV_RMSE	0,593	0,595	0,595	0,595
Q^2	**0,744**	**0,779**	**0,779**	**0,779**
TEST_MAE	0,409	0,390	0,390	0,390
TEST_MSE	0,322	0,277	0,277	0,277
TEST_RMSE	0,567	0,527	0,527	0,527

Table 4.1: PCM models in 10-fold CV on ChEMBL31 dataset.

In addition, an outlier analysis was carried out in order to identify and analyse datapoints with a predicted activity value that differs significantly from the experimental one. Figure 4.9 depicts the plot of matrix1, other matrices gave very similar results. The X-axis represents the experimental values of pchembl and the Y-axis represents the predicted ones. If the model is accurate, it is expected to observe cluster points around the bisector line, where the predicted value coincides with the experimental one.

Colouring various datapoints based on a parameter called delta pchembl (absolute difference between experimental pchembl value and predicted one, it is a scaled measure of deviation), it is evident as there is a large predominance of good predictions (evaluated as good when the delta pchembl is lower than 1) corresponding to 91% of all predictions, while a very limited number of datapoints reported a delta pchembl comprised between 1 and 3 (6% in range 1-1.5 and 3% comprised between 1.5 and 3) and no extreme outliers – with a delta pchembl higher than 3 – were found.

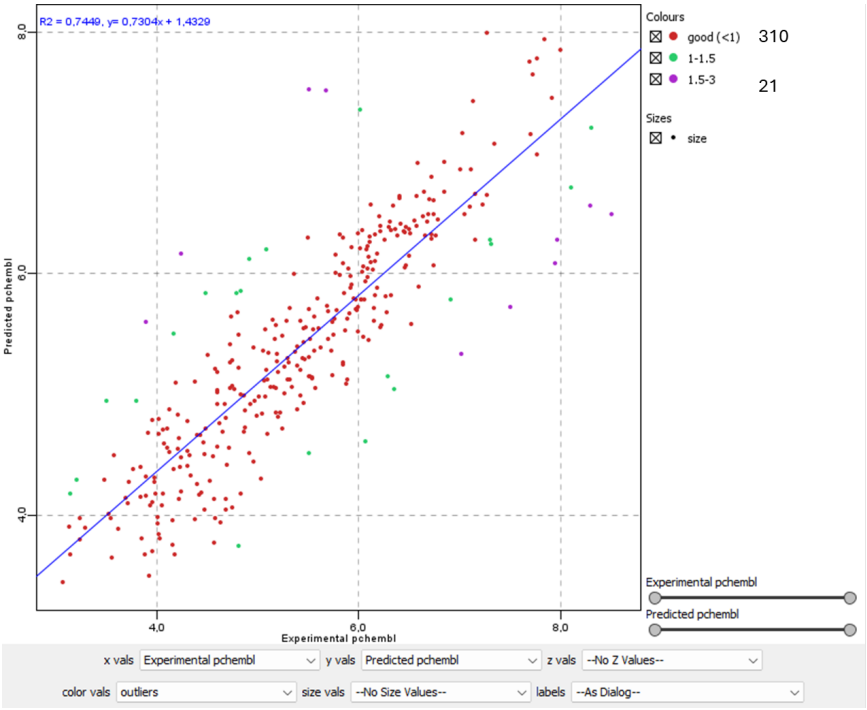

Figure 4.9: Outlier analysis – CV ChEMBL31 model (matrix1).

Matrix2, matrix3 and matrix4 resulted in slightly different distribution: 93% of datapoints reported a delta pchembl lower than 1 and 2% between 1.5 and 3. This difference could be not significant, even if it is relevant to notice that these enhanced matrices included many target descriptors, contrary to matrix1.

4.4.2.2 10-fold Cross Validation on ChEMBL33

A second CV validation was carried out on the entire ChEMBL33 dataset.

As described by the evaluation metrics depicted in Table 4.2, the training resulted slightly better when the new version of ChEMBL33 was employed, contrary to test where Q^2 showed lower performance.

The outlier analysis reported in Figure 4.10 resulted very close to the ChEMBL31 CV model; the only difference was the presence of an extreme outlier, in detail a hP-gp inhibitor reporting a delta pchembl of 3.24. The Tanimoto similarity was calculated, searching for similar compound in the related training set; the molecule more similar presented a Tanimoto value of 0.15, indicating a high similarity; this molecule was well predicted in the training, giving a delta pchembl value lower than 1. Since hP-gp molecules appeared as clusters in t-SNE analysis (Figure 4.7) and

the bad predicted molecule was in the boundaries of one of them, it is possible to conclude that this instance constitutes an outlier due to some features that the model failed to handle.

XG Boost model - ChEMBL33 CV

	only activity and binary classification	mean value of helices descriptors	BP with 45% presence in all BPs	BP with all residues
	Matrix 1	**Matrix 2**	**Matrix 3**	**Matrix 4**
R^2	**0,721**	**0,719**	**0,719**	**0,719**
CV_MAE	0,413	0,416	0,416	0,416
CV_MSE	0,323	0,325	0,325	0,325
CV_RMSE	0,568	0,570	0,570	0,570
Q^2	**0,712**	**0,704**	**0,704**	**0,704**
TEST_MAE	0,436	0,445	0,445	0,445
TEST_MSE	0,330	0,339	0,339	0,339
TEST_RMSE	0,574	0,582	0,582	0,582

Table 4.2: PCM models in 10-fold CV on ChEMBL33 dataset.

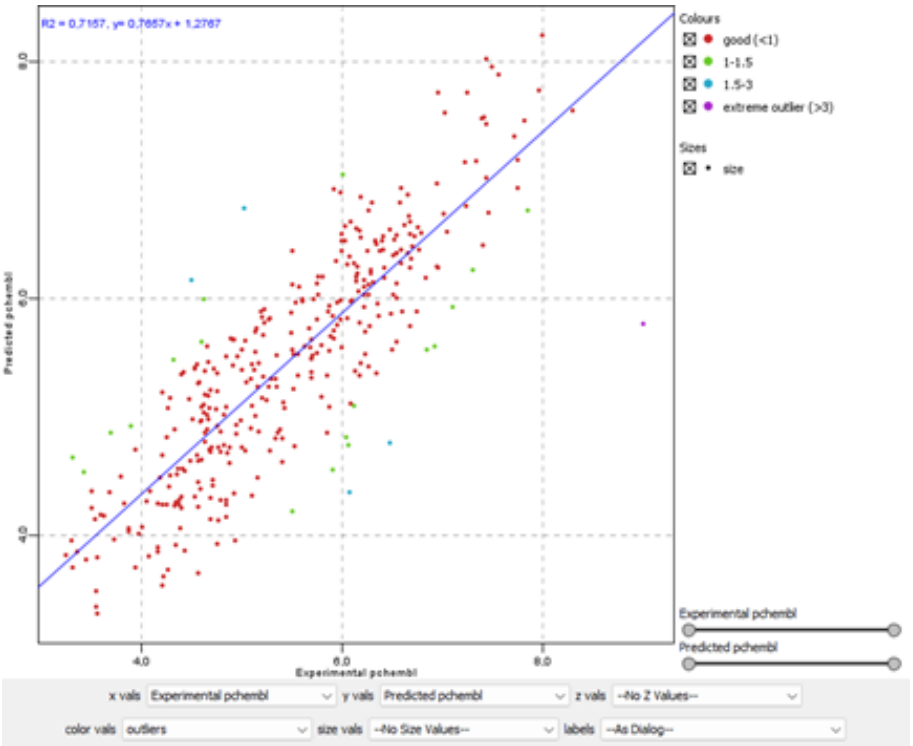

Figure 4.10: Outlier analysis – CV ChEMBL33 model (matrix1).

4.4.2.3 External Validation: DeltaChEMBL33 Dataset

In order to test the ability of a model built on the entire ChEMBL31 to predict completely new data, new molecules of version ChEMBL33 (not already in ChEMBL31) were considered as external test set – hereinafter this dataset will be called DeltaChEMBL33. As done in previous models, a 10-fold CV was used in training phase (entire ChEMBL31) and, after a hyperparametrisation tuning, the best performing setting was used to predict the DeltaChEMBL33 dataset. The results are illustrated in Table 4.3.

XG Boost model - DeltaChEMBL33

	only activity and binary classification	**mean value of helices descriptors**	**BP with 45% presence in all BPs**	**BP with all residues**
	Matrix 1	**Matrix 2**	**Matrix 3**	**Matrix 4**
R^2	**0,732**	**0,733**	**0,732**	**0,732**
CV_MAE	0,405	0,407	0,406	0,406
CV_MSE	0,312	0,311	0,311	0,311
CV_RMSE	0,559	0,558	0,558	0,558
Q^2	**0,240**	**0,237**	**0,240**	**0,240**
TEST_MAE	0,706	0,719	0,706	0,706
TEST_MSE	0,694	0,708	0,694	0,694
TEST_RMSE	0,833	0,842	0,833	0,833

Table 4.3: PCM models tested on DeltaChEMBL33.

The overall performance of the models is quite good in training, while testing is less performing than previous cases. The low Q^2 (around 0.24) suggests that the models are able to describe part of the system, but some aspects are still unknown. It is worth highlighting that there is a quantitative unbalance between the training set (1703 datapoints) and the external test set (98 datapoint, corresponding to just 6% of total).

In all described predictive models, the performance of all matrices remains almost the same. ABC transporters promiscuity and structural diversity across isoforms and organisms could be one of the main reasons why the inclusion of amino acid descriptors seems to be not relevant on the model performance.

The outlier analysis was carried out for DeltaChEMBL33 too. As Figure 4.11 depicts (taken matrix1 case), the first main difference with previous analysis is the

lower quantity of datapoints (98), and they appear to be more scattered across the plot. Notwithstanding this, the colour scale highlights that the majority of them (74%) are not considered outliers since their delta pchembl were lower than 1; the 19% of all predictions reported a delta pchembl comprised between 1 and 1.5 and 6% of them showed a value between 1.5 and 3. As in previous analysis, the behaviour of the outliers is both under-predicting and over-predicting and no extreme outliers were found.

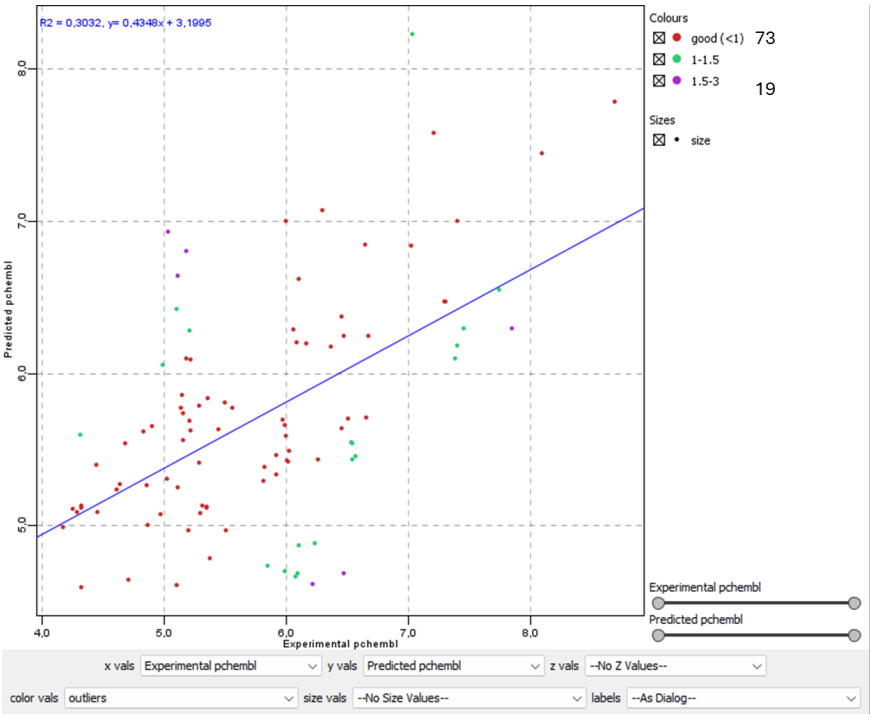

Figure 4.11: Outlier analysis – DeltaChEMBL33 model (matrix1).

The major outliers (delta pchembl from 1.5 to 1.8) were structurally checked and analysed. In Figure 4.12, the outliers are highlighted in different colours based on delta pchembl value (as shown by the legend on the right), on the x-axis the pchembl is reported, while the y-axis represents the delta pchembl. On the top of the plot, there are the major outliers, on the left those with a low pchembl value.

As highlighted by the 2D structure representation, the outliers could be divided into two clusters: molecules with no basic nitrogen atoms and those with a tetrazole moiety in their structure. These two moieties have an important role in the design of ABCB transporters novel inhibitors. They contribute to the stability and binding affinity thanks to π-stacking interactions with residues such as PHE314 or PHE732, or hydrogen bonds with acidic amino acids.[143]

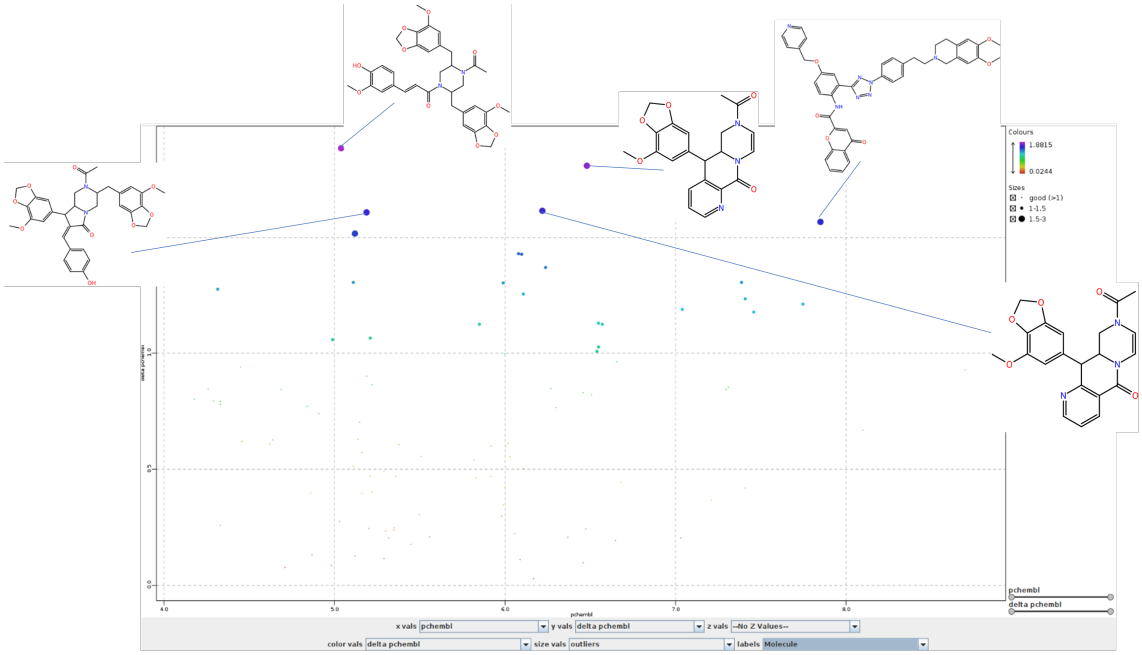

Figure 4.12: Scatter plot of outliers, coloured as legend reports. The 2D structure of major outliers is depicted.

4.4.3 Feature Importance Analysis

As last evaluation, in order to better investigate the relevance of various descriptors on model performance, a gain-based feature importance analysis was carried out on CV models as well as on DeltaChEMBL33 ones. In this method, the gain implies the average gain across all splits the feature is used in. A higher value of this metric when compared to another feature implies it is more important for generating a prediction. [80]

The most important 20 features are shown below in Figure 4.13, 4.14 and 4.15, for ChEMBL31 CV model, ChEMBL33 CV model and DeltaChEMBL33 one, respectively.

In all matrices of ChEMBL31 models, MQN1 and MQN38 (Molecular Quantum Numbers, they describe structural properties of a molecule) resulted among the most important ligand features, and the most important feature in matrix2, matrix3 and matrix4 was a component of Z-scales. In matrix2 the Z-scales descriptor was the lipophilicity of a helix, in matrix2 it was the lipophilicity of the residue 3 in considered binding pocket alignment, while in matrix4 it was the polarity of the first residue.

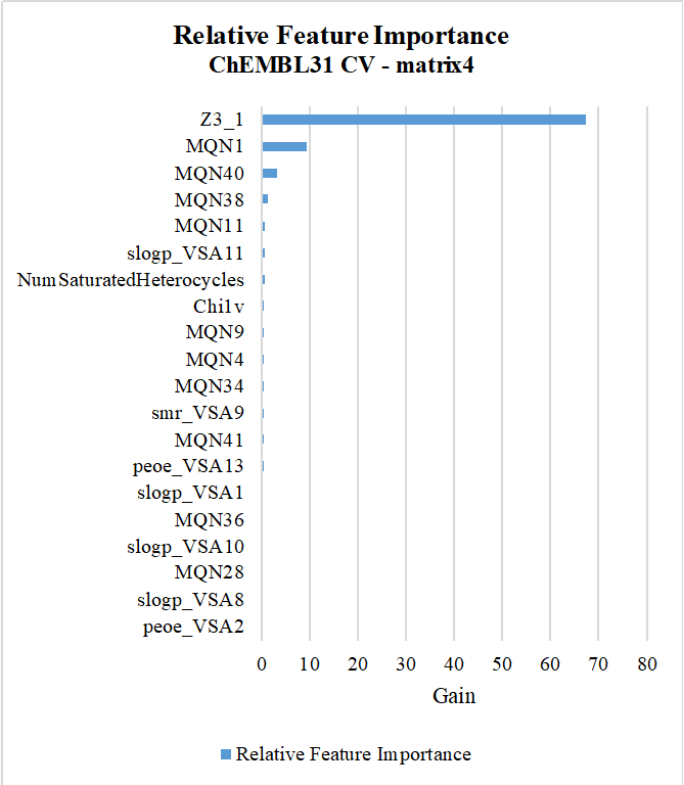

Figure 4.13: Top 20 feature importance gain-based of the model tested on ChEMBL31 CV model (matrix4).

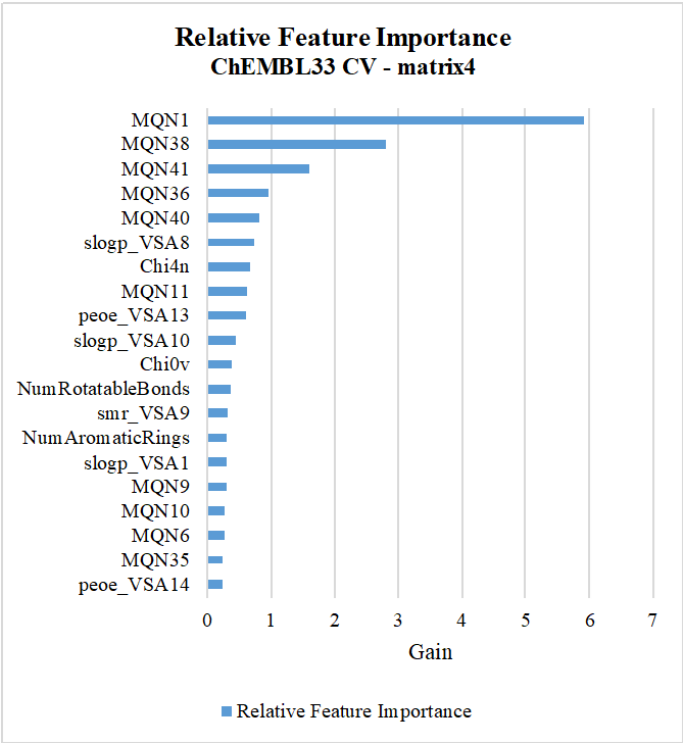

Figure 4.14: Top 20 feature importance gain-based of the model tested on ChEMBL33 CV model (matrix4).

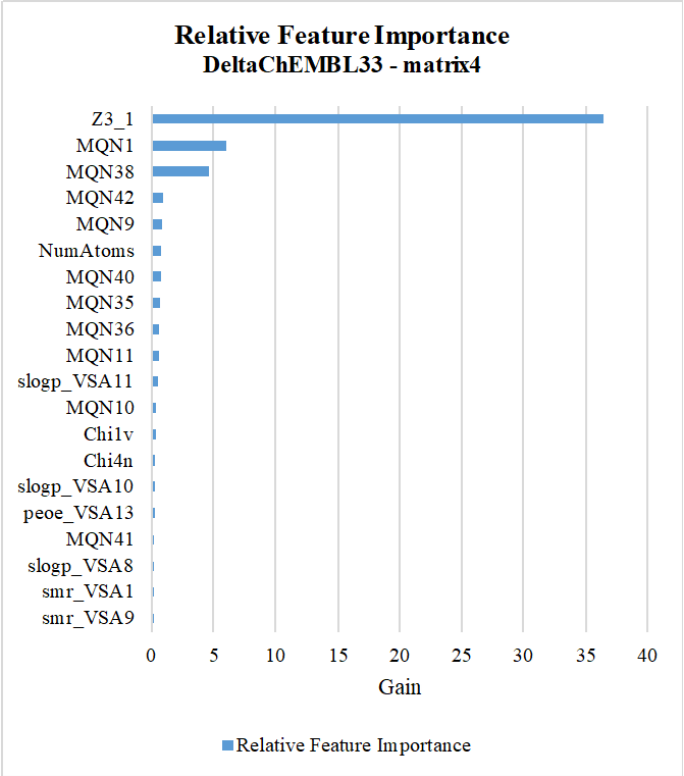

Figure 4.15: Top 20 feature importance gain-based of the model tested on DeltaChEMBL33 (matrix4).

ChEMBL33 cross-validation models gave a slightly different output of Feature Importance analysis. Indeed, no protein descriptor was defined as most important features, except for matrix2, where the first selected property was Z1-scale descriptor (the lipophilicity aspect) of a helix. Even in this case, the most important ligand descriptors were MQN type, *i.e.* MQN1, MQN38 and MQN41.

Finally, DeltaChEMBL33 models gave similar results to ChEMBL31 CV models. Indeed, the most important feature was a Z-scale descriptor in matrix2 (Z1), matrix3 (Z1_2) and matrix4 (Z3_1). Other important properties were MQN1 and MQN38; in addition, matrix1 highlighted the importance of MQN40 too.

Globally, all these findings suggest that PCM models on ABCB transporters benefit from structure activity relationship between the molecules and the proteins.

4.5 Conclusions

ABC-transporters represent an integral and important part of the human transportome. Known for their polyspecific nature, they represent a challenge for traditional molecular modelling and QSAR methods, due to their limited understanding of considered targets multispecificity.

They fulfil important roles and are strongly linked to drug absorption, distribution, and elimination. Besides cytochromes, they are also involved in drug-drug interactions and thus also drug toxicity.

However, the challenges in the field of ABC-transporter modelling are manifold, and the main question – what is the molecular basis for the polyspecificity – shows a big interest.

Furthermore, the transporters undergo a substantial conformational change when progressing through the transport cycle, which renders all structures available only snapshots of a very distinct point in the whole conformational space.

Based on these grounds, this work aimed to deep investigate the ABCB subfamily of transporters, through a PCM approach, combining information of the ligand with target ones to carry out some regression analysis, trying to provide new insights on this field.

The overall performances of generated predictive models were satisfactory, in detail, ChEMBL33 datasets gave slightly better results in training compared to ChEMBL31, that showed good outcomes too.

The analysis on DeltaChEMBL33 gave worst results and this could be due to the limited dimension of this external test set compared to ChEMBL31 training set (94%-6% instead of the common 80%-20%).

All the conducted outlier analysis proved the absence of consistent outliers in prediction, that could have had a negative impact on the model performance.

The ABC-transporters are very promiscuous proteins, and their binding site is fuzzy, so the investigation about key amino acids that drive the interactions could be

complex. As Feature Importance analyses highlighted, it is a diffuse pattern, which contribution comes from ligand features as well as target ones.

This work could represent a starting point to future investigations shedding lights on the binding mode and transport behaviour of this protein family. The increasing number of crystal structure releases contributes to improve structure-based approach, allowing soon the binding site mapping of other B subfamily members for following PCM studies. Finally, the analysis of the on- and off-kinetics and the multiple interplay of ABC-transporters depict additional challenges the scientific community will face in the coming future.

Chapter 5

Novel Structure-Based Descriptors: FingerInt and HyperVolume

Providing labelled data is essential for training a ML algorithm using a supervised approach. As detailed in Section 2.3, in the context of drug design, labels are used to describe specific characteristics of the input data and they are referred to as descriptors or features. Among the most commonly employed and crucial descriptors, there are scoring functions, molecular descriptors and fingerprints.

The present work aims to develop novel kinds of descriptors: protein-ligand interaction fingerprints and volume-based descriptors.

5.1 Interaction Fingerprints (IFP)

Interaction fingerprints (IFP) are a peculiar class of fingerprints that describes the interactions between a receptor and a small molecule, known also as protein-ligand interaction fingerprints. These advanced descriptors provide structural insights into protein targets, increasing their applicability due to the simplicity and the richness of the information they encode. IFPs are stored as one-dimensional vectors or matrices (booleans, integers, or floating-point numbers), capturing biologically relevant interactions such as hydrogen bonds, hydrophobic contacts, salt bridges,

and π-cation interactions.[144]

IFPs are particularly suitable for machine learning algorithms, providing predictive models with strong discriminative abilities due to the robustness of these descriptors.[145] Investigating protein-ligand interaction characteristics is crucial in drug discovery and can reveal key molecular biology mechanisms. IFP approaches can predict promising targets, design inhibitors, and uncover binding mechanisms for drug resistance.[146]

During last years, various IFP algorithms have been developed, each offering different methods of interaction representation. Some published ones are reported below, in chronological order of publication:

- SIFt (Structural Interaction Fingerprint)[147]: pioneers among IFP algorithms, they are composed of a one-dimensional string generated by concatenating, for each interacting binding pocket residue, a series of 7 bits. Each bit represents a specific type of interaction: the first one id referred to the presence of a contact with the ligand, the second one indicates whether the interaction involves the amino acid's backbone or side chain, the remaining bits represent the kind of interaction – *e.g.* polar, hydrophobic, hydrogen bond acceptor or donor
- APIF (Atomic Pairwise Interaction Fingerprint)[148]: this method encodes protein-ligand interactions independently of the binding site size. Each interaction pair is classified into one of six encoded interaction types and sorted according to distance ranges between interacting pairs. The resulting binary IFP is composed of 294 bits
- TIFP (Triplet Interaction Fingerprint)[149]: they are protein-ligand interaction FPs of a fixed length of 210 bits. The interaction is characterised by a triplet composed of two interacting atoms and a pseudo-atom representing the interaction (ionic bond, hydrogen bond or metal coordination)
- PyPLIF (Python-based Protein-Ligand Interaction Fingerprint)[150]: an open-source Python tool converts molecular docking interaction data into one-dimensional strings where each bit encodes the presence or absence of one of seven specific interaction types, and the similarity to a reference ligand's string is measured using the Tanimoto distance. This allows for selecting the best poses from docking results without relying on energy scores only
- SILIRID (Simple Ligand-Receptor Interaction Descriptor)[151]: SILIRID encodes binding site features using a fixed-length vector of 168 bits. It sums binary IFPs for identical amino acids, identifying ligand-target interaction types

independently of the binding site, facilitating comparisons between binding sites of different sizes

- SPLIF (Structural Protein–Ligand Interaction Fingerprints)[152]: they are independently from the binding site, but they explicitly encode the 3D structures of interacting ligand and protein fragments. For each interacting atom, circular fragments are considered, which include nearby atoms within a specific radius. These atoms are assigned to some identifiers that allow their 3D coordinates to be retrieved, allowing the comparison with a string of a reference ligand, in order to assess the similarity
- PLEC (Protein–Ligand Extended Connectivity Fingerprint)[153]: PLEC fingerprints are based on the concept of extended connectivity too. Their construction involves identifying pairs of contact atoms and characterising the surrounding atomic environment within a specific bond depth. The final fingerprint can reach a length of up to 232 bits. The primary limitation of this algorithm lies in the difficulty of interpreting the generated output.
- ProLIF (Protein–Ligand Interaction Fingerprints)[154]: Python library designed to generate interaction fingerprints for molecular complexes derived from molecular dynamics trajectories, experimental structures and docking studies. The interaction is stored within the fingerprint as a map between a pair of residues, "ligand" and "protein", along with the corresponding interaction vector. ProLIF distinguishes 14 types of interactions.

The main drawbacks of using these known approaches are, primarily, the lack of prediction interpretability that often occurs and, in parallel, the requirement for known structural information about ligand-protein interactions. Additionally, these methods often do not take into account a proper incorporation of energetic terms, which are essential for fully characterising the interactions between proteins and ligands. Finally, the plasticity of the binding site and the induced fit effects introduce further complexity.[144]

5.2 Volume-Based Descriptors

Besides IFPs, there are various other kind of features, each focusing on specific aspects of the examined compound. Among these, some descriptors analyse the available space within the binding pocket of the target protein or on the molecular overlapping rate of similar compounds.

According to the basic "lock and key" model [155], the protein and the ligand must exhibit a certain structural complementarity to allow the complex formation and proper binding. In this context, the ligand volume and shape assume a crucial role. Indeed, these properties are defined by the 3D-positioning of the ligand atoms, which must be arranged in a proper way as to determine a certain correspondence within the binding site.[156]

Unfortunately, especially for highly flexible molecules, the number of possible geometries is elevated. This means that many poses generated by molecular docking programs need to be discarded due to high-energy binding modes.[157] In fact, despite the generation of poses similar to the experimentally observed active conformation, the current scoring functions are not yet reliable or robust enough to assign the highest scores to these poses.[158]

For these reasons, some descriptors able to compare the 3D shapes of ligands and proteins have been developed. In this way, the evaluation of their molecular shape similarity is possible, and it can be used as a kind of filter to exclude structures with certain energetically unfavourable properties.

For this purpose, Muegge *et al.* [36][159] developed a correction factor for their scoring function PMF, which considers the volume occupied by the ligand relative to a reference state – *i.e.* the crystallised ligand. This factor was found to be crucial for deriving relevant interaction potentials, although its effect on predictive power appears to be lower.

ROCS (Rapid Overlay of Chemical Structures)[160] is a tool for calculating 3D shape and chemical similarity, that is measured by the overlap between dummy atoms marking relevant chemical functionalities – *i.e.* hydrogen bond donors and acceptors, charged functional groups, rings, and hydrophobic groups.

The maximum volume overlap (MVO)[161] method utilises the structural information of experimental ligand-protein complexes to select other complexes with greater geometric accuracy and high molecular similarity. The obtained pharmacophore is used to evaluate the overlap between the ligand and the reference compound coordinates, and so to select the correct structure, *i.e.* the one with an RMSD value of less than 2Å compared to the crystallographic structure. MVO has also been coupled with molecular dynamics simulations [162] and docking programs [163] in order to select reliable poses by calculating the volume overlaps between the

generated pose and that of the known ligand.

Phase Shape [164] is another shape-based method for superimposing flexible ligands. It has been used for virtual compound screening, where similar molecules are expected to have similar binding properties. Phase Shape has enriched the results showing a high accuracy in aligning compounds with their respective crystal structures and maximising volume overlap.

Martin Stahl and Hans-Joachim Böhm [158] proposed another approach to filter the docking complexes, improving the quality of predictions. For each generated poses, four properties are calculated: (a) the fraction of the ligand volume unexposed within the binding pocket, (b) the size of lipophilic cavities on protein-ligand interface, (c) the solvent-accessible surface (SAS) of the non-polar portions of the ligand, and (d) the number of contacts between non-hydrogen-bonded polar atoms of the ligand and the protein. The authors used a reference pose to build a grid model of the active site.

Although initially designed for use in virtual screening approaches, this class of descriptors showed successful applications in many other fields, such as molecular optimisation, library design, pose prediction or target active site identification.[165]

However, it is worth noticing that these kinds of descriptors are not so capable of identifying compounds that bind to proteins in different regions – *i.e.* allosteric ligands – or when great induced-fit effects are present.[164]

In conclusion, geometric complementarity alone is not sufficient to lead molecular recognition, since many other factors occur.

5.3 Aim

Understanding the underlying mechanisms of protein-ligand recognition and binding is essential in drug discovery. As highlighted in previous paragraph, among various kinds of molecular descriptors and fingerprints, the protein-ligand interaction fingerprints provide valuable insights into interaction pattern studies.[15]

Based on these grounds and with the aim to propose new tools to better describe protein-ligand interactions in docking studies, in this Chapter two novel types of

descriptors are proposed: some interaction fingerprints, called FingerInt (FI) and a volume-based set of descriptors called HyperVolume (HV) descriptors. Furthermore, an assessing benchmark was carried out to evaluate their ability to enhance the predictive power in SBVS.

5.4 Methods

5.4.1 FingerInt (FI)

FI fingerprints represent a condensed yet comprehensive description of the interactions occurring within the binding pocket between a protein and a ligand.

Typically, published IFPs provide a scheme that allows the understanding of the interaction type and the identification of the interaction pair, including the specific interacting residue and, in some cases, the atom of the ligand involved. While these details are crucial for a clear understanding of the ligand-protein complex formation mechanism, they result too detailed and difficult to interpret.

By mapping all contacts that a ligand establishes with the nearby residues, the FI algorithm gives a simplified "signature" for each protein-ligand complex encoding for the residues involved in clear contacts.

In this context, a contact is considered as established when the distance between a ligand and the protein is less than 2.5Å. The Python script for FingerInt calculation embedded in the VEGA ZZ suite [78] is based on molecular docking studies and analyses all generated poses within the considered binding pocket. As input, the script requires the docked poses database (in mdb or mol2 format file) and the receptor file used for the docking procedure (in mol2 format file).

FI fingerprints comprehend a set of 24 descriptors:

- Contacts: no. of contacts established by the ligand
- Contacts_NORM_HEVATMS: Contacts score normalised on no. of ligand heavy atoms
- Contacts_NORM_WEIGHT: Contacts score normalised on ligand molecular weight

- Contacts_FINGERPRINT: a 20 bitmap in decimal representation in which each bit corresponds to one of the twenty amino acids. A bit is set to 1 when there is at least one contact with that specific amino acid
- 20 descriptors, in "FP_XXX" name format (where XXX is one of the 20 natural amino acids in 3-letter code): how many contacts with each amino acid type the ligand establishes within 2.5Å.

FI calculation has been implemented also in a tool called Rescore+ [88], that computes some docking-derived descriptors, with the following command line:

Rescore+ -e -p "Contactsex" -d poses.db -r receptor.mol2

Where:

- -e: to set the decimal separator according to system settings
- -p: to specify which function to be calculated – here "Contactsex" identifies FI
- -d: to specify the docked poses database or a single molecule (allowed formats: Access, mol2, sdf, sqlite, zip)
- -r: minimised receptor file (allowed formats: Alchemy, AMMP, Arc, AutoDock 4 DLG, BioDock, CAR, CHARMM CRD, CIF, CML, CML 2.0, CPMD XYZ, CRT, CHARMM DCD, Chem3D, ChemDraw CDX, ChemSol, CSSR, EMPIRE, ESCHER NG, Fasta, GAMESS, Gaussian In/Out, GRAMM, Gromacs/Gromos, mol, Gromacs TRR, Gromacs XTC, HIN, IFF, InChI, LiGen pocket, MDL, MDL V3000, Mol2, Mopac cartesian, Mopac Gaussian Z-matrix, Mopac internal, MSF, NAMD binary, PDB, PDBA, PDBF, PDBL, PDBQT, PQR, PQRXML, PSFX, QMC, Quanta CSR, RIFF, SDF, TINKER XYZ, XYZ, ZIP.

5.4.2 HyperVolume (HV)

HyperVolume is a new descriptor based on the volume that a ligand occupies within the binding pocket. The C script – called *HyperVolume analyzer.c* – embedded in the VEGA ZZ suite is based on molecular docking studies (here performed using PLANTS software [86]) and analyses the shape of all generated poses within the considered binding pocket.

As input, the script requires the docked poses database (in mdb or mol2 format file). Considering a set of docked poses, HyperArea and HyperVolume are expressed

by computation of, respectively, shared area and shared volume and describe how much all poses differ compared to the top ranked pose. In this way, it is clear that only a HV descriptor is calculated for each ligand (and not for each pose, as FI fingerprints do).

HV descriptors comprehend a set of 8 descriptors:

- Area: best docked pose area (ranking based on docking scores)
- HyperArea: shared area of docked poses set
- DeltaArea: difference between HyperArea and Area
- RatioArea: ratio between HyperArea and Area
- Volume: best docked pose volume
- HyperVolume: shared volume of docked poses set
- DeltaVolume: difference between HyperVolume and Volume
- RatioArea: ratio between HyperVolume and Volume

5.4.3 Datasets Building

The evaluating benchmark started from the searching of proper datasets capable of providing a suitable assessment of developed descriptors. DUD-E database [166] was taken into account, since it collects various ligand datasets for each considered receptor. It is composed of 102 receptors, and for each one experimental active (hereinafter referred to as Actives), experimental inactive (hereinafter referred to as Inactives) and decoy molecules are categorised.

A preliminary benchmark was carried out considering the 25 datasets with complete experimental data (for both Actives and Inactives), containing at least 50 instances per class after cleaning and pre-processing steps.

All molecules – retrieved by ChEMBL [167] – with an affinity lower than 1μM were considered as Actives, while molecules with an affinity higher than 20μM were considered as Inactives. Then, a clusterisation step for inactive molecules was carried out, in order to obtain a heterogenous dataset. Their binary RDKit Morgan fingerprints [140] were compared through the Tanimoto coefficient and a threshold of 0.6 was set (molecules with a Tanimoto similarity higher than 0.6 were discarded). Since inactive molecules were stored in DUD-E database as SMILES strings, they were converted to 3D structure, hydrogen atoms were added and an

AMMP optimisation step [168] was carried out. In order to standardise the chemical space, reducing the bias as much as possible, DUD-E Actives dataset was filtered by considering the physicochemical properties (computed by VEGA ZZ software) of the inactive molecules, since they were the minority class. So, the Actives dataset was reduced about a third of the initial one and the receptor nos1 was excluded from further studies since its dataset became too small. Then, in order to test the novel descriptors on a larger dataset, an extended benchmark was assessed on 102 receptors. The positive samples were the same of previous benchmark, while decoys were considered as inactive set, based on DUD-E classification, from ZINC database [169]. For this dataset no cleaning step was needed, since already clusterised 3D-structures were downloadable from DUD-E webserver.

5.4.4 Descriptors Calculation

Structure-based descriptors were calculated starting from molecular docking studies.

Molecular docking was carried out using PLANTS software as implemented in VEGA ZZ suite with Docking+ tool. For each receptor, the published PDB structure was considered (as reported in DUD-E), a sphere radius of 8Å around the co-crystallised ligand was chosen as binding site and 5 poses for each ligand were generated through ChemPLP scoring function.

Then, Rescore+ was employed to compute 25 additional scoring function (in a similar way as done in Section 3.2.3), FingerInt descriptors, while the tailored C script calculated HyperVolume ones.

Since, in DUD-E database, for active molecules and decoys all possible protonation states are reported, a filtering step was needed. To select only one protonation state for each molecule, the top ranked tautomer according to docking results was chosen.

The resulting features were analysed in two different ways to train ML models. In the first case, for each descriptor, the value of the top ranked poses was considered for each ligand (BP approach). In the second case, all the docked poses were considered by computing the mean values of each descriptor for each ligand (AV approach).

5.4.5 Model Building and Evaluation

All the models described in this work were built by means tailored KNIME [80] workflows by using Random Forest algorithm. A Random Under Sampling step was carried out to balance each receptor dataset, based on the minority class; the generation of 5 samplings was set after few examinations (5, 10, 20 or 50 random sets) on which number of random samplings was the most appropriate.

Random Forest models were built with these settings: 10-fold cross-validation with stratified sampling, information gain ratio as quality measure, 100 decision trees maximum depth of trees as 10.

Since the primary objective was to evaluate the impact of novel descriptors in SBVS, in this work the contribution of these descriptors was assessed compared to Rescore+ analysis alone.

5.5 Results and Discussion

To analyse such a variety of generated models, among the resulted evaluation metrics, the MCC value was mainly taken into consideration.

5.5.1 Preliminary Benchmark (Experimental Data)

Rescore+ analysis alone showed a various performance among receptor sets, showing a mean MCC value of 0.468 in BP approach and 0.482 in AV approach, maybe due to the large target diversity and those related to molecule dataset.
While FI fingerprints or HV descriptors alone may not provide a substantial contribution, their potential was assessed in conjunction with Rescore+ analysis, showing an overall improvement of the SB models performance as depicted in Table 5.1.

In this way, FI fingerprints demonstrated a satisfactory performance, with a mean MCC value of 0.479 (BP) and 0.505 (AV) on the entire set of 25 receptors, corresponding to a percentage improvement – compared to Rescore+ analysis alone – of 2.48% and 4.85% respectively. 18 out of 25 (72%) receptors showed an evident

(a) **BP approach**

	Rescore+	**Rescore+ & FI**	**Rescore+ & HV**	**Rescore+ & FI & HV**
Mean MCC value	0.468	0.479	0.471	0.473
Improvement (%) (compared to Rescore+)		**2.48%**	0.57%	1.08%

(b) **AV approach**

	Rescore+	**Rescore+ & FI**	**Rescore+ & HV**	**Rescore+ & FI & HV**
Mean MCC value	0.482	0.505	0.486	0.503
Improvement (%) (compared to Rescore+)		**4.85%**	0.71%	4.41%

Table 5.1: Mean MCC values of the entire experimental set, for each scenario (Rescore+ analysis alone, "Rescore+ & FI", "Rescore+ & HV" and "Rescore+ & FI & HV"), (a) in BP and (b) AV approaches, with corresponding percentage improvement. The best results are bolded.

improved predicting power compared to Rescore+ analysis alone, resulting in a 6.03% in BP approach and 8.58% in AV one (Table 5.2). 7 receptors for BP and 14 for AV gave an improvement higher than 5%.

	Rescore+	**Rescore+ & FI**
Mean MCC value	0.460	0.499
Improvement (%) (compared to Rescore+)		8.58%

Table 5.2: Assessment of FI descriptors impact: mean MCC values and related percentage considering 18 out of 25 receptors, in AV approach (best scenario).

Considering the entire set of 25 receptors, HyperVolume descriptors reported a mean MCC value of 0.471 (BP) and 0.486 (AV), corresponding to a percentage improvement of 0.57% and 0.71% compared to Rescore+ analysis alone. Indeed, they showed only a slight performance improvement, since in BP approach, 12 out of 25 (48%) receptors showed a percentage improvement of 5.74% compared to Rescore+

analysis alone (Table 5.3), while in AV approach 15 out of receptors (60%) showed an improvement of 4.95%. 7 (28%) receptors for BP and 5 for AV gave an improvement higher than 5%.

	Rescore+	**Rescore+ & HV**
Mean MCC value	0.476	0.502
Improvement (%) (compared to Rescore+)		5.74%

Table 5.3: Assessment of HV descriptors impact: mean MCC values and related percentage considering 12 out of 25 receptors, in BP approach (best scenario).

Finally, a global model was assessed, comprehending both the new kind of descriptors, FI and HV. The conjunction of FI and HV with Rescore+ descriptors gave a percentage improvement of 1.08% for BP approach and 4.41% for AV approach (Table 5.1).

In Figures 5.1 and 5.2, histograms describe the MCC value on vertical axis for each receptor, for each receptor set.

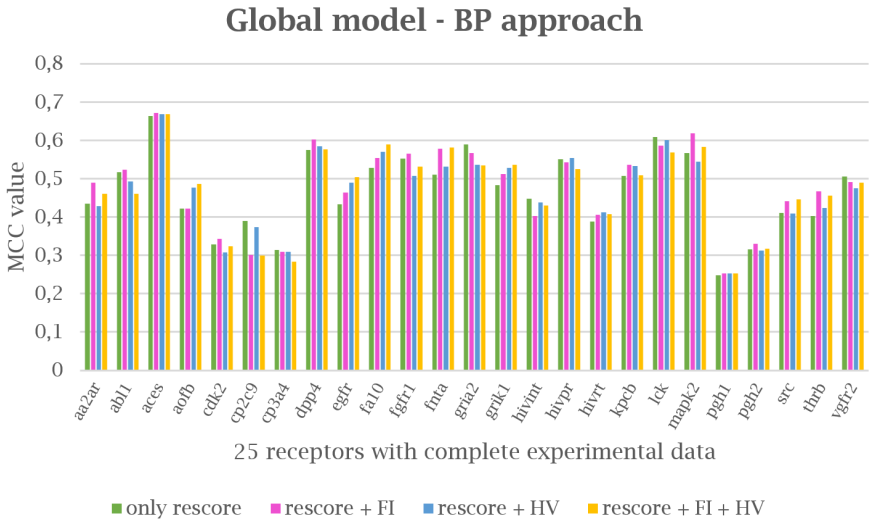

Figure 5.1: MCC values for each experimental set, considering the BP approach.

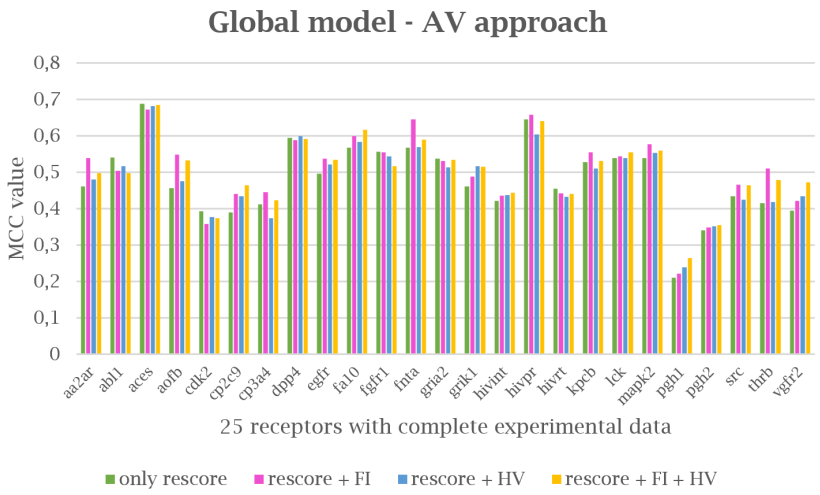

Figure 5.2: MCC values for each experimental set, considering the AV approach.

Considering the entire 25 receptors set, FI fingerprints showed a more satisfactory performance than HV descriptors gave. Taken individually, only some receptors performed in a better way in the combined mode "Rescore+ & HV" than "Rescore+ & FI" – *e.g.* aofb, cp2c9, egfr, hivint demonstrated a clear preference (mean MCC value difference higher than 5%) for "Rescore+ & FI" in BP approach while cdk2, grik1 and pgh1 in AV approach. Some receptors (4 for BP approach and 4 for AV approach), instead, suggested a clear positive contribution of HV descriptors in conjunction with "Rescore+" and FI fingerprints, rather than "Rescore+ & FI" alone (mean MCC value difference higher than 5%).

Since their promising results, to better investigate the FI fingerprints ability to enhance the predicting power in structure-based virtual screening, the new protein-ligand fingerprints were compared with published alternatives, such as ProLIF [154] and PLEC [153].
ProLIF (variable feature number, due to established interactions) gave a comparable result as they showed, in conjunction with Rescore+ descriptors, a percentage improvement of 3.96% (BP approach) and 12.32% (AV approach) compared to "Rescore+ & FI" analysis (Table 5.4).

It is worth noticing that the adenosine A2 receptor (aa2ar) and cp2c9 and cp3a4 cytochromes reported a considerable difference between ProLIF and FI, influencing the mean values in extensive manner. Hypothesising these receptors as outliers and discarding them, 22 out of 25 ones demonstrates MCC improvements of 3.11% for FI fingerprints and 3.17% for ProLIF using the BP approach, and 3.92% vs 9.69% using the AV approach (Table 5.5).

(a) **BP approach**

	Rescore+	**Rescore+ & FI**	**Rescore+ & HV**	**Rescore+ & FI & HV**	**ProLIF**	**ProLIF & Rescore+**
Mean MCC value	0.468	0.479	0.471	0.473	0.407	0.498
Improvement (%) (compared to Rescore+)		**2.48%**	0.57%	1.08%	-13.05%	**6.53%**
Improvement (%) (compared to Rescore+ & FI)					-15.15%	**3.96%**

(b) **AV approach**

	Rescore+	**Rescore+ & FI**	**Rescore+ & HV**	**Rescore+ & FI & HV**	**ProLIF**	**ProLIF & Rescore+**
Mean MCC value	0.482	0.505	0.486	0.503	0.480	0.568
Improvement (%) (compared to Rescore+)		**4.85%**	0.71%	4.41%	-0.35%	**17.77%**
Improvement (%) (compared to Rescore+ & FI)					-4.96%	**12.32%**

Table 5.4: Mean MCC values of the entire experimental set, for each already seen scenario (Rescore+ analysis alone, "Rescore+ & FI", "Rescore+ & HV" and "Rescore+ & FI & HV"), compared to ProLIF fingerprints, (a) in BP and (b) AV approaches, with corresponding percentage improvement. The best results are red, while the main percentage differences are bolded.

(a) **BP approach**

	Rescore+	**Rescore+ & FI**	**Rescore+ & HV**	**Rescore+ & FI & HV**	**ProLIF**	**ProLIF & Rescore+**
Mean MCC value	0.480	0.495	0.484	0.490	0.394	0.495
Improvement (%) (compared to Rescore+)		**3.11%**	0.88%	2.08%	-17.84%	**3.17%**
Improvement (%) (compared to Rescore+ & FI)					-20.31%	**0.06%**

(b) **AV approach**

	Rescore+	**Rescore+ & FI**	**Rescore+ & HV**	**Rescore+ & FI & HV**	**ProLIF**	**ProLIF & Rescore+**
Mean MCC value	0.490	0.510	0.493	0.509	0.442	0.538
Improvement (%) (compared to Rescore+)		**3.92%**	0.56%	3.79%	-9.85%	**9.69%**
Improvement (%) (compared to Rescore+ & FI)					-13.25%	**5.55%**

Table 5.5: Mean MCC values of 22 out of 25 receptors, for each already seen scenario (Rescore+ analysis alone, "Rescore+ & FI", "Rescore+ & HV" and "Rescore+ & FI & HV"), compared to ProLIF fingerprints, (a) in BP and (b) AV approaches, with corresponding percentage improvement. The best results are red, while the main percentage differences are bolded.

It is possible to conclude that FingerInt and ProLIF performances on considered datasets appeared as comparable descriptors; although ProLIF showed a slightly better predictive improvement, their computational and data processing time represents a significant drawback, while FI fingerprints benefit from fast and simple calculations, returning always the same number of interpretable features.

PLEC set (4092 features) showed a huge difference compared to FI and ProLIF performances (Table 5.6).

(a) **BP approach**

	Rescore+	**Rescore+ & FI**	**Rescore+ & HV**	**Rescore+ & FI & HV**	**PLEC**	**PLEC & Rescore+**
Mean MCC value	0.468	0.479	0.471	0.473	0.693	0.694
Improvement (%) (compared to Rescore+)		2.48%	0.57%	1.08%	48.12%	48.24%
Improvement (%) (compared to Rescore+ & FI)					44.54%	44.65%

(b) **AV approach**

	Rescore+	**Rescore+ & FI**	**Rescore+ & HV**	**Rescore+ & FI & HV**	**PLEC**	**PLEC & Rescore+**
Mean MCC value	0.482	0.505	0.486	0.503	0.733	0.739
Improvement (%) (compared to Rescore+)		4.85%	0.71%	4.41%	52.85%	53.23%
Improvement (%) (compared to Rescore+ & FI)					45.78%	46.14%

Table 5.6: Mean MCC values of the entire experimental set, for each already seen scenario (Rescore+ analysis alone, "Rescore+ & FI", "Rescore+ & HV" and "Rescore+ & FI & HV"), compared to PLEC fingerprints, (a) in BP and (b) AV approaches, with corresponding percentage improvement. The best results are red.

Indeed, the resulting models proved an excellent predictive power of PLEC even if the only descriptors used (Rescore+ descriptors improved marginally these models). The percentage difference between PLEC fingerprints and "Rescore+ & FI" is higher than 44%, in both BP and AV approaches. In conclusion, PLEC IFP provided meticulous descriptions of protein-ligand interaction, maybe due to their extensive feature mapping – or an overfitting phenomenon too, since their high numerousness vs small datasets – but their interpretation remains challenging.

5.5.2 Extended Benchmark (Actives and Decoys)

After the preliminary evaluation, an extended benchmark was assessed on entire DUD-E database, comprehending 102 receptor datasets where Actives were the positive samples, while decoys were considered for negative class.

The so generated models confirmed the experimental set trend: indeed, "Rescore+ & FI" or in conjunction with HV descriptors showed satisfactory performance, although the overall metrics are enhanced – starting from Rescore+ analysis alone – compared to previous benchmark (Table 5.7). This may be due to potential biases occurring in decoy datasets, as explained in Section 1.3.

(a) **BP approach**

	Rescore+	**Rescore+ & FI**	**Rescore+ & HV**	**Rescore+ & FI & HV**
Mean MCC value	0.645	0.658	0.651	0.664
Improvement (%) (compared to Rescore+)		**2.02%**	0.95%	**2.87**%

(b) **AV approach**

	Rescore+	**Rescore+ & FI**	**Rescore+ & HV**	**Rescore+ & FI & HV**
Mean MCC value	0.672	0.694	0.675	0.694
Improvement (%) (compared to Rescore+)		**3.27%**	0.43%	**3.16**%

Table 5.7: Mean MCC values of the extended set, for each scenario (Rescore+ analysis alone, "Rescore+ & FI", "Rescore+ & HV" and "Rescore+ & FI & HV"), (a) in BP and (b) AV approaches, with corresponding percentage improvement. The best results are bolded.

The "Rescore+ & FI" combination has been confirmed as the best in AV approach, with a percentage improvement of 3.27% on to "Rescore+" alone, compared to 2.02% in BP approach. HyperVolume descriptors alone didn't show a significant impact on the model performance (less than 1%), while the conjunction with Rescore+ and FI showed a slightly better performance: 2.87% in BP approach and 3.16% in AV one.

Building models with Rescore+ descriptors, in conjunction with FI, 74 out of 102

receptors (73%) reported a mean percentage improvement 3.68%, considering BP approach, while 82 out of 102 (80%) targets reported an improvement of 4.79%. 19 receptors for BP and 33 for AV gave an improvement higher than 5%.

On the other hand, HyperVolume descriptors gave a positive impact on model performance in 58 targets (57%), with a mean percentage improvement of 3.69%, in BP approach, and 2.75% for 51 receptors (50%) in AV approach. 14 receptors for BP and 9 for AV gave an improvement higher than 5%.

5.6 Conclusion

This work focused on the development of two novel structure-based descriptors, called FingerInt (FI) and HyperVolume (HV), and their validation using the freely-accessible molecular database DUD-E.

Several classification models for activity prediction assessment were developed using various combinations of descriptors, applying Random Forest algorithms. The analysis of the results demonstrated that FingerInt significantly improved performance compared to using Rescore+ alone.

FingerInt offers a balanced trade-off between computational cost and predictive performance, showing a significant improvement for the majority of tested receptors. Moreover, the models generated by using FI fingerprints are easily interpretable and thus are well suited for explainable artificial intelligence (XAI) applications.

Furthermore, combining FI fingerprints with HyperVolume descriptors – with the aim to explore this novel descriptor space – in most cases the predictive performance increases, even if the prediction enhancement by HV alone may not be significant enough. Surely, the accessible way to calculate HV and the low required computational cost represent a great plus point that renders this descriptor set worthy of further deepening and optimisation. Finally, as discussed for FI fingerprints, the interpretability of models generated by using HV descriptors is evident and so valuable in XAI applications. Further exploration and optimisation of the novel type of descriptors, enhancing their descriptive and predictive power, could lead to broader applications and deeper insights in molecular recognition studies.

Chapter 6

Conclusions and perspectives

Artificial Intelligence has become an integral part of modern drug discovery, improving and enhancing various crucial phases, such as hit identification, lead optimisation and ADMET assessment. The most desired outcomes are the analysis of large datasets and accurate but still reliable predictions.

The main aim of this Thesis was to investigate and improve the use of AI methods in drug discovery, focusing on creating predictive models able to accurately identify potential drug candidates, through various strategies.

Starting from the investigation of a peculiar metabolic reaction, the GSH conjugation, Chapter 3 focuses on xenobiotics metabolism prediction. The occurrence of this phase II reaction is traditionally determinated through the identification of toxicophores; in order to overcome all the related drawbacks, this work started from a rigorous curation step, with an appropriate selection of input data and a precise pre-processing step. The main focus was on investigating the involvement of the enzyme structure, as well as the metabolite structure, in the metabolism prediction. Globally, the obtained results were satisfactory, where LB models overperformed compared to SB ones. As well discussed in Section 3.2.1, the low predictive power of SB models could be ascribed to docking intrinsic static – and so approximate – treatment of underlying interactions, that do not consider the solvation and entropic effect of the system, as well as the promiscuity of metabolic enzymes. Future directions and possible improvements could be achieved using a PCM approach, taking into account the polyspecificity and diverse substrate preferences of the GST family.

Furthermore, the xenobiotic metabolism assessment was expanded to the entire set of known metabolic reaction classes. MetaSpot approach, based on MetaQSAR-derived data, provided satisfactory performance, highlighting the possibility of successfully recognising the reactive atoms. The already published MetaClass and MetaSpot can have a synergistic and complementary role, since the latter can be used to recognise the reactive atoms for the potential substrates identified by the former. Surely, enriched metabolic data will improve the predictive power of models by increasing the range of reactions they cover. Furthermore, this data can be used to create adequate external validation sets for refining the models.

In order to consider a multi-ligand and multi-target system, during my internship at the University of Vienna, I performed a predictive modelling on the ABCB transporter subfamily with a PCM approach, described in Chapter 4. Given this complex system, PCM allowed the combination of ligand information, with target ones, added to cross-term features. The generated predictive models showed good performances, but due to the fuzziness of the binding site, the investigation of crucial interacting residues proved challenging. In all models, from Feature Importance analysis emerged that ligands, as well as targets, contribute to the predicting power of the model. Future investigations, in conjunction with an increased availability of experimental data, could contribute to a major knowledge of this complex subfamily.

Finally, given the extreme importance of chemical descriptors in predictive modelling, Chapter 5 described the development and validation of novel structure-based descriptors, FingerInt (FI) and HyperVolume (HV).
The results of several classification models highlighted that FI performed significantly better than other known SB descriptors in most cases. When they were combined with HV descriptors, the predictive performance was improved in most cases. FingerInt emerged with their good balance between computational cost and predictive accuracy. Both FI and HV are highly interpretable, and surely new refinements could lead to more powerful models and new insights into molecular recognition.

As demonstrated by case studies presented in this Thesis, AI approaches have proven to be a highly advantageous tool in the drug discovery process, offering enhanced predictive accuracy, efficiency and reliability of the results. The predictive models developed provided valuable insights into molecular interactions and posed the basis for further advances in the field.

Looking ahead, continuing refinement of AI algorithms and integration with

emerging technologies can accelerate and enhance the entire drug discovery process. Among the most promising AI directions, Graph Neural Networks (GNNs) and neural sequence-to-sequence (Seq2Seq) models are worth mentioning since their remarkable demonstrated potential in addressing complex challenges in molecular design and drug development. [170][171][172]

By exploiting GNNs, molecular properties can be predicted, as well as the lead compound optimisation and identification of novel drugs, thanks to their ability to describe both the local chemical environment and global molecular topology.

Originally developed for natural language processing tasks, the sequence generation method of Seq2Seq architectures has been adapted for chemical and protein structures modelling, ADMET prediction and synthetic pathways rationalisation.

Combining GNNs, Seq2Seq models with other emerging techniques such as transfer learning, reinforcement learning or multi-task learning can enhance their applicability. The integration of these tools with high-throughput experimental screenings and the real-world data expansion will be pivotal to fill the gap between computational predictions and experimental validations.

Appendix I

MetaQSAR classification of the Metabolic Reactions according to Main Class, Class, Subclass Scheme with the corresponding relative abundance of each Subclass as monitored in the here collected database. [50]

main classes		classes		subclasses
code	description	code	description	code
01	redox	01	oxidation of Csp^3	01, 02, 03, 04, 05
		02	oxidation of Csp^2 and Csp	01, 02, 03, 04
		03	$-CHOH \leftrightarrow >C=O \rightarrow -COOH$	01, 02, 03, 04
		04	various redox reactions of carbon atoms	01, 02, 03, 04
		05	redox reactions of R_3N	01, 02, 03
		06	oxidation of $>NH$, $>NOH$, $-N=O$ /reduction of $-NO_2$, $-N=O$, $>NOH$, etc.	01, 02, 03, 04, 05, 06, 07, 08
		07	oxidation to quinones or analogues/reduction of quinones and analogues	01, 02, 03, 04, 05

		08	oxidation and reduction of S atoms	01, 02, 03, 04, 05, 06, 07, 08, 09
		09	redox reactions of other atoms	01, 02
02	hydrolysis and other	11	hydrolysis of esters, lactones, and inorganic esters	01, 02, 03, 04, 05, 06, 07, 08
		12	hydrolysis of amides, lactams and peptides	01, 02, 03, 04
		13	epoxide hydration	01
		14	other hydrolyses/hydration reactions/ nonenzymatic eliminations and rearrangements	01, 02, 03, 04, 05, 06, 07, 08, 09, 10
03	conjugations	21	O-glucuronidations and glycosylations	01, 02, 03, 04
		22	N- and S-glucuronidations/all other glycosilations	01, 02, 03, 04
		23	sulfonations ($O-$, $N-$, ...)	01, 02, 03
		24	GSH and RSH conjugations + sequels/GSH-mediated reductions	01, 02, 03, 04, 05, 06, 07

		25	acetylations and acylations	01, 02, 03, 04
		26	CoASH-ligation followed by amino acid conjugations or other sequels	01, 02, 03, 04, 05
		27	methylations ($O-$, $N-$, $S-$)	01, 02, 03, 04
		28	other conjugations (PO_4, CO_2, ...)/transaminations	01, 02, 03

Appendix II

MOPAC keywords and their description.[173]

Keyword	**Description**
++	Extend the line of keywords to the next line of the input file
0SCF	Read input file, then stop before the SCF cycle
1ELECTRON	Print the matrix of one-electron resonance integrals
1SCF	Converge one SCF cycle, then stop before geometry relaxation
A0	Read the input geometry in atomic units (bohr)
ADD-H	Add hydrogen atoms to close the valence shell of an organic molecule
AIGIN	Read the input geometry in Gaussian Z-matrix format
AIGOUT	Write the output geometry in Gaussian Z-matrix format to the .arc file

ALLBONDS	Print the final bond-order matrix, including weak bonds and H atoms (MOZYME only)
ALLVEC	Print all MO eigenvectors
ALT_A	In PDB files with alternative atoms, select atoms A
AM1	Use the AM1 model
ANGSTROMS	Read the input geometry in angstroms
AUTOSYM	Automatically impose a subset of SYMMETRY constraints
AUX	Output auxiliary information for use by other programs
BANANA	Generate localised molecular orbitals with hybrid orbitals for double bonds
BAR	reduce bar length by a maximum of n.nn%
BCC	Only even unit cells used (used by BZ)
BFGS	Use the Flepo or BFGS geometry optimiser

BIGCYCLES	Do a maximum of n big steps
BIRADICAL	Target a biradical ground state
BONDS	Print the final bond-order matrix
BZ	Output data for Program BZ - analysis of the Brillouin Zones
C.A.S.	Define a complete active space for MRCI (INDO only)
C.I.	Define an active space for CI calculations (INDO compatible)
C.I.D.	Define an active space for double excitations (INDOonly)
CAMP	Use Camp-King converger in SCF
CARTAB	Print point-group character table
CHAINS	In a protein, explicitely define the letters of chains.
CHARGE	Total electric charge on the system (MOZYME compatible)

CHARGES	Print net charge on system, and all charges in the system
CHARST	Print details of working in CHARST
CHECK	Report possible faults in input geometry
CIS	Restrict an active space to 1-electron excitations
CISD	Restrict an active space to 1 & 2-electron excitations (INDO compatible)
CISDT	Restrict an active space to 1, 2, & 3-electron excitations
COMPARE	Compare the geometries of two systems
COMPFG	Print heat of formation calculated in COMPFG
COSCCH	Add in COSMO charge corrections
COSWRT	Write details of the solvent accessible surface to a file
CUTOFF	In MOZYME, the interatomic distance where the NDDO approximation stops (default: 10 Ångstroms)

CUTOFP	Madelung distance cutoff is n.nn Ångstroms. This can speed up the calculations (default: 30 Ångstroms)
CUTOFS	In MOZYME, the interatomic distance beyond which overlap integrals are ignored (default: 7 Ångstroms)
CVB	In MOZYME. add and remove specific bonds to allow a Lewis or PDB structure.
CYCLES	Do a maximum of n steps
DAMP	n MOZYME. damp SCF oscillations using a factor of n.nn
DATA	Input data set is re-defined to text
DCART	Print part of working in DCART
DDMAX	See EF code
DDMIN	Minimum trust radius in a EF/TS calculation
DEBUG	Debug option turned on
DEBUGPULAY	Print working in PULAY

DENOUT	Density matrix output
DENOUTF	Formatted density matrix output
DENSITY	Print final density matrix
DERI1	Print part of working in DERI1
DERI2	Print part of working in DERI2
DERITR	Print part of working in DERIT
DERIV	Print part of working in DERIV
DERNVO	Print part of working in DERNVO
DFORCE	Force calculation specified, also print force matrix.
DFP	Use Davidson-Fletcher-Powell method to optimise geometries
DIPOLE	In animations graphs, replace ΔHf with dipole
DISEX	Distance for interactions in fine grid in COSMO

DISP	Print post-SCF corrections to the heat of formation
DMAX	Maximum stepsize in eigenvector following
DOUBLET	Target a doublet spin state
DRC	Dynamic reaction coordinate calculation
DUMP	Write restart files every n seconds
ECHO	Data are echoed back before calculation starts
EF	Use the EigenFollowing routine for geometry optimisation
EIGEN	Print canonical eigenvectors instead of LMOs in MOZYME calculations
EIGS	Print all eigenvalues in ITER
ENPART	Partition energy into components
EPS	Dielectric constant in COSMO calculation

ESP	Do not use. Use GRAPHF instead.
ESPGRID	Do not use. Use GRAPHF instead.
ESR	Calculate RHF spin density
EXCITED	Target the first excited singlet state
EXTERNAL	Read parameters off disk
FIELD	An external electric field is to be used
FILL	In RHF open and closed shell, force M.O. n to be filled
FLEPO	Print details of geometry optimisation
FMAT	Print details of working in FMAT
FOCK	Print last Fock matrix
FORCE	Calculate vibrational frequencies
FORCETS	Calculate vibrational frequencies for a transition state

FREQCY	Print symmetrised Hessian in a FORCE calculation
GEO-OK	Override some safety checks
GEO_DAT	Read in geometry from the file <text>
GEO_REF	Read in a second geometry from the file <text>
GNORM	Exit when gradient norm drops below n .n kcal/mol/Angstrom
GRADIENTS	Print all gradients
GRAPH	Generate unformatted file for graphics
GRAPHF	Generate formatted file for graphics suitable for Jmol and MOPETE.
H-PRIORITY	Heat of formation takes priority in DRC
HCORE	Print all parameters used, the one-electron matrix, and two-electron integrals
HESS	Options for calculating Hessian matrices in EF

HESSIAN	Print Hessian from geometry optimisation
HTML	Write a web-page for displaying and editing a protein
HYPERFINE	Hyperfine coupling constants to be calculated
INDO	Use the INDO/S model for excited states
INT	Make all coordinates internal coordinates
INVERT	Reverse all optimisation flags
IONIZE	Do not use - use SITE=(IONIZE) instead
IRC	Intrinsic reaction coordinate calculation
ISOTOPE	Force matrix written to disk (channel 9)
ITER	Print details of working in ITER
ITRY	Set limit of number of SCF iterations to n
IUPD	Mode of Hessian update in eigenvector following

KINETIC	Excess kinetic energy added to DRC calculation
KING	Use Camp-King converger for SCF
LARGE	Print expanded output
LBFGS	Use the low-memory version of the BFGS optimiser
LET	Override certain safety checks
LEWIS	Print the Lewis structure
LINMIN	Print details of line minimisation
LOCAL	Print localised orbitals. These are also called Natural Bond Orbitals or NBO
LOCATE-TS	Given reactants and products, locate the transition state connecting them
LOG	Generate a log file
MAXCI	Maximum number of configurations in the active space (INDO only)
MECI	Print details of MECI calculations

MERS	Keyword generated by MAKPOL for use with programBZ
METAL	Make specified atoms 100% ionic
MICROS	Restrict an active space to a list of microstates
MINI	Reduce the size of the output by only printing specified atoms
MINMEP	Minimise MEP minima in the plane defined
MMOK	Use molecular mechanics correction to CONH bonds
MNDO	Use the MNDO model
MNDOD	Use the MNDO-d model
MODE	In EF, follow Hessian mode no. n
MOLDAT	Print details of working in MOLDAT
MOLSYM	Print details of working in MOLSYM
MOPAC	Use old MOPAC definition for 2nd and 3rd atoms

MOZYME	Use the Localised Molecular Orbital method to speed up the SCF
MRCI	Use an active space of excitations from multiple reference states (INDO only)
MS	Target a spin state by its spin quantum number
MULLIK	Print the Mulliken population analysis
N**2	In excited state COSMO calculations, set the value of N**2
NLLSQ	Minimise gradients using NLLSQ
NOANCI	Use numerical CI derivatives (analytical derivatives are the default)
NOCOMMENTS	Ignore all lines except ATOM, HETATM, and TER in PDB files
NOLOG	Suppress log file trail, where possible
NOMM	Do not use molecular mechanics correction to CONH bonds
NONET	Target a nonet spin state

NONR	Do not use Newton-Raphson method in EF
NOOPT	Do not optimise the coordinates of all atoms of type X
NOREOR	In symmetry work, FORCE calculations, and whenGEO_REF is used, use the supplied orientation
NORESEQ	Suppress the default re-sequencing of atoms to the PDB sequence
NOSWAP	Do not allow atom swapping when GEO_REF is used
NOSYM	Reduce point-group symmetry to C1
NOSYM	Do not put "TER"s in PDB files
NOTHIEL	Do not use Thiel's FSTMIN technique
NOTXT	Remove any text from atom symbols
NOXYZ	Do not print Cartesian coordinates
NSPA	Sets number of geometric segments in COSMO

NSURF	Number of surfaces in an ESP calculation
OCTET	Target an octet spin state
OLDCAV	In COSMO, use the old Solvent Accessible Surface calculation
OLDENS	Read initial density matrix off disk
OLDFPC	Use the old fundamental physical constants
OLDGEO	Previous geometry to be used
OMIN	In TS, minimum allowed overlap of eigenvectors
OPEN	Distribute electrons over partially occupied orbitals
OPT, OPT-X	Optimise coordinates of all atoms within n.nn Ångstroms of atoms labelled "text"
OPT, OPT-X	Optimise the coordinates of all atoms of type X
OUTPUT	Reduce the amount of output (useful for large systems)

P	An applied pressure of n.nn Newtons/m2 to be used
PDB=(text)	Input geometry is in protein data bank format
PDB=(text)	User defined chemical symbols in protein data base
PDBOUT	Output geometry in pdb format
PECI	Restrict an active space to 1-electron and paired 2-electron excitations
PI	Resolve density matrix into σ, π, and δ components
PKA	Print the pKa for ionisable hydrogen atoms attached to oxygen atoms
PL	Monitor convergence of density matrix in ITER
PM3	Use the PM3 model
PM6	Use the PM6 model
PM6-D3	Use PM6 with the D3 model for dispersion

PM6-D3H4	Use PM6 with the D3H4 model for dispersion and hydrogen bonding
PM6-D3H4X	Use PM6 with the D3H4X model for dispersion and hydrogen/halogen bonding
PM6-DH+	Use PM6 with the DH+ model for dispersion and hydrogen bonding
PM6-DH2	Use PM6 with the DH2 model for dispersion and hydrogen bonding
PM6-DH2X	Use PM6 with the DH2X model for dispersion and hydrogen/halogen bonding
PM7	Use the PM7 model (default)
PM7-TS	Use the PM7-TS model for transition states
PMEP	Complete semiempirical MEP calculation
PMEPR	Complete semiempirical MEP in a plane to be defined
POINT	Number of points in reaction path
POINT1	Number of points in first direction in grid calculation

POINT2	Number of points in second direction in grid calculation
POLAR	Calculate first, second and third order polarisabilities
POTWRT	In ESP, write out electrostatic potential to unit 21
POWSQ	Print details of working in POWSQ
PRECISE	More stringent criteria are used
P=n.nn	Apply pressure or tension to a solid or polymer
PRNT	Print details of geometry optimisation in EF
PRTCHAR	Print charges in ARC file
PRTINT	Print interatomic distances
PRTMEP	MEP contour data output to <filename>.mep
PRTXYZ	Print Cartesian coordinates
PULAY	Use Pulay's converger to obtain a SCF

QMMM	Incorporate environmental effects in the QM/MM approach
QPMEP	Charges derived from Wang-Ford type AM1 MEP
QUARTET	Target a quartet spin state
QUINTET	Target a quintet spin state
RABBIT	Generate localised molecular orbitals with hybrid orbitals for double bonds
RAMA	Print Ramachandra angles for the residues in a protein
RAPID	In MOZYME geometry optimisations, only use atoms being optimised in the SCF
RE-LOCAL	During and at end of MOZYME calculation, re-localise the LMOs
RECALC	In EF, recalculate Hessian every n steps
RELSCF	Default SCF criterion multiplied by n
REORTH	In MOZYME, re-orthogonalise LMO's each 10 SCF calculations.

RESEQ	Re-arrange the atoms to match the PDB convention
RESIDUES	Label each atom in a polypeptide with the amino acid residue
RESTART	Calculation restarted
RHF	Use a restricted Hartree-Fock Hamiltonian
RM1	Use the RM1 model
RMAX	In TS, maximum allowed ratio for energy change
RMIN	In TS, minimum allowed ratio for energy change
ROOT	Target an excited state by energy or symmetry
RSCAL	In EF, scale p-RFO to trust radius
RSOLV	Effective radius of solvent in COSMO
SADDLE	Optimise transition state

SCALE	Scaling factor for van der waals distance in ESP
SCFCRT	Default SCF criterion replaced by the value supplied
SCINCR	Increment between layers in ESP
SEPTET	Target a septet spin state
SETPI	In MOZYME, some π bonds are explicitly set by the user
SETUP	Extra keywords to be read from setup file
SEXTET	Target a sextet spin state
SHIFT	a damping factor of n defined to start SCF
SIGMA	Minimise gradients using SIGMA
SINGLET	Target a singlet spin state
SITE	Define ionisation state of residues in proteins
SLOG	In L-BFGS optimisation, use fixed step of length n .nn

SLOPE	Multiplier used to scale MNDO charges
SMOOTH	In a GRID calculation, remove artifacts caused by the order in which points are calculated
SNAP	Snap Z-matrix angles to common high-symmetry values
SPARKLE	Use sparkles instead of atoms with basis sets
SPIN	Print final UHF spin matrix
START_RES	Define starting residue numbers in a protein, if different from the default
STATIC	Calculate Polarisability using electric fields
STEP	Step size in path
STEP1	Step size n for first coordinate in grid calculation
STEP2	Step size n for second coordinate in grid calculation
STO3G	Deorthogonalise orbitals in STO-3G basis

SUPER	Print superdelocalisabilities
SYBYL	Output a file for use by Tripos's SYBYL program
SYMAVG	Average symmetry equivalent ESP charges
SYMMETRY	Impose Z-matrix and Cartesian coordinate constraints
SYMOIR	Print characters of eigenvectors and print number of I.R.s
SYMTRZ	Print details of working in subroutine SYMTRZ.
T	A time of n seconds requested
T-PRIORITY	Time takes priority in DRC
TDIP	Output transition-dipole moments between excited states (INDO only)
THERMO	Perform a thermodynamics calculation
TIMES	Print times of various stages

TRANS	The system is a transition state (used in thermodynamics calculation)
TRIPLET	Target a triplet spin state
TS	Using EF routine for TS search
UHF	Use an unrestricted Hartree-Fock Hamiltonian
VDW	Van der waals radius for atoms in COSMO defined by user
VDWM	Van der waals radius for atoms in MOZYME defined by user
VECTORS	Print final vectors
VELOCITY	Supply the initial velocity vector in a DRC calculation
WILLIAMS	Use Williams surface
WRTCI	Maximum number of excited states to print (INDO only)
WRTCONF	Coefficient threshold for printing components of an excited state (INDO only)

X-PRIORITY	Geometry changes take priority in DRC
XENO	Allow non-standard residues in proteins to be labelled.
XYZ	Do all geometric operations in Cartesian coordinates
Z	Number of mers in a cluster

Appendix III

Full list of LB descriptors: VEGA ZZ [78], MOPAC [173] and BlueDesc [79].

VEGA ZZ	
Descriptor Name	**Description**
Angles	Number of bond angles
Atoms	Number of atoms
Bonds	Number of bonds
Charge	Total charge
ChiralAtms	Number of chiral atoms
Dipole	Dipole moment
EzBnds	Number of atoms with E/Z geometric isomerism
FlexTorsions	Number of flexible torsions
Gyrrad	Gyration radius in Å
HbAcc	Number of H-bond acceptors atoms

HbDon	Number of H-bond donors atoms
HeavyAtoms	Number of heavy atoms
Impropers	Number of improper angles (out of plane or pyramidal angles)
Lipole	Lipophilic moment (it indicates the lipophilic distribution of the molecule)
Mass	Mass in Daltons
MassMI	Monoisotopic mass in Daltons
Ovality	Ovality of the molecule (if it's one, the molecule has a spherical shape)
PSA	Polar surface area in $Å^2$
Rings	Number of rings
SAS	Solvent accessible surface in $Å^2$
SAV	Solvent accessible volume in $Å^3$
Sdiam	Surface diameter (diameter of the sphere having the same surface of the molecule)
Surface	Surface area in $Å^2$

Torsions	Number of torsions
Vdiam	Volume diameter (diameter of the sphere having the same volume of the molecule)
VirtualLogP	Bernard Testa's Virtual LogP
Volume	Molecular volume in $Å^3$

MOPAC	
Descriptor Name	**Description**
HEAT_OF_FORMATION	Enthalpy variation, when one mole of a system is formed from its elements.
DIELECTRIC_ENERGY	Stabilisation energy from the interaction of the charges in the solute with the induced charges on the solvent accessible surface plus the electrostatic energy.
GRADIENT_NORM	Scalar of the vector of derivatives of the energy with respect to the geometric variables flagged for optimisation.
DIPOLE	Measure of the separation of positive and negative charges in the molecule, indicating its polarity.
NO._OF_FILLED_LEVELS	Number of filled levels: the count of occupied molecular orbitals in the electronic structure of the molecule.
IONIZATION_POTENTIAL	Minimum energy required to eject an electron out of a neutral atom or molecule in its ground state.

HOMO_ENERGY	Energy of the highest occupied molecular orbital.
LUMO_ENERGY	Energy of the lowest unoccupied molecular orbital.
MOLECULAR_WEIGHT	Mass of a given substance divided by the amount of a substance, defined as mol.
COSMO_AREA	Surface of the molecule that can be reached by the by the center of charge of a solvent molecule.
COSMO_VOLUME	Volume included in the COSMO surface.
CHARGE_ON_SYSTEM	Net charge of the molecule.
MULLIKEN_ELECTRONEGATIVITY	-α
PARR_&POPLE_ABSOLUTE_HARDNESS	$$\eta = \frac{\epsilon_{homo} - \epsilon_{lumo}}{2}$$
SCHUURMANN_MO_SHIFT_ALPHA	Avegage of the Homo and Lumo energies. $$\alpha = \frac{\epsilon_{homo} + \epsilon_{lumo}}{2}$$

EHOMO	Energy of the highest occupied molecular orbital.
ELUMO	Energy of the lowest unoccupied molecular orbital.
Dn_TOTAL	Nucleophilic delocalisabilities, Total sums of all DN(r) (nucleophilic delocalisabilities) for each reactive center (r) of the molecule. It's based on the molecular orbital expansion coefficients, also as a measure of energy stabilisation due to nucleophilic attack.
De_TOTAL	Electrophilic delocalisabilities, Total sums of all DE(r) (electrophilic delocalisabilities) for each reactive centre (r) of a molecule. It's based on the molecular orbital expansion coefficients, also as a measure for the energy stabilisation due to electrophilic attack.
piS_TOTAL	Self-polarisability πS of an atom.

BlueDesc	
Descriptor Name	**Description**
ATSc1	AutocorrelationDescriptorCharge
ATSc2	AutocorrelationDescriptorCharge
ATSc3	AutocorrelationDescriptorCharge
ATSc4	AutocorrelationDescriptorCharge
ATSc5	AutocorrelationDescriptorCharge
ATSm1	AutocorrelationDescriptorMass
ATSm2	AutocorrelationDescriptorMass
ATSm3	AutocorrelationDescriptorMass
ATSm4	AutocorrelationDescriptorMass
ATSm5	AutocorrelationDescriptorMass
SC-3	ConnectivityDescriptor (Cluster)
SC-4	ConnectivityDescriptor (Cluster)

SC-5	ConnectivityDescriptor (Cluster)
SC-6	ConnectivityDescriptor (Cluster)
VC-3	ConnectivityDescriptor (Cluster)
VC-4	ConnectivityDescriptor (Cluster)
VC-5	ConnectivityDescriptor (Cluster)
VC-6	ConnectivityDescriptor (Cluster)
SCH-3	ConnectivityDescriptor (Chain)
SCH-4	ConnectivityDescriptor (Chain)
SCH-5	ConnectivityDescriptor (Chain)
SCH-6	ConnectivityDescriptor (Chain)
VCH-3	ConnectivityDescriptor (Chain)
VCH-4	ConnectivityDescriptor (Chain)
VCH-5	ConnectivityDescriptor (Chain)
VCH-6	ConnectivityDescriptor (Chain)

SP-0	ConnectivityDescriptor (Path)
SP-1	ConnectivityDescriptor (Path)
SP-2	ConnectivityDescriptor (Path)
SP-3	ConnectivityDescriptor (Path)
SP-4	ConnectivityDescriptor (Path)
SP-5	ConnectivityDescriptor (Path)
SP-6	ConnectivityDescriptor (Path)
SP-7	ConnectivityDescriptor (Path)
VP-8	ConnectivityDescriptor (Path)
VP-9	ConnectivityDescriptor (Path)
VP-10	ConnectivityDescriptor (Path)
VP-11	ConnectivityDescriptor (Path)
VP-12	ConnectivityDescriptor (Path)
VP-13	ConnectivityDescriptor (Path)

VP-14	ConnectivityDescriptor (Path)
VP-15	ConnectivityDescriptor (Path)
SPC-4	ConnectivityDescriptor (PathCluster)
SPC-5	ConnectivityDescriptor (PathCluster)
SPC-6	ConnectivityDescriptor (PathCluster)
VPC-4	ConnectivityDescriptor (PathCluster)
VPC-5	ConnectivityDescriptor (PathCluster)
VPC-6	ConnectivityDescriptor (PathCluster)
chi0C	ConnectivityDescriptor (Chain)
chi1C	ConnectivityDescriptor (Chain)
chi0vC	ConnectivityDescriptor (Chain)
chi1vC	ConnectivityDescriptor (Chain)
AcidicGroups	NumberOfAcidicGroups
AliphaticOHGroups	NumberOfAliphaticOHGroups

AromaticBonds	NumberOfAromaticBonds
AromaticOHGroups	NumberOfAromaticOHGroups
BasicGroups	NumberOfBasicGroups
HBA1	NumberOfHBA1
HBA2	NumberOfHBA2
HBD1	NumberOfHBD1
HBD2	NumberOfHBD2
HeavyBonds	NumberOfHeavyBonds
HeteroCycles	NumberOfHeteroCycles
HydrophobicGroups	NumberOfHydrophobicGroups
NO2Groups	NumberOfNO2Groups
NumberOfAtoms	NumberOfAtoms
NumberOfB	NumberOfB
NumberOfBonds	NumberOfBonds

NumberOfBr	NumberOfBr
NumberOfC	NumberOfC
NumberOfCl	NumberOfCl
NumberOfF	NumberOfF
NumberOfHal	NumberOfHal
NumberOfI	NumberOfI
NumberOfN	NumberOfN
NumberOfO	NumberOfO
NumberOfP	NumberOfP
NumberOfS	NumberOfS
OSOGroups	NumberOfOSOGroups
SO2Groups	NumberOfSO2Groups
SOGroups	NumberOfSOGroups
nAtomLAC	NumberOfAtomsInLongest Aliphatic Chain

nAtomLC	NumberOfAtomsInLargestChain
nAtomP	NumberOfAtomsInLargestPiSystem
FractionRotatableBonds	FractionOfRotatableBonds
PPSA-1	ChargedPartialSurfaceArea (CPSA) Descriptor
PPSA-2	ChargedPartialSurfaceArea (CPSA) Descriptor
PPSA-3	ChargedPartialSurfaceArea (CPSA) Descriptor
PNSA-1	ChargedPartialSurfaceArea (CPSA) Descriptor
PNSA-2	ChargedPartialSurfaceArea (CPSA) Descriptor
PNSA-3	ChargedPartialSurfaceArea (CPSA) Descriptor
DPSA-1	ChargedPartialSurfaceArea (CPSA) Descriptor
DPSA-2	ChargedPartialSurfaceArea (CPSA) Descriptor

DPSA-3	ChargedPartialSurfaceArea (CPSA) Descriptor
FPSA-1	ChargedPartialSurfaceArea (CPSA) Descriptor
FPSA-2	ChargedPartialSurfaceArea (CPSA) Descriptor
FPSA-3	ChargedPartialSurfaceArea (CPSA) Descriptor
FNSA-1	ChargedPartialSurfaceArea (CPSA) Descriptor
FNSA-2	ChargedPartialSurfaceArea (CPSA) Descriptor
FNSA-3	ChargedPartialSurfaceArea (CPSA) Descriptor
WPSA-1	ChargedPartialSurfaceArea (CPSA) Descriptor
WPSA-2	ChargedPartialSurfaceArea (CPSA) Descriptor
WPSA-3	ChargedPartialSurfaceArea (CPSA) Descriptor

WNSA-1	ChargedPartialSurfaceArea (CPSA) Descriptor
WNSA-2	ChargedPartialSurfaceArea (CPSA) Descriptor
WNSA-3	ChargedPartialSurfaceArea (CPSA) Descriptor
RPCG	ChargedPartialSurfaceArea (CPSA) Descriptor
RNCG	ChargedPartialSurfaceArea (CPSA) Descriptor
RHSA	ChargedPartialSurfaceArea (CPSA) Descriptor
RNCS	ChargedPartialSurfaceArea (CPSA) Descriptor
RPCS	ChargedPartialSurfaceArea (CPSA) Descriptor
RPSA	ChargedPartialSurfaceArea (CPSA) Descriptor
THSA	ChargedPartialSurfaceArea (CPSA) Descriptor

TPSA	Molecular property descriptors
LOBMAX	LengthOverBreadthDescriptor
LOBMIN	LengthOverBreadthDescriptor
MOMI-R	MomentOfInertiaDescriptor
MOMI-X	MomentOfInertiaDescriptor
MOMI-Y	MomentOfInertiaDescriptor
MOMI-Z	MomentOfInertiaDescriptor
MOMI-XY	MomentOfInertiaDescriptor
MOMI-XZ	MomentOfInertiaDescriptor
MOMI-YZ	MomentOfInertiaDescriptor
GeometricalDiameter	GeometricalDiameter
GeometricalRadius	GeometricalRadius
LipinksiFailures	Number Of Failurues Of The Lipinksi's Rule Of Five
XLogP	XLogPDescriptor

apol	SumOfAtomicPolarizabilities
bpol	BPolDescriptor
LogP	LogP
MolarRefractivity	MolarRefractivity
MolecularWeight	MolecularWeight
PolarSurfaceArea	PolarSurfaceArea (PSA)
ECCEN	EccentricConnectivityIndexDescriptor
GRAV-1	GravitationalIndexDescriptor
GRAV-2	GravitationalIndexDescriptor
GRAV-3	GravitationalIndexDescriptor
GRAV-4	GravitationalIndexDescriptor
GRAV-5	GravitationalIndexDescriptor
GRAV-6	GravitationalIndexDescriptor
GRAVH-1	GravitationalIndexDescriptor

GRAVH-2	GravitationalIndexDescriptor
GRAVH-3	GravitationalIndexDescriptor
WPATH	WienerNumbersDescriptor
WPOL	WienerNumbersDescriptor
WTPT-1	WeightedPathDescriptor
WTPT-2	WeightedPathDescriptor
WTPT-3	WeightedPathDescriptor
WTPT-4	WeightedPathDescriptor
WTPT-5	WeightedPathDescriptor
Zagreb	ZagrebIndexDescriptor
GlobalTopologicalChargeIndex	GlobalTopologicalChargeIndex
GraphShapeCoefficient	GraphShapeCoefficient
KierShape1	KierShapeWithLength1
KierShape2	KierShapeWithLength2

KierShape3	KierShapeWithLength3
TopologicalDiameter	TopologicalDiameter
WA.unity	WHIMDescriptor
WD.unity	WHIMDescriptor
WG.unity	WHIMDescriptor
WK.unity	WHIMDescriptor
WT.unity	WHIMDescriptor
WV.unity	WHIMDescriptor
Weta1.unity	WHIMDescriptor
Weta2.unity	WHIMDescriptor
Weta3.unity	WHIMDescriptor
Wgamma1.unity	WHIMDescriptor
Wgamma2.unity	WHIMDescriptor
Wgamma3.unity	WHIMDescriptor

Wlambda1.unity	WHIMDescriptor
Wlambda2.unity	WHIMDescriptor
Wlambda3.unity	WHIMDescriptor
Wnu1.unity	WHIMDescriptor
Wnu2.unity	WHIMDescriptor

Appendix IV

Rescore+ descriptors [88].

Descriptor Name	**Description**
PLANTS_CHEMPLP	CHEMPLP score after the complex minimisation
PLANTS_CHEMPLP_NORM_HEVATMS	CHEMPLP normalised by the number of heavy atoms
PLANTS_CHEMPLP_NORM_WEIGHT	CHEMPLP normalised by the molecular mass of the ligand
PLANTS_PLP	PLP score after the complex minimisation
PLANTS_PLP_NORM_HEVATMS	PLP normalised by the number of heavy atoms
PLANTS_PLP_NORM_WEIGHT	PLP normalised by the molecular mass of the ligand
PLANTS_PLP95	PLP95 score after the complex minimisation
PLANTS_PLP95_NORM_HEVATMS	PLP95 normalised by the number of heavy atoms

PLANTS_PLP95_NORM_WEIGHT	PLP95 normalised by the molecular mass of the ligand
XS_HPScore	Hydrophobic pair score calculated by X-Scoresoftware
XS_HMScore	Hydrophobic match score calculated by X-Score software
XS_HSScore	Steric score calculated by X-Score software
XS_Average	Average score calculated by X-Score ((HPScore+ HMScore + HSScore) / 3)
XS_Binding	X-score binding energy
CHARMM	Lennard-Jones energy computed by CHARMM force-field
ELECT	Electrostatic energy as computed by Coulomb's law
ELECTDD	Elect with and without distance-dependent dielectric function

MLPINS	Molecular Lipophilicity Potential Interaction Score: it indicates the hydro/lipophilic complementarity between the ligand and the receptor
MLPINS_2	Like MLPinS but take in account the square distance between the interacting atoms
MLPINS_3	Like MLPinS but take in account the cubic distance between the interacting atoms
MLPINS_F	Like MLPinS but uses the Fermi equation to establish the effect of the distance between the interacting atoms
Contacts	Contacts score based on the number of surroundingresidues
Contacts_NORM_HEVATMS	Contacts score normalised per heavy atoms
Contacts_NORM_WEIGHT	Contacts score normalised per-molecular weight

Appendix V

RDKit descriptors: constitutional and molecular descriptors [140].

Constitutional (count-based) descriptors	
NumRotatableBonds	NumAmideBonds
NumAromaticCarbocycles	NumAliphaticHeterocycles
NumHeavyAtoms	NumAtoms
NumRings	NumAliphaticRings
NumAliphaticCarbocycles	NumAromaticHeterocycles
NumAromaticRings	NumSaturatedRings
NumHeteroAtoms	NumStereoCenters
NumUnspecifiedStereoCenters	NumSaturatedHeterocycles
NumSaturatedCarbocycles	

Molecule descriptors retrieved from RDKit	
Descriptor	**Description**
LogP	Octanol - water partition coefficient is a standard property determined for potential drug molecules by indicating a measure of lipophilicity.
SMR	Molar refractivity, intended to capture polarisability
LabuteASA	Approximate surface area description
TPSA	Topological Polar Surface Area of a molecule (surface belonging to polar atoms). It describes molecular transport through membranes in an identical way to the 3D polar surface area.
ExactMW	Exact molecular weight
NumLipinskiHBA	Number of Hydrogen bond acceptor atoms
NumLipinskiHBD	Number of Hydrogen bond donator atoms
FractionCSP3	Fraction of the carbon atoms that are SP3 hybridised
Chi0v - Chi4v - Chi0n - Chi4n	Molecular connectivity indexes that characterise the structural attributes of molecules
Kappa1 - Kappa3	Molecular shape indexes

HallKierAlpha	A descriptor developed by Hall and Kier to indicate the presence of non SP3 hybridised carbon atoms in molecule branches
slogp_VSA1 - 12	A set of 12 subdivided surface area descriptors that capture hydrophobic and hydrophilic effects and indicate the contribution to lipophilicity.
SMR_VSA1 - 10	A set of 10 subdivided surface descriptors that capture polarisability and indicate the contribution to the molar refractivity.
peoe_VSA1 - 14	A set of 14 subdivided surface descriptors that capture direct electrostatic interactions. peoe stands for Partial Equalisation of Orbital Electronegativities.
MQN1-42	A set of 42 Molecular Quantum Numbers that is defined as counts of simple structural features such as atom, bond, and ring type.

Bibliography

[1] J. Jiménez-Luna, F. Grisoni, N. Weskamp, and G. Schneider, "Artificial intelligence in drug discovery: recent advances and future perspectives," Expert Opinion on Drug Discovery, vol. 16, no. 9, pp. 949–959, 2021.

[2] L. Yang, Q. Guo, and L. Zhang, "AI-assisted chemistry research: a comprehensive analysis of evolutionary paths and hotspots through knowledge graphs," 6 2024.

[3] S. Singh, N. Kaur, and A. Gehlot, "Application of artificial intelligence in drug design: A review," 9 2024.

[4] A. A. Arabi, "Artificial Intelligence in Drug Design: Algorithms, Applications, Challenges and Ethics," Future Drug Discovery, vol. 3, 6 2021.

[5] marketsandmarkets.com, "Artificial Intelligence in Drug Discovery Market by Offering, Process (Target selection, Validation, Lead generation, optimization), Drug Design (Small molecule, Vaccine, Antibody, PK/PD), Dry Lab, Wet Lab (Single Cell analysis) & Region - Global Forecast to 2028," 11 2023.

[6] K. K. Kırboğa, S. Abbasi, and E. U. Küçüksille, "Explainability and white box in drug discovery," 7 2023.

[7] R. Alizadehsani, S. S. Oyelere, S. Hussain, R. R. Calixto, V. H. C. de Albuquerque, M. Roshanzamir, M. Rahouti, and S. K. Jagatheesaperumal, "Explainable Artificial Intelligence for Drug Discovery and Development – A Comprehensive Survey," 9 2023.

[8] I. Muegge and S. Oloff, "Advances in virtual screening," 12 2006.

[9] D. B. Kitchen, H. Decornez, J. R. Furr, and J. Bajorath, "Docking and scoring in virtual screening for drug discovery: Methods and applications," 11 2004.

[10] W. Tong, H. Hong, Q. Xie, L. Shi, H. Fang, and R. Perkins, "Assessing QSAR Limitations-A Regulatory Perspective," tech. rep., 2005.

[11] W. Sippl, "3D-QSAR – Applications, recent advances, and limitations," in Challenges and Advances in Computational Chemistry and Physics, vol. 8, pp. 103–125, Springer, 2010.

[12] I. Wallach, D. Bernard, K. Nguyen, G. Ho, A. Morrison, A. Stecula, A. Rosnik, A. M. O'Sullivan, A. Davtyan, B. Samudio, B. Thomas, B. Worley, B. Butler, C. Laggner, D. Thayer, E. Moharreri, G. Friedland, H. Truong, H. van den Bedem, H. L. Ng, K. Stafford, K. Sarangapani, K. Giesler, L. Ngo, M. Mysinger, M. Ahmed, N. J. Anthis, N. Henriksen, P. Gniewek, S. Eckert, S. de Oliveira, S. Suterwala, S. V. K. PrasadPrasad, S. Shek, S. Contreras, S. Hare, T. Palazzo, T. E. O'Brien, T. Van Grack, T. Williams, T. R. Chern, V. Kenyon, A. H. Lee, A. B. Cann, B. Bergman, B. M. Anderson, B. D. Cox, J. M. Warrington, J. M. Sorenson, J. M. Goldenberg, M. A. Young, N. DeHaan, R. P. Pemberton, S. Schroedl, T. M. Abramyan, T. Gupta, V. Mysore, A. G. Presser, A. A. Ferrando, A. D. Andricopulo, A. Ghosh, A. G. Ayachi, A. Mushtaq, A. M. Shaqra, A. K. L. Toh, A. V. Smrcka, A. Ciccia, A. S. de Oliveira, A. Sverzhinsky, A. M. de Sousa, A. I. Agoulnik, A. Kushnir, A. N. Freiberg, A. V. Statsyuk, A. R. Gingras, A. Degterev, A. Tomilov, A. Vrielink, A. A. Garaeva, A. Bryant-Friedrich, A. Caflisch, A. K. Patel, A. V. Rangarajan, A. Matheeussen, A. Battistoni, A. Caporali, A. Chini, A. Ilari, A. Mattevi, A. T. Foote, A. Trabocchi, A. Stahl, A. B. Herr, A. Berti, A. Freywald, A. G. Reidenbach, A. Lam, A. R. Cuddihy, A. White, A. Taglialatela, A. K. Ojha, A. M. Cathcart, A. A. Motyl, A. Borowska, A. D'Antuono, A. K. Hirsch, A. M. Porcelli, A. Minakova, A. Montanaro, A. Müller, A. Fiorillo, A. Virtanen, A. J. O'Donoghue, A. Del Rio Flores, A. E. Garmendia, A. Pineda-Lucena, A. T. Panganiban, A. Samantha, A. K. Chatterjee, A. L. Haas, A. S. Paparella, A. L. John, A. Prince, A. ElSheikh, A. M. Apfel, A. Colomba, A. O'Dea, B. N. Diallo, B. M. R. M. Ribeiro, B. A. Bailey-Elkin, B. L. Edelman, B. Liou, B. Perry, B. S. K. Chua, B. Kováts, B. Englinger, B. Balakrishnan, B. Gong, B. Agianian, B. Pressly, B. P. Salas, B. M. Duggan, B. V. Geisbrecht, B. W. Dymock, B. C. Morten, B. D. Hammock, B. E. F. Mota, B. C. Dickinson, C. Fraser, C. Lempicki, C. D. Novina, C. Torner, C. Ballatore, C. Bon, C. J. Chapman, C. L. Partch, C. T. Chaton, C. Huang, C. Y. Yang, C. M. Kahler, C. Karan, C. Keller, C. L. Dieck, C. Huimei, C. Liu, C. Peltier, C. K. Mantri, C. M. Kemet, C. E. Müller, C. Weber, C. M. Zeina, C. S. Muli, C. Moris-

seau, C. Alkan, C. Reglero, C. A. Loy, C. M. Wilson, C. Myhr, C. Arrigoni, C. Paulino, C. Santiago, D. Luo, D. J. Tumes, D. A. Keedy, D. A. Lawrence, D. Chen, D. Manor, D. J. Trader, D. A. Hildeman, D. H. Drewry, D. J. Dowling, D. J. Hosfield, D. M. Smith, D. Moreira, D. P. Siderovski, D. Shum, D. T. Krist, D. W. Riches, D. M. Ferraris, D. H. Anderson, D. R. Coombe, D. S. Welsbie, D. Hu, D. Ortiz, D. Alramadhani, D. Zhang, D. Chaudhuri, D. J. Slotboom, D. R. Ronning, D. Lee, D. Dirksen, D. A. Shoue, D. W. Zochodne, D. Krishnamurthy, D. Duncan, D. M. Glubb, E. L. M. Gelardi, E. C. Hsiao, E. G. Lynn, E. B. Silva, E. Aguilera, E. Lenci, E. T. Abraham, E. Lama, E. Mameli, E. Leung, E. M. Christensen, E. R. Mason, E. Petretto, E. F. Trakhtenberg, E. J. Rubin, E. Strauss, E. W. Thompson, E. Cione, E. M. Lisabeth, E. Fan, E. G. Kroon, E. Jo, E. M. García-Cuesta, E. Glukhov, E. Gavathiotis, F. Yu, F. Xiang, F. Leng, F. Wang, F. Ingoglia, F. van den Akker, F. Borriello, F. J. Vizeacoumar, F. Luh, F. S. Buckner, F. S. Vizeacoumar, F. B. Bdira, F. Svensson, G. M. Rodriguez, G. Bognár, G. Lembo, G. Zhang, G. Dempsey, G. Eitzen, G. Mayer, G. L. Greene, G. A. Garcia, G. L. Lukacs, G. Prikler, G. C. G. Parico, G. Colotti, G. De Keulenaer, G. Cortopassi, G. Roti, G. Girolimetti, G. Fiermonte, G. Gasparre, G. Leuzzi, G. Dahal, G. Michlewski, G. L. Conn, G. D. Stuchbury, G. R. Bowman, G. M. Popowicz, G. Veit, G. E. de Souza, G. Akk, G. Caljon, G. Alvarez, G. Rucinski, G. Lee, G. Cildir, H. Li, H. E. Breton, H. Jafar-Nejad, H. Zhou, H. P. Moore, H. Tilford, H. Yuan, H. Shim, H. Wulff, H. Hoppe, H. Chaytow, H. K. Tam, H. Van Remmen, H. Xu, H. M. Debonsi, H. B. Lieberman, H. Jung, H. Y. Fan, H. Feng, H. Zhou, H. J. Kim, I. R. Greig, I. Caliandro, I. Corvo, I. Arozarena, I. N. Mungrue, I. M. Verhamme, I. A. Qureshi, I. Lotsaris, I. Cakir, J. J. P. Perry, J. Kwiatkowski, J. Boorman, J. Ferreira, J. Fries, J. M. Kratz, J. Miner, J. L. Siqueira-Neto, J. G. Granneman, J. Ng, J. Shorter, J. H. Voss, J. M. Gebauer, J. Chuah, J. J. Mousa, J. T. Maynes, J. D. Evans, J. Dickhout, J. P. MacKeigan, J. N. Jossart, J. Zhou, J. Lin, J. Xu, J. Wang, J. Zhu, J. Liao, J. Xu, J. Zhao, J. Lin, J. Lee, J. Reis, J. Stetefeld, J. B. Bruning, J. B. Bruning, J. G. Coles, J. J. Tanner, J. M. Pascal, J. So, J. L. Pederick, J. A. Costoya, J. B. Rayman, J. J. Maciag, J. A. Nasburg, J. J. Gruber, J. M. Finkelstein, J. Watkins, J. M. Rodríguez-Frade, J. A. S. Arias, J. J. Lasarte, J. Oyarzabal, J. Milosavljevic, J. Cools, J. Lescar, J. Bogomolovas, J. Wang, J. M. Kee, J. M. Kee, J. Liao, J. C. Sistla, J. S. Abrahão, K. Sishtla, K. R. Francisco, K. B. Hansen, K. A. Molyneaux, K. A. Cunningham, K. R. Martin, K. Gadar, K. K. Ojo, K. S. Wong, K. L. Wentworth, K. Lai, K. A. Lobb, K. M.

Hopkins, K. Parang, K. Machaca, K. Pham, K. Ghilarducci, K. S. Sugamori, K. J. McManus, K. Musta, K. M. Faller, K. Nagamori, K. J. Mostert, K. V. Korotkov, K. Liu, K. S. Smith, K. Sarosiek, K. H. Rohde, K. K. Kim, K. H. Lee, L. Pusztai, L. Lehtiö, L. M. Haupt, L. E. Cowen, L. J. Byrne, L. Su, L. Wert-Lamas, L. Puchades-Carrasco, L. Chen, L. H. Malkas, L. Zhuo, L. Hedstrom, L. Hedstrom, L. D. Walensky, L. Antonelli, L. Iommarini, L. Whitesell, L. M. Randall, M. D. Fathallah, M. H. Nagai, M. L. Kilkenny, M. Ben-Johny, M. P. Lussier, M. P. Windisch, M. Lolicato, M. L. Lolli, M. Vleminckx, M. C. Caroleo, M. J. Macias, M. Valli, M. M. Barghash, M. Mellado, M. A. Tye, M. A. Wilson, M. Hannink, M. R. Ashton, M. V. C. Cerna, M. Giorgis, M. K. Safo, M. S. Maurice, M. A. McDowell, M. Pasquali, M. Mehedi, M. S. M. Serafim, M. B. Soellner, M. G. Alteen, M. M. Champion, M. Skorodinsky, M. L. O'Mara, M. Bedi, M. Rizzi, M. Levin, M. Mowat, M. R. Jackson, M. Paige, M. Al-Yozbaki, M. A. Giardini, M. M. Maksimainen, M. De Luise, M. S. Hussain, M. Christodoulides, N. Stec, N. Zelinskaya, N. Van Pelt, N. M. Merrill, N. Singh, N. A. Kootstra, N. Singh, N. S. Gandhi, N. L. Chan, N. M. Trinh, N. O. Schneider, N. Matovic, N. Horstmann, N. Longo, N. Bharambe, N. Rouzbeh, N. Mahmoodi, N. J. Gumede, N. C. Anastasio, N. B. Khalaf, O. Rabal, O. Kandror, O. Escaffre, O. Silvennoinen, O. T. Bishop, P. Iglesias, P. Sobrado, P. Chuong, P. O'Connell, P. Martin-Malpartida, P. Mellor, P. V. Fish, P. O. L. Moreira, P. Zhou, P. Liu, P. Liu, P. Wu, P. Agogo-Mawuli, P. L. Jones, P. Ngoi, P. Toogood, P. Ip, P. von Hundelshausen, P. H. Lee, R. B. Rowswell-Turner, R. Balaña-Fouce, R. E. O. Rocha, R. V. Guido, R. S. Ferreira, R. K. Agrawal, R. K. Harijan, R. Ramachandran, R. Verma, R. K. Singh, R. K. Tiwari, R. Mazitschek, R. K. Koppisetti, R. T. Dame, R. N. Douville, R. C. Austin, R. E. Taylor, R. G. Moore, R. H. Ebright, R. M. Angell, R. Yan, R. Kejriwal, R. A. Batey, R. Blelloch, R. J. Vandenberg, R. J. Hickey, R. J. Kelm, R. J. Lake, R. K. Bradley, R. M. Blumenthal, R. Solano, R. M. Gierse, R. E. Viola, R. R. McCarthy, R. M. Reguera, R. V. Uribe, R. L. do Monte-Neto, R. Gorgoglione, R. T. Cullinane, S. Katyal, S. Hossain, S. Phadke, S. A. Shelburne, S. E. Geden, S. Johannsen, S. Wazir, S. Legare, S. M. Landfear, S. K. Radhakrishnan, S. Ammendola, S. Dzhumaev, S. Y. Seo, S. Li, S. Zhou, S. Chu, S. Chauhan, S. Maruta, S. R. Ashkar, S. L. Shyng, S. G. Conticello, S. Buroni, S. Garavaglia, S. J. White, S. Zhu, S. Tsimbalyuk, S. H. Chadni, S. Y. Byun, S. Park, S. Q. Xu, S. Banerjee, S. Zahler, S. Espinoza, S. Gustincich, S. Sainas, S. L. Celano, S. J. Capuzzi, S. N. Waggoner, S. Poirier, S. H. Olson, S. O. Marx, S. R. Van Doren, S. Sarilla, S. M. Brady-Kalnay,

S. Dallman, S. M. Azeem, T. Teramoto, T. Mehlman, T. Swart, T. Abaffy, T. Akopian, T. Haikarainen, T. L. Moreda, T. Ikegami, T. R. Teixeira, T. D. Jayasinghe, T. H. Gillingwater, T. Kampourakis, T. I. Richardson, T. J. Herdendorf, T. J. Kotzé, T. R. O'Meara, T. W. Corson, T. Hermle, T. H. Ogunwa, T. Lan, T. Su, T. Banjo, T. A. O'Mara, T. Chou, T. F. Chou, U. Baumann, U. R. Desai, V. P. Pai, V. C. Thai, V. Tandon, V. Banerji, V. L. Robinson, V. Gunasekharan, V. Namasivayam, V. F. Segers, V. Maranda, V. Dolce, V. G. Maltarollo, V. C. Scoffone, V. A. Woods, V. P. Ronchi, V. Van Hung Le, W. B. Clayton, W. T. Lowther, W. A. Houry, W. Li, W. Tang, W. Zhang, W. C. Van Voorhis, W. A. Donaldson, W. C. Hahn, W. G. Kerr, W. H. Gerwick, W. J. Bradshaw, W. E. Foong, X. Blanchet, X. Wu, X. Lu, X. Qi, X. Xu, X. Yu, X. Qin, X. Wang, X. Yuan, X. Zhang, Y. J. Zhang, Y. Hu, Y. A. Aldhamen, Y. Chen, Y. Li, Y. Sun, Y. Zhu, Y. K. Gupta, Y. Pérez-Pertejo, Y. Li, Y. Tang, Y. He, Y. C. Tse-Dinh, Y. A. Sidorova, Y. Yen, Y. Li, Z. J. Frangos, Z. Chung, Z. Su, Z. Wang, Z. Zhang, Z. Liu, Z. Inde, Z. Artía, and A. Heifets, "AI is a viable alternative to high throughput screening: a 318-target study," Scientific Reports, vol. 14, 12 2024.

[13] X. Qi, Y. Zhao, Z. Qi, S. Hou, and J. Chen, "Machine Learning Empowering Drug Discovery: Applications, Opportunities and Challenges," 2 2024.

[14] L. Chen, A. Cruz, S. Ramsey, C. J. Dickson, J. S. Duca, V. Hornak, D. R. Koes, and T. Kurtzman, "Hidden bias in the DUD-E dataset leads to misleading performance of deep learning in structure-based virtual screening," PLoS ONE, vol. 14, 8 2019.

[15] A. Cereto-Massagué, M. J. Ojeda, C. Valls, M. Mulero, S. Garcia-Vallvé, and G. Pujadas, "Molecular fingerprint similarity search in virtual screening," Methods, vol. 71, no. C, pp. 58–63, 2015.

[16] B. Mahesh, "Machine Learning Algorithms - A Review," International Journal of Science and Research (IJSR), vol. 9, pp. 381–386, 1 2020.

[17] W. Y. Loh, "Classification and regression trees," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 1, pp. 14–23, 1 2011.

[18] P. Tatjewski and M. Ławryńczuk, "Algorithms with state estimation in linear and nonlinear model predictive control," Computers and Chemical Engineering, vol. 143, 12 2020.

[19] Vujović, "Classification Model Evaluation Metrics," International Journal of Advanced Computer Science and Applications, vol. 12, no. 6, pp. 599–606, 2021.

[20] J. Fan, S. Upadhye, and A. Worster, "Understanding receiver operating characteristic (ROC) curves," Canadian Journal of Emergency Medicine, vol. 8, no. 1, pp. 19–20, 2006.

[21] V. Consonni, D. Ballabio, and R. Todeschini, "Evaluation of model predictive ability by external validation techniques," Journal of Chemometrics, vol. 24, no. 3-4, pp. 194–201, 2010.

[22] R. Todeschini, D. Ballabio, and F. Grisoni, "Beware of Unreliable Q2! A Comparative Study of Regression Metrics for Predictivity Assessment of QSAR Models," Journal of Chemical Information and Modeling, vol. 56, pp. 1905–1913, 10 2016.

[23] "Daylight Theory Manual."

[24] J. L. Durant, B. A. Leland, D. R. Henry, and J. G. Nourse, "Reoptimization of MDL Keys for Use in Drug Discovery," Journal of Chemical Information and Computer Sciences, vol. 42, pp. 1273–1280, 11 2002.

[25] D. Rogers and M. Hahn, "Extended-Connectivity Fingerprints," Journal of Chemical Information and Modeling, vol. 50, pp. 742–754, 5 2010.

[26] G. J. Van Westen, R. F. Swier, J. K. Wegner, A. P. Jzerman, H. W. Van Vlijmen, and A. Bender, "Benchmarking of protein descriptor sets in proteochemometric modeling (part 1): Comparative study of 13 amino acid descriptor sets," Journal of Cheminformatics, vol. 5, 9 2013.

[27] G. J. Van Westen, R. F. Swier, I. Cortes-Ciriano, J. K. Wegner, J. P. Overington, A. P. Jzerman, H. W. Van Vlijmen, and A. Bender, "Benchmarking of protein descriptor sets in proteochemometric modeling (part 2): Modeling performance of 13 amino acid descriptor sets," Journal of Cheminformatics, vol. 5, 9 2013.

[28] O. Korb, T. Stützle, and T. E. Exner, "Empirical scoring functions for advanced Protein-Ligand docking with PLANTS," Journal of Chemical Information and Modeling, vol. 49, no. 1, pp. 84–96, 2009.

[29] E. C. Meng, B. K. Shoichet, and I. D. Kuntz, "Automated docking with grid-based energy evaluation," Journal of Computational Chemistry, vol. 13, 1992.

[30] S. F. Sousa, P. A. Fernandes, and M. J. Ramos, "Protein-ligand docking: Current status and future challenges," 10 2006.

[31] V. Salmaso and S. Moro, "Bridging molecular docking to molecular dynamics in exploring ligand-protein recognition process: An overview," 8 2018.

[32] R. Wang, L. Lai, and S. Wang, "Further development and validation of empirical scoring functions for structure-based binding affinity prediction," tech. rep., 2002.

[33] G. Vistoli, A. Pedretti, A. Mazzolari, and B. Testa, "In silico prediction of human carboxylesterase-1 (hCES1) metabolism combining docking analyses and MD simulations," Bioorganic and Medicinal Chemistry, vol. 18, pp. 320–329, 1 2010.

[34] M. D. Eldridge, C. W. Murray, T. R. Auton, G. V. Paolini, and R. P. Mee, "Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes," tech. rep., 1997.

[35] I. A. Guedes, F. S. Pereira, and L. E. Dardenne, "Empirical scoring functions for structure-based virtual screening: Applications, critical aspects, and challenges," 9 2018.

[36] I. Muegge and Y. C. Martin, "A General and Fast Scoring Function for ProteinLigand Interactions: A Simplified Potential Approach," Journal of Medicinal Chemistry, vol. 42, pp. 791–804, 3 1999.

[37] M. Wójcikowski, P. J. Ballester, and P. Siedlecki, "Performance of machine-learning scoring functions in structure-based virtual screening," Scientific Reports, vol. 7, 2017.

[38] J. D. Durrant and J. A. McCammon, "NNScore 2.0: A neural-network receptor-ligand scoring function," Journal of Chemical Information and Modeling, vol. 51, pp. 2897–2903, 11 2011.

[39] J. Li, D. Q. Huang, B. Zou, H. Yang, W. Z. Hui, F. Rui, N. T. S. Yee, C. Liu, S. N. Nerurkar, J. C. Y. Kai, M. L. P. Teng, X. Li, H. Zeng, J. A. Borghi, L. Henry, R. Cheung, and M. H. Nguyen, "Epidemiology of COVID-19: A systematic review and meta-analysis of clinical characteristics, risk factors, and outcomes," Journal of Medical Virology, vol. 93, pp. 1449–1458, 3 2021.

[40] C. Shen, Y. Hu, Z. Wang, X. Zhang, H. Zhong, G. Wang, X. Yao, L. Xu, D. Cao, and T. Hou, "Can machine learning consistently improve the scoring power of classical scoring functions? Insights into the role of machine learning in scoring functions," Briefings in Bioinformatics, vol. 22, pp. 497–514, 1 2021.

[41] G. J. Van Westen, J. K. Wegner, A. P. Ijzerman, H. W. Van Vlijmen, and A. Bender, "Proteochemometric modeling as a tool to design selective compounds and for extrapolating to novel targets," 1 2011.

[42] P. Soucek, "Xenobiotics," in Encyclopedia of Cancer (M. Schwab, ed.), pp. 3964–3967, Berlin, Heidelberg: Springer Berlin Heidelberg, 2011.

[43] L. Di, "The role of drug metabolizing enzymes in clearance," 3 2014.

[44] A.-C. Macherey and P. M. Dansette, "Chapter 33 - Biotransformations Leading to Toxic Metabolites: Chemical Aspect," in The Practice of Medicinal Chemistry (Third Edition) (C. G. Wermuth, ed.), pp. 674–696, New York: Academic Press, 2008.

[45] P. Czodrowski, J. M. Kriegl, S. Scheuerer, and T. Fox, "Computational approaches to predict drug metabolism," 1 2009.

[46] B. Testa, A.-L. Balmat, A. Long, and P. Judson, "Predicting Drug Metabolism ± An Evaluation of the Expert System METEOR," tech. rep.

[47] D. S. Wishart, S. Tian, D. Allen, E. Oler, H. Peters, V. Lui, V. Gautam, Y. Djoumbou-Feunang, R. Greiner, and T. Metz, "BioTransformer 3.0—a web server for accurately predicting metabolic transformation products," Nucleic Acids Research, vol. 50, pp. W115–W123, 7 2022.

[48] M. Šícho, C. Stork, A. Mazzolari, C. de Bruyn Kops, A. Pedretti, B. Testa, G. Vistoli, D. Svozil, and J. Kirchmair, "FAME 3: Predicting the Sites of Metabolism in Synthetic Compounds and Natural Products for Phase 1 and Phase 2 Metabolic Enzymes," Journal of Chemical Information and Modeling, vol. 59, pp. 3400–3412, 8 2019.

[49] M. K. Matlock, T. B. Hughes, and S. J. Swamidass, "XenoSite server: a web-available site of metabolism prediction tool," Bioinformatics, vol. 31, pp. 1136–1137, 4 2015.

[50] A. Pedretti, A. Mazzolari, G. Vistoli, and B. Testa, "MetaQSAR: An Integrated Database Engine to Manage and Analyze Metabolic Data," Journal of Medicinal Chemistry, vol. 61, pp. 1019–1030, 2 2018.

[51] A. Mazzolari, L. Sommaruga, A. Pedretti, and G. Vistoli, "MetaTREE, a novel database focused on metabolic trees, predicts an important detoxification mechanism: The glutathione conjugation," Molecules, vol. 26, no. 7, 2021.

[52] K. Satoh, "The high non-enzymatic conjugation rates of some glutathione S-transferase (GST) substrates at high glutathione concentrations," Tech. Rep. 4, 1995.

[53] A. J. L. Cooper and M. H. Hanigan, "10.17 - Metabolism of Glutathione S-Conjugates: Multiple Pathways," in Comprehensive Toxicology (Third Edition) (C. A. McQueen, ed.), pp. 363–406, Oxford: Elsevier, third edition ed., 2018.

[54] P. J. Van Bladeren, "Glutathione conjugation as a bioactivation reaction," tech. rep., 2000.

[55] W. Dekant, "The role of biotransformation and bioactivation in toxicity," tech. rep., 2009.

[56] J. D. Hayes, D. J. Pulford, and K. D. Tew, "The Glutathione S-Transferase Supergene Family: Regulation of GST* and the Contribution of the lsoenzymes to Cancer Chemoprotection and Drug Resistance," Tech. Rep. 6.

[57] B. Mannervik, "The isoenzymes of glutathione transferase," Adv Enzymol Relat Areas Mol Biol, vol. 57, pp. 357–417, 1985.

[58] J. D. Hayes, R. C. Strange, and J. D. Hayes, "Glutathione S-Transferase Polymorphisms and Their Biological Consequences," tech. rep., 2000.

[59] P.-J. Jakobsson, R. Morgenstern, J. Mancini, A. Ford-Hutchinson, and B. Persson, "Common structural features of MAPEG-A widespread superfamily of membrane associated proteins with highly divergent functions in eicosanoid and glutathione metabolism," tech. rep., 1999.

[60] P. Reinemer, H. Dirr, R. Ladenstein, J. Schäffer, O. Gallay, and R. Huber, "The three-dimensional structure of class pi glutathione S-transferase in complex with glutathione sulfonate at 2.3 A resolution.," The EMBO Journal, vol. 10, pp. 1997–2005, 8 1991.

[61] A. Oakley, "Glutathione transferases: A structural perspective," 5 2011.

[62] J. E. Ladner, J. F. Parsons, C. L. Rife, G. L. Gilliland, and R. N. Armstrong, "Parallel evolutionary pathways for glutathione transferases: structure and

mechanism of the mitochondrial class kappa enzyme rGSTK1-1," Biochemistry, vol. 43, no. 2, pp. 352–361, 2004.

[63] J. Li, Z. Xia, and J. Ding, "Thioredoxin-like domain of human κ class glutathione transferase reveals sequence homology and structure similarity to the θ class enzyme," Protein science, vol. 14, no. 9, pp. 2361–2369, 2005.

[64] X. Ji, M. Tordova, J. F. Parsons, J. B. Hayden, G. L. Gilliland, P. Zimniak, and M. X. Memorial Hospital, "Downloaded via UNIV OF MILAN on," tech. rep., 1997.

[65] A.-M. Abdalla, C. M. Bruns, J. A. Tainer, B. Mannervik, and G. Stenberg, "Design of a monomeric human glutathione transferase GSTP1, a structurally stable but catalytically inactive protein," Tech. Rep. 10, 2002.

[66] R. Fabrini, A. De Luca, L. Stella, G. Mei, B. Orioni, S. Ciccone, G. Federici, M. Lo Bello, and G. Ricci, "Monomer-dimer equilibrium in glutathione transferases: A critical re-examination," Biochemistry, vol. 48, pp. 10473–10482, 11 2009.

[67] B. B. Bi Ochi ˜i C˜a, T. Kura, Y. Takahashi, T. Takayama, N. Ban, T. Saito, T. Kuga, and Y. Niitsu, "Glutathione S-transferase-Tr is secreted as a monomer into human plasma by platelets and tumor cells," tech. rep., 1996.

[68] A. Potęga, "Glutathione-Mediated Conjugation of Anticancer Drugs: An Overview of Reaction Mechanisms and Biological Significance for Drug Detoxification and Bioactivation," 8 2022.

[69] B. Coles, "Effects of Modifying Structure on Electrophilic Reactions with Biological Nucleop hi les*," Tech. Rep. 7, 1985.

[70] B. Ketterer, "Detoxication reactions of glutathione and glutathione transferases For personal use only," tech. rep., 1986.

[71] B. Testa and S. Krämer, "The Biochemistry of Drug Metabolism – An Introduction," Chemistry & Biodiversity, vol. 5, no. 11, pp. 2171–2336, 2008.

[72] A. E. Salinas and M. G. Wong, "Glutathione S-transferases–a review," Current medicinal chemistry, vol. 6, p. 279—309, 4 1999.

[73] J. Kazius, R. McGuire, and R. Bursi, "Derivation and validation of toxicophores for mutagenicity prediction," Journal of Medicinal Chemistry, vol. 48, pp. 312–320, 1 2005.

[74] T. J. Athersuch, D. J. Antoine, A. R. Boobis, M. Coen, A. K. Daly, L. Possamai, J. K. Nicholson, and I. D. Wilson, "Paracetamol metabolism, hepatotoxicity, biomarkers and therapeutic interventions: A perspective," Toxicology Research, vol. 7, no. 3, pp. 347–357, 2018.

[75] J. L. Raucy, J. M. Lasker, C. S. Lieber, and M. Blacks, "Acetaminophen Activation by Human Liver Cytochromes P450llEl and P450lA2'," Tech. Rep. 2, 1989.

[76] G. H. Hakimelahi and G. A. Khodarahmi, "The Identification of Toxicophores for the Prediction of Mutagenicity, Hepatotoxicity and Cardiotoxicity," Tech. Rep. 4, 2005.

[77] J. J. P. Stewart, "MOPAC: A semiempirical molecular orbital program," Journal of Computer-Aided Molecular Design, vol. 4, no. 1, pp. 1–103, 1990.

[78] A. Pedretti, A. Mazzolari, S. Gervasoni, L. Fumagalli, and G. Vistoli, "The VEGA suite of programs: An versatile platform for cheminformatics and drug design projects," Bioinformatics, vol. 37, pp. 1174–1175, 4 2021.

[79] J. Dong, D. S. Cao, H. Y. Miao, S. Liu, B. C. Deng, Y. H. Yun, N. N. Wang, A. P. Lu, W. B. Zeng, and A. F. Chen, "ChemDes: An integrated web-based platform for molecular descriptor and fingerprint computation," Journal of Cheminformatics, vol. 7, 12 2015.

[80] M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, C. Sieb, K. Thiel, and B. Wiswedel, "KNIME: The Konstanz Information Miner," in Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007), Springer, 2007.

[81] D. M. Townsend and K. D. Tew, "The role of glutathione-S-transferase in anti-cancer drug resistance," Oncogene, vol. 22, pp. 7369–7375, 10 2003.

[82] F. Morel, C. Rauch, E. Petit, A. Piton, N. Theret, B. Coles, and A. Guillouzo, "Gene and Protein Characterization of the Human Glutathione S-Transferase Kappa and Evidence for a Peroxisomal Localization," Journal of Biological Chemistry, vol. 279, pp. 16246–16253, 4 2004.

[83] M. Schwartz, F. Menetrier, J. M. Heydel, E. Chavanne, P. Faure, M. Labrousse, F. Lirussi, F. Canon, B. Mannervik, L. Briand, and F. Neiers, "Interactions between Odorants and Glutathione Transferases in the Human Olfactory Cleft," Chemical Senses, vol. 45, pp. 645–654, 10 2020.

[84] R. Anandakrishnan, B. Aguilar, and A. V. Onufriev, "H++ 3.0: Automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations," Nucleic Acids Research, vol. 40, 7 2012.

[85] J. C. Phillips, D. J. Hardy, J. D. C. Maia, J. E. Stone, J. V. Ribeiro, R. C. Bernardi, R. Buch, G. Fiorin, J. Hénin, W. Jiang, R. McGreevy, M. C. R. Melo, B. K. Radak, R. D. Skeel, A. Singharoy, Y. Wang, B. Roux, A. Aksimentiev, Z. Luthey-Schulten, L. V. Kalé, K. Schulten, C. Chipot, and E. Tajkhorshid, "Scalable molecular dynamics on CPU and GPU architectures with NAMD," The Journal of Chemical Physics, vol. 153, p. 044130, 7 2020.

[86] O. Korb, T. Stützle, and T. E. Exner, "PLANTS: Application of Ant Colony Optimization to Structure-Based Drug Design," in Ant Colony Optimization and Swarm Intelligence (M. Dorigo, L. M. Gambardella, M. Birattari, A. Martinoli, R. Poli, and T. Stützle, eds.), (Berlin, Heidelberg), pp. 247–258, Springer Berlin Heidelberg, 2006.

[87] G. Jones, P. Willett, R. C. Glen, A. R. Leach, and R. Taylor, "Development and Validation of a Genetic Algorithm for Flexible Docking," tech. rep., 1997.

[88] A. Pedretti, C. Granito, A. Mazzolari, and G. Vistoli, "Structural Effects of Some Relevant Missense Mutations on the MECP2-DNA Binding: A MD Study Analyzed by Rescore+, a Versatile Rescoring Tool of the VEGA ZZ Program," Molecular Informatics, pp. 424–433, 9 2016.

[89] G. Vistoli, A. Mazzolari, B. Testa, and A. Pedretti, "Binding Space Concept: A New Approach to Enhance the Reliability of Docking Scores and Its Application to Predicting Butyrylcholinesterase Hydrolytic Activity," Journal of Chemical Information and Modeling, vol. 57, pp. 1691–1702, 7 2017.

[90] L. Breiman, "Random Forests," tech. rep., 2001.

[91] V. Le Guilloux, P. Schmidtke, and P. Tuffery, "Fpocket: An open source platform for ligand pocket detection," BMC Bioinformatics, vol. 10, 5 2009.

[92] A. Mazzolari, A. Scaccabarozzi, G. Vistoli, and A. Pedretti, "MetaClass, a Comprehensive Classification System for Predicting the Occurrence of Metabolic Reactions Based on the MetaQSAR Database," Molecules, vol. 26, no. 19, 2021.

[93] B. Testa, A. Pedretti, and G. Vistoli, "Reactions and enzymes in the metabolism of drugs and other xenobiotics," Drug Discovery Today, vol. 17, no. 11, pp. 549–560, 2012.

[94] A. Mazzolari, A. M. Afzal, A. Pedretti, B. Testa, G. Vistoli, and A. Bender, "Prediction of UGT-mediated Metabolism Using the Manually Curated MetaQSAR Database," ACS Medicinal Chemistry Letters, vol. 10, pp. 633–638, 4 2019.

[95] L. H. Hall, B. Mohney, and L. B. Kier, "The electrotopological state: structure information at the atomic level for molecular graphs," Journal of Chemical Information and Computer Sciences, vol. 31, pp. 76–82, 2 1991.

[96] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," SIGKDD Explor. Newsl., vol. 11, pp. 10–18, 11 2009.

[97] A. Pedretti, A. Mazzolari, S. Gervasoni, and G. Vistoli, "Tree2C: A Flexible Tool for Enabling Model Deployment with Special Focus on Cheminformatics Applications," Applied Sciences, vol. 10, p. 7704, 10 2020.

[98] P. Broto, G. Moreau, and C. Vandycke, "Molecular structures: perception, autocorrelation descriptor and SAR studies. Autocorrelation descriptor," European journal of medicinal chemistry, vol. 19, no. 1, pp. 66–70, 1984.

[99] T. B. Hughes, G. P. Miller, and S. J. Swamidass, "Site of Reactivity Models Predict Molecular Reactivity of Diverse Chemicals with Glutathione," Chemical Research in Toxicology, vol. 28, pp. 797–809, 4 2015.

[100] L. R. Domingo, "Molecular Electron Density Theory: A Modern View of Reactivity in Organic Chemistry," Molecules, vol. 21, no. 10, 2016.

[101] C. F. Higgins, "ABC Transporters: From Microorganisms to Man," Annual Review of Cell and Developmental Biology, vol. 8, no. Volume 8, 1992, pp. 67–113, 1992.

[102] K. D. Bunting, "ABC transporters as phenotypic markers and functional regulators of stem cells," Stem cells, vol. 20, no. 1, pp. 11–20, 2002.

[103] G. Ferro-Luzzi Ames, C. S. Mimura, and V. Shyamala, "Bacterial periplasmic permeases belong to a family of transport proteins operating from Escherichia coli to human: traffic ATPases," FEMS microbiology reviews, vol. 6, no. 4, pp. 429–446, 1990.

[104] A. Alam and K. P. Locher, "Structure and Mechanism of Human ABC Transporters," Annual Review of Biophysics Downloaded from www.annualreviews.org. Guest, p. 10, 2024.

[105] R. L. Tatusov, E. V. Koonin, and D. J. Lipman, "A genomic perspective on protein families," Science, vol. 278, no. 5338, pp. 631–637, 1997.

[106] M. Dean, Y. Hamon, and G. Chimini, "The human ATP-binding cassette (ABC) transporter superfamily," 2001.

[107] R. Gaudet and D. C. Wiley, "Structure of the ABC ATPase domain of human TAP1, the transporter associated with antigen processing," The EMBO Journal, vol. 20, no. 17, pp. 4964–4972, 2001.

[108] C. Thomas, S. G. Aller, K. Beis, E. P. Carpenter, G. Chang, L. Chen, E. Dassa, M. Dean, F. Duong Van Hoa, D. Ekiert, R. Ford, R. Gaudet, X. Gong, I. B. Holland, Y. Huang, D. K. Kahne, H. Kato, V. Koronakis, C. M. Koth, Y. Lee, O. Lewinson, R. Lill, E. Martinoia, S. Murakami, H. W. Pinkett, B. Poolman, D. Rosenbaum, B. Sarkadi, L. Schmitt, E. Schneider, Y. Shi, S.-L. Shyng, D. J. Slotboom, E. Tajkhorshid, D. P. Tieleman, K. Ueda, A. Váradi, P.-C. Wen, N. Yan, P. Zhang, H. Zheng, J. Zimmer, and R. Tampé, "Structural and functional diversity calls for a new classification of ABC transporters," FEBS Letters, vol. 594, no. 23, pp. 3767–3775, 2020.

[109] M. L. Fitzgerald, Z. Mujawar, and N. Tamehiro, "ABC transporters, atherosclerosis and inflammation," Atherosclerosis, vol. 211, pp. 361–370, 8 2010.

[110] N. Tanaka, S. Abe-Dohmae, N. Iwamoto, M. L. Fitzgerald, and S. Yokoyama, "Helical apolipoproteins of high-density lipoprotein enhance phagocytosis by stabilizing ATP-binding cassette transporter A7," Journal of Lipid Research, vol. 51, pp. 2591–2599, 9 2010.

[111] L. Yvan-Charvet, N. Wang, and A. R. Tall, "Role of HDL, ABCA1, and ABCG1 transporters in cholesterol efflux and immune responses," 2 2010.

[112] W. A. Chutkow, M. C. Simon, M. M. Le Beau, and C. F. Burant, "Cloning, Tissue Expression, and Chromosomal Localization of SUR2, the Putative Drug-Binding Subunit of Cardiac, Skeletal Muscle, and Vascular Channels," tech. rep., 1996.

[113] L. Shyng, T. Ferrigni, J. B. Shepard, A. Nestorowicz, B. Glaser, M. A. Permutt, and C. G. Nichols, "Functional Analyses of Novel Mutations in the Sulfonylurea

Receptor 1 Associated With Persistent Hyperinsulinemic Hypoglycemia of Infancy," tech. rep., 1998.

[114] M. Harakalova, J. J. Van Harssel, P. A. Terhal, S. Van Lieshout, K. Duran, I. Renkens, D. J. Amor, L. C. Wilson, E. P. Kirk, C. L. Turner, D. Shears, S. Garcia-Minaur, M. M. Lees, A. Ross, H. Venselaar, G. Vriend, H. Takanari, M. B. Rook, M. A. Van Der Heyden, F. W. Asselbergs, H. M. Breur, M. E. Swinkels, I. J. Scurr, S. F. Smithson, N. V. Knoers, J. J. Van Der Smagt, I. J. Nijman, W. P. Kloosterman, M. M. Van Haelst, G. Van Haaften, and E. Cuppen, "Dominant missense mutations in ABCC9 cause Cantúsyndrome," Nature Genetics, vol. 44, pp. 793–796, 7 2012.

[115] J. F. Hunt, C. Wang, and R. C. Ford, "Cystic fibrosis transmembrane conductance regulator (ABCC7) structure," Cold Spring Harbor Perspectives in Medicine, vol. 3, 2 2013.

[116] A. Pujol, I. Ferrer, C. Camps, E. Metzger, C. Hindelang, N. Callizot, M. Ruiz, T. Pàmpols, M. Giròs, and J. L. Mandel, "Functional overlap between ABCD1 (ALD) and ABCD2 (ALDR) transporters: a therapeutic target for X-adrenoleukodystrophy," Human Molecular Genetics, vol. 13, pp. 2997–3006, 12 2004.

[117] D. Coelho, J. C. Kim, I. R. Miousse, S. Fung, M. Du Moulin, I. Buers, T. Suormala, P. Burda, M. Frapolli, M. Stucki, P. Nürnberg, H. Thiele, H. Robenek, W. Höhne, N. Longo, M. Pasquali, E. Mengel, D. Watkins, E. A. Shoubridge, J. Majewski, D. S. Rosenblatt, B. Fowler, F. Rutsch, and M. R. Baumgartner, "Mutations in ABCD4 cause a new inborn error of vitamin B12 metabolism," Nature Genetics, vol. 44, pp. 1152–1155, 10 2012.

[118] A. Kobayashi, Y. Takanezawa, T. Hirata, Y. Shimizu, K. Misasa, N. Kioka, H. Arai, K. Ueda, and M. Matsuo, "Efflux of sphingomyelin, cholesterol, and phosphatidylcholine by ABCG1," Journal of lipid research, vol. 47, no. 8, pp. 1791–1802, 2006.

[119] E. M. Leslie, R. G. Deeley, and S. P. C. Cole, "Multidrug resistance proteins: role of P-glycoprotein, MRP1, MRP2, and BCRP (ABCG2) in tissue defense," Toxicology and applied pharmacology, vol. 204, no. 3, pp. 216–237, 2005.

[120] Z. Wang, L. D. Stalcup, B. J. Harvey, J. Weber, M. Chloupkova, M. E. Dumont, M. Dean, and I. L. Urbatsch, "Purification and ATP Hydrolysis of

the Putative Cholesterol Transporters ABCG5 and ABCG8," Biochemistry, vol. 45, pp. 9929–9939, 8 2006.

[121] P. B. A. N. D. A. H. Schinkel, "P-glycoprotein ABCB1: a major player in drug handling by mammals," The Journal of Clinical Investigation, vol. 123, pp. 4131–4133, 10 2013.

[122] Annese V, Valvano M, Palmieri O, Latiano A, Bossa F, and Andriulli A, "Multidrug resistance 1 gene in inflammatory bowel disease: A meta-analysis," World J Gastroenterol., vol. 12, pp. 3636–3644, 6 2006.

[123] US Food Drug Admin. Cent.Drug Eval. Res. (US FDA)., "In vitro drug interaction studies—cytochrome P450 enzyme- and transporter-mediated drug interactions: guidance for industry.," 2020.

[124] R. P. J. O. Elferink, G. N. J. Tytgat, and A. K. Groen, "The role of mdr2 P-glycoprotein in hepatobiliary lipid transport," The FASEB journal, vol. 11, no. 1, pp. 19–28, 1997.

[125] J. M. L. De Vree, E. Jacquemin, E. Sturm, D. Cresteil, P. J. Bosma, J. Aten, J.-F. Deleuze, M. Desrochers, M. Burdelski, O. Bernard, and others, "Mutations in the MDR 3 gene cause progressive familial intrahepatic cholestasis," Proceedings of the National Academy of Sciences, vol. 95, no. 1, pp. 282–287, 1998.

[126] N. Y. Frank, S. S. Pendse, P. H. Lapchak, A. Margaryan, D. Shlain, C. Doeing, M. H. Sayegh, and M. H. Frank, "Regulation of progenitor cell fusion by ABCB5 P-glycoprotein, a novel human ATP-binding cassette transporter," Journal of Biological Chemistry, vol. 278, no. 47, pp. 47156–47165, 2003.

[127] C. Zhang, D. Li, J. Zhang, X. Chen, M. Huang, S. Archacki, Y. Tian, W. Ren, A. Mei, Q. Zhang, and others, "Mutations in ABCB6 cause dyschromatosis universalis hereditaria," Journal of Investigative Dermatology, vol. 133, no. 9, pp. 2221–2228, 2013.

[128] L. Wang, F. He, J. Bu, X. Liu, W. Du, J. Dong, J. D. Cooney, S. K. Dubey, Y. Shi, B. Gong, and others, "ABCB6 mutations cause ocular coloboma," The American Journal of Human Genetics, vol. 90, no. 1, pp. 40–48, 2012.

[129] A. Maguire, K. Hellier, S. Hammans, and A. May, "X-linked cerebellar ataxia and sideroblastic anaemia associated with a missense mutation in the ABC7

gene predicting V411L," British journal of haematology, vol. 115, no. 4, pp. 910–917, 2001.

[130] Y. Ichikawa, M. Bayeva, M. Ghanefar, V. Potini, L. Sun, R. K. Mutharasan, R. Wu, A. Khechaduri, T. Jairaj Naik, and H. Ardehali, "Disruption of ATP-binding cassette B8 in mice leads to cardiomyopathy through a decrease in mitochondrial iron export," Proceedings of the National Academy of Sciences, vol. 109, no. 11, pp. 4152–4157, 2012.

[131] J. C. Wolters, R. Abele, and R. Tampé, "Selective and ATP-dependent translocation of peptides by the homodimeric ATP binding cassette transporter TAP-like (ABCB9)," Journal of biological chemistry, vol. 280, no. 25, pp. 23631–23636, 2005.

[132] A. Mahajan, D. Taliun, M. Thurner, N. R. Robertson, J. M. Torres, N. W. Rayner, A. J. Payne, V. Steinthorsdottir, R. A. Scott, N. Grarup, and others, "Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps," Nature genetics, vol. 50, no. 11, pp. 1505–1513, 2018.

[133] A. P. Morris, B. F. Voight, T. M. Teslovich, T. Ferreira, A. V. Segrè, V. Steinthorsdottir, R. J. Strawbridge, H. Khan, H. Grallert, A. Mahajan, I. Prokopenko, H. M. Kang, C. Dina, T. Esko, R. M. Fraser, S. Kanoni, A. Kumar, V. Lagou, C. Langenberg, J. Luan, C. M. Lindgren, M. Müller-Nurasyid, S. Pechlivanis, N. W. Rayner, L. J. Scott, S. Wiltshire, L. Yengo, L. Kinnunen, E. J. Rossin, S. Raychaudhuri, A. D. Johnson, A. S. Dimas, R. J. Loos, S. Vedantam, H. Chen, J. C. Florez, C. Fox, C. T. Liu, D. Rybin, D. J. Couper, W. H. L. Kao, M. Li, M. C. Cornelis, P. Kraft, Q. Sun, R. M. Van Dam, H. M. Stringham, P. S. Chines, K. Fischer, P. Fontanillas, O. L. Holmen, S. E. Hunt, A. U. Jackson, A. Kong, R. Lawrence, J. Meyer, J. R. Perry, C. G. Platou, S. Potter, E. Rehnberg, N. Robertson, S. Sivapalaratnam, A. Stančáková, K. Stirrups, G. Thorleifsson, E. Tikkanen, A. R. Wood, P. Almgren, M. Atalay, R. Benediktsson, L. L. Bonnycastle, N. Burtt, J. Carey, G. Charpentier, A. T. Crenshaw, A. S. Doney, M. Dorkhan, S. Edkins, V. Emilsson, E. Eury, T. Forsen, K. Gertow, B. Gigante, G. B. Grant, C. J. Groves, C. Guiducci, C. Herder, A. B. Hreidarsson, J. Hui, A. James, A. Jonsson, W. Rathmann, N. Klopp, J. Kravic, K. Krjutškov, C. Langford, K. Leander, E. Lindholm, S. Lobbens, S. MäNnistö, G. Mirza, T. W. MüHleisen, B. Musk, M. Parkin, L. Rallidis, J. Saramies, B. Sennblad, S. Shah, G. Sigursson, A. Silveira, G. Steinbach,

B. Thorand, J. Trakalo, F. Veglia, R. Wennauer, W. Winckler, D. Zabaneh, H. Campbell, C. Van Duijn, A. G. Uitterlinden, A. Hofman, E. Sijbrands, G. R. Abecasis, K. R. Owen, E. Zeggini, M. D. Trip, N. G. Forouhi, A. C. Syvänen, J. G. Eriksson, L. Peltonen, M. M. Nöthen, B. Balkau, C. N. Palmer, V. Lyssenko, T. Tuomi, B. Isomaa, D. J. Hunter, L. Qi, A. R. Shuldiner, M. Roden, I. Barroso, T. Wilsgaard, J. Beilby, K. Hovingh, J. F. Price, J. F. Wilson, R. Rauramaa, T. A. Lakka, L. Lind, G. Dedoussis, I. NjøLstad, N. L. Pedersen, K. T. Khaw, N. J. Wareham, S. M. Keinanen-Kiukaanniemi, T. E. Saaristo, E. Korpi-HyöVäLti, J. Saltevo, M. Laakso, J. Kuusisto, A. Metspalu, F. S. Collins, K. L. Mohlke, R. N. Bergman, J. Tuomilehto, B. O. Boehm, C. Gieger, K. Hveem, S. Cauchi, P. Froguel, D. Baldassarre, E. Tremoli, S. E. Humphries, D. Saleheen, J. Danesh, E. Ingelsson, S. Ripatti, V. Salomaa, R. Erbel, K. H. JöCkel, S. Moebus, A. Peters, T. Illig, U. D. Faire, A. Hamsten, A. D. Morris, P. J. Donnelly, T. M. Frayling, A. T. Hattersley, E. Boerwinkle, O. Melander, S. Kathiresan, P. M. Nilsson, P. Deloukas, U. Thorsteinsdottir, L. C. Groop, K. Stefansson, F. Hu, J. S. Pankow, J. Dupuis, J. B. Meigs, D. Altshuler, M. Boehnke, and M. I. McCarthy, "Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes," Nature Genetics, vol. 44, pp. 981–990, 9 2012.

[134] Y. Zhang, F. Li, A. D. Patterson, Y. Wang, K. W. Krausz, G. Neale, S. Thomas, D. Nachagari, P. Vogel, M. Vore, and others, "Abcb11 deficiency induces cholestasis coupled to impaired β-fatty acid oxidation in mice," Journal of Biological Chemistry, vol. 287, no. 29, pp. 24784–24794, 2012.

[135] N. Shaikh, M. Sharma, and P. Garg, "Selective Fusion of Heterogeneous Classifiers for Predicting Substrates of Membrane Transporters," Journal of Chemical Information and Modeling, vol. 57, pp. 594–607, 3 2017.

[136] S. Kickinger, A. Seiler, D. Digles, and G. F. Ecker, "Proteochemometric modeling strengthens the role of Q299 for GABA transporter subtype selectivity,"

[137] Fabian Kaiser, "BSc Thesis - University of Vienna. Pharmacoinformatics Research Group."

[138] D. Young, T. Martin, R. Venkatapathy, and P. Harten, "Are the Chemical Structures in Your QSAR Correct?," QSAR & Combinatorial Science, vol. 27, no. 11-12, pp. 1337–1345, 2008.

[139] T. U. Consortium, "UniProt: the Universal Protein Knowledgebase in 2023," Nucleic Acids Research, vol. 51, pp. D523–D531, 1 2023.

[140] "RDKit: Open-source cheminformatics. https://www.rdkit.org."

[141] M. Sandberg, L. Eriksson, J. Jonsson, M. Sjöström, and S. Wold, "New Chemical Descriptors Relevant for the Design of Biologically Active Peptides. A Multivariate Characterization of 87 Amino Acids," Journal of Medicinal Chemistry, vol. 41, pp. 2481–2491, 7 1998.

[142] A. Bender, N. Schneider, M. Segler, W. Patrick Walters, O. Engkvist, and T. Rodrigues, "Evaluation guidelines for machine learning tools in the chemical sciences," Nature Reviews Chemistry, vol. 6, no. 6, pp. 428–442, 2022.

[143] L. M. Lagares, N. Minovski, A. Y. C. Alfonso, E. Benfenati, S. Wellens, M. Culot, F. Gosselet, and M. Novič, "Homology modeling of the human p-glycoprotein (Abcb1) and insights into ligand binding through molecular docking studies," International Journal of Molecular Sciences, vol. 21, pp. 1–37, 6 2020.

[144] Z. Li, R. Huang, M. Xia, T. A. Patterson, and H. Hong, "Fingerprinting Interactions between Proteins and Ligands for Facilitating Machine Learning in Drug Discovery," 1 2024.

[145] D. D. Wang, M. T. Chan, and H. Yan, "Structure-based protein–ligand interaction fingerprints for binding affinity prediction," 1 2021.

[146] Z. Zhao and P. E. Bourne, "Harnessing systematic protein–ligand interaction fingerprints for drug discovery," 10 2022.

[147] Z. Deng, C. Chuaqui, and J. Singh, "Structural Interaction Fingerprint (SIFt): A Novel Method for Analyzing Three-Dimensional ProteinLigand Binding Interactions," Journal of Medicinal Chemistry, vol. 47, pp. 337–344, 1 2004.

[148] V. I. Pérez-Nueno, O. Rabal, J. I. Borrell, and J. Teixidó, "APIF: A New Interaction Fingerprint Based on Atom Pairs and Its Application to Virtual Screening," Journal of Chemical Information and Modeling, vol. 49, pp. 1245–1260, 5 2009.

[149] J. Desaphy, E. Raimbaud, P. Ducrot, and D. Rognan, "Encoding Protein–Ligand Interaction Patterns in Fingerprints and Graphs," Journal of Chemical Information and Modeling, vol. 53, pp. 623–637, 3 2013.

[150] M. Radifar, N. Yuniarti, and E. P. Istyastono, "PyPLIF: Python-based protein-ligand interaction fingerprinting," Bioinformation, vol. 9, no. 6, p. 325, 2013.

[151] V. Chupakhin, G. Marcou, H. Gaspar, and A. Varnek, "Simple Ligand–Receptor Interaction Descriptor (SILIRID) for alignment-free binding site comparison," Computational and Structural Biotechnology Journal, vol. 10, no. 16, pp. 33–37, 2014.

[152] C. Da and D. Kireev, "Structural Protein–Ligand Interaction Fingerprints (SPLIF) for Structure-Based Virtual Screening: Method and Benchmark Study," Journal of Chemical Information and Modeling, vol. 54, pp. 2555–2561, 9 2014.

[153] M. Wójcikowski, M. Kukiełka, M. M. Stepniewska-Dziubinska, and P. Siedlecki, "Development of a protein-ligand extended connectivity (PLEC) fingerprint and its application for binding affinity predictions," Bioinformatics, vol. 35, pp. 1334–1341, 4 2019.

[154] C. Bouysset and S. Fiorucci, "ProLIF: a library to encode molecular interactions as fingerprints," Journal of Cheminformatics, vol. 13, 12 2021.

[155] F. Cramer, "Emil Fischer's Lock-and-Key Hypothesis after 100 years—Towards a Supracellular Chemistry," in Perspectives in Supramolecular Chemistry: The Lock-and-Key Principle, vol. 1, pp. 1 – 23, 10 2007.

[156] S. Putta and P. Beroza, "Shapes of Things: Computer Modeling of Molecular Shape in Drug Discovery," tech. rep., 2007.

[157] A. C. Good, T. J. A. Ewing, D. A. Gschwend, and I. D. Kuntz, "New molecular shape descriptors: Application in database screening," Journal of Computer-Aided Molecular Design, vol. 9, no. 1, pp. 1–12, 1995.

[158] M. Stahl and H.-J. Böhm, "Development of filter functions for protein–ligand docking," Journal of Molecular Graphics and Modelling, vol. 16, no. 3, pp. 121–132, 1998.

[159] I. Muegge, "Effect of ligand volume correction on PMF scoring," Journal of Computational Chemistry, vol. 22, no. 4, pp. 418–425, 2001.

[160] S. Kearnes and V. Pande, "ROCS-derived features for virtual screening," Journal of Computer-Aided Molecular Design, vol. 30, pp. 609–617, 8 2016.

[161] Y. Fukunishi and H. Nakamura, "Prediction of protein–ligand complex structure by docking software guided by other complex structures," Journal of Molecular Graphics and Modelling, vol. 26, no. 6, pp. 1030–1033, 2008.

[162] Y. Fukunishi and H. Nakamura, "A new method for in-silico drug screening and similarity search using molecular dynamics maximum volume overlap (MD-MVO) method," Journal of Molecular Graphics and Modelling, vol. 27, no. 5, pp. 628–636, 2009.

[163] Y. Fukunishi and H. Nakamura, "Integration of ligand-based drug screening with structure-based drug screening by combining maximum volume overlapping score with ligand docking," Pharmaceuticals, vol. 5, pp. 1332–1345, 12 2012.

[164] G. M. Sastry, S. L. Dixon, and W. Sherman, "Rapid Shape-Based Ligand Alignment and Virtual Screening Method Based on Atom/Feature-Pair Similarities and Volume Overlap Scoring," Journal of Chemical Information and Modeling, vol. 51, pp. 2455–2466, 10 2011.

[165] A. Nicholls, G. B. McGaughey, R. P. Sheridan, A. C. Good, G. Warren, M. Mathieu, S. W. Muchmore, S. P. Brown, J. A. Grant, J. A. Haigh, N. Nevins, A. N. Jain, and B. Kelley, "Molecular Shape and Medicinal Chemistry: A Perspective," Journal of Medicinal Chemistry, vol. 53, pp. 3862–3886, 5 2010.

[166] M. M. Mysinger, M. Carchia, J. J. Irwin, and B. K. Shoichet, "Directory of useful decoys, enhanced (DUD-E): Better ligands and decoys for better benchmarking," Journal of Medicinal Chemistry, vol. 55, pp. 6582–6594, 7 2012.

[167] B. Zdrazil, E. Felix, F. Hunter, E. J. Manners, J. Blackshaw, S. Corbett, M. de Veij, H. Ioannidis, D. M. Lopez, J. F. Mosquera, M. P. Magarinos, N. Bosc, R. Arcila, T. Kizilören, A. Gaulton, A. P. Bento, M. F. Adasme, P. Monecke, G. A. Landrum, and A. R. Leach, "The ChEMBL Database in 2023:ã drug disco v ery platf orm spanning multiple bioactivity data typesãnd time periods," Nucleic Acids Research, vol. 52, pp. D1180–D1192, 1 2024.

[168] R. W. Harrison, "Stiffness and energy conservation in molecular dynamics: An improved integrator," Journal of Computational Chemistry, vol. 14, no. 9, pp. 1112–1122, 1993.

[169] J. J. Irwin and B. K. Shoichet, "ZINC-A Free Database of Commercially Available Compounds for Virtual Screening," tech. rep.

[170] A. Lavecchia, "Advancing drug discovery with deep attention neural networks," 8 2024.

[171] X. Zeng, P. K. Feng, S. J. Li, S. Q. Lv, M. L. Wen, and Y. Li, "GNN-DDAS: Drug discovery for identifying anti-schistosome small molecules based on graph neural network," Journal of Computational Chemistry, 12 2024.

[172] R. A. Jacob, O. Wieder, Y. Chen, A. Mazzolari, A. Bergner, K.-J. Schleifer, and J. Kirchmair, "aweSOM: a GNN-based Site-of-Metabolism Predictor with Aleatoric and Epistemic Uncertainty Estimation," 10 2024.

[173] "MOPAC2016, James J. P. Stewart, Stewart Computational Chemistry, Colorado Springs, CO, USA,"