ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

«ΤΕΧΝΗΤΗ ΝΟΗΜΟΣΥΝΗ ΚΑΙ ΕΦΑΡΜΟΓΕΣ»

«MASTER OF SCIENCE IN ARTIFICIAL INTELLIGENCE AND APPLICATIONS»

Reliable ADME-Tox Prediction with Classical Machine Learning: Probability Calibration, OOD Detection and Drift Monitoring

ΑΔΑΜ ΜΑΡΙΑ του ΒΑΣΙΛΕΙΟΥ

ΙΟΥΝΙΟΣ 2026

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
«ΤΕΧΝΗΤΗ ΝΟΗΜΟΣΥΝΗ ΚΑΙ ΕΦΑΡΜΟΓΕΣ»

«Master of Science in Artificial Intelligence and Applications»

«Reliable ADME-Tox Prediction with Classical Machine Learning: Probability Calibration, OOD Detection and Drift Monitoring»

Διπλωματική Εργασία που υποβλήθηκε στο Τμήμα Πληροφορικής και Τηλεπικοινωνιών του Πανεπιστημίου Θεσσαλίας ως μέρος των απαιτήσεων για την απόκτηση Διπλώματος Μεταπτυχιακών Σπουδών Τεχνητή Νοημοσύνη και Εφαρμογές από τον/την

ΑΔΑΜ ΜΑΡΙΑ του ΒΑΣΙΛΕΙΟΥ

Δήλωση Αυθεντικότητας, ζητήματα Copyright

«Ο/Η μεταπτυχιακός/ή φοιτητής/τρια που εκπόνησε την παρούσα διπλωματική εργασία φέρει ολόκληρη την ευθύνη προσδιορισμού της δίκαιης χρήσης του υλικού, η οποία ορίζεται στη βάση των εξής παραγόντων: του σκοπού και χαρακτήρα της χρήσης (μη-εμπορικός, μη-κερδοσκοπικός, αλλά εκπαιδευτικός-ερευνητικός), της φύσης του υλικού που χρησιμοποιεί (τμήμα του κειμένου, πίνακες, σχήματα, εικόνες κ.λπ.), του ποσοστού και της σημαντικότητας του τμήματος που χρησιμοποιεί σε σχέση με το όλο κείμενο υπό copyright, και των πιθανών συνεπειών της χρήσης αυτής στην αγορά ή την γενικότερη αξία του υπό copyright κειμένου».

Ο/Η Μεταπτυχιακός/ή Φοιτητής/τρια

ΙΟΥΝΙΟΣ 2026

Τριμελής Εξεταστική Επιτροπή

Η παρούσα Μ.Δ.Ε. εγκρίθηκε ομόφωνα από την τριμελή εξεταστική επιτροπή η οποία ορίστηκε από την Σ.Ε. / Συνέλευση του Τμήματος Πληροφορικής και Τηλεπικοινωνιών του Πανεπιστημίου Θεσσαλίας, σύμφωνα με το νόμο και τον εγκεκριμένο Οδηγό Σπουδών του Π.Μ.Σ. «Τεχνητή Νοημοσύνη και Εφαρμογές». Τα μέλη της Επιτροπής ήταν:

1ο Μέλος (Επιβλέπων):	ΚΟΛΟΜΒΑΤΣΟΣ ΚΩΝΣΤΑΝΤΙΝΟΣ
2ο Μέλος:	
3ο Μέλος:	

Η έγκριση της Μ.Δ.Ε. από το Τμήμα Πληροφορικής και Τηλεπικοινωνιών του Πανεπιστημίου Θεσσαλίας, δεν υποδηλώνει αποδοχή των απόψεων του συγγραφέα.

Abstract

In recent years, the interest in the emerging technologies of Artificial Intelligence (AI) and Machine Learning (ML) has increased significantly across a wide range of fields, due to the constant demand for improvement and innovation. As a result, significant efforts have been made to adopt AI and ML technologies to increase the quality, efficiency and robustness of processes in various fields, including the fields of chemistry, biomedicine and pharmaceuticals. Key challenges associated with pharmaceutical R&D are optimizing the procedures of experimental testing in terms of time and cost and prioritizing candidate compounds more efficiently. A critical aspect of drug discovery is the evaluation of absorption, distribution, metabolism, excretion (ADME) and toxicity properties. Toxicity risk assessment refers to the systematic evaluation of a compound's potential to cause adverse biological effects, including both the likelihood and severity of toxic responses under specific exposure scenarios. Its strong dependence on dose and assay conditions increases the complexity of the process. Thus, the development and adoption of reliable ML models in addressing these challenges is of high importance. In this thesis toxicity and ADME prediction are investigated. The main objective of the thesis is the development of reliable ML models that not only produce well calibrated and accurate probability estimates, identify outliers and novel samples, but also trigger alerts if the data distribution shifts over time and performance degrades. To achieve that, high fidelity classical ML models were developed using visualizations and evaluation metrics, along with the out-of-distribution (OOD) detection and drift monitoring. The ML models were based on publicly available molecular datasets, and their performance was constantly monitored to further increase the quality and reliability of the results.

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

1.1 Motivation and Problem Statement

Prediction of absorption, distribution, metabolism, excretion and toxicity (ADME-Tox) is an important step in the drug discovery and development process, as it enables the early detection and evaluation of compounds with higher likelihood of safe and effective use in humans, while reducing the risk of costly failures in advanced stages of development [1]. At its core, drug development is a complex systems engineering endeavor in which the primary challenge lies in balancing therapeutic efficacy against rigorous safety thresholds [2]. In practical applications, toxicological evaluation is not just an intermediate stage but constitutes a significant link between fundamental research and clinical translation, acting as a critical ethical and economic filter. This filter not only optimizes development timelines and the overall costs but also protects public health by avoiding adverse reactions. This is underscored by the fact that approximately 30% to 40% of preclinical candidate compounds fail due to toxicity issues or insufficient ADMET profiles, making adverse toxicological reactions the leading cause of drug withdrawal from the market. [2]

For an extended period, in vitro assays and in vivo animal models have been the basis of drug safety assessments. However, their predictive capacity for human adverse drug effects remains limited. Animal models often fail to reproduce the complexity of human biological systems, leading to marked inconsistencies in ADME characteristics and toxicity patterns between preclinical and clinical stages [3]. This lack of translatability is compounded by substantial moral and financial burden. Traditional ADME and toxicity testing methods are not only fraught with ethical controversies regarding animal suffering but are also hindered by protracted testing durations and high resource requirements [4]. Within industrial pharmaceutical frameworks, these constraints have shifted the focus toward front-loading ADME evaluations. Shifting screening to the earliest possible stages is now a strategic priority. This necessity becomes even more pressing when considering the scale of modern chemical exploration. With thousands of novel molecules requiring assessment annually, even a marginal rate of unforeseen safety liabilities creates a prohibitive workload for research teams [5]. Consequently, reliance on physical assays

alone is becoming unsustainable. Drug discovery is increasingly pivoting toward scalable, computational decision support frameworks, leading to a transition toward Artificial Intelligence (AI) and Machine Learning (ML) architecture for predictive ADME-Tox modelling.

1.2 Thesis Scope and Methodology

While the adoption of Machine Learning has accelerated the molecular screening process, a critical reliability gap remains between algorithmic outputs and decision-making in the pharmaceutical industry. In high-risk environments such as drug discovery, a simple binary classification (e.g. toxic vs non-toxic) is rarely sufficient for comprehensive risk assessment. However, most contemporary ML architectures, especially high performing ensemble methods, suffer from poor probability calibration. This results in overconfident predictions that do not reflect the actual empirical risk. Such behavior can lead to a false sense of security, potentially promoting suboptimal compounds and undermining the very cost-saving goals discussed previously.

To address these concerns about reliability, this research examines the predictive integrity of three distinct categories of classical ML algorithms, each selected for its unique behavior regarding uncertainty. To determine a transparent and interpretable baseline, Logistic regression (LR) is first used, primarily due to its inherent probabilistic foundations which facilitate direct calibration assessments [6]. This linear approach is contrasted with the geometric strategy of Support Vector Machines (SVM). As margin-based classifiers, SVMs offer significant predictive power but do not natively produce calibrated probabilities, thus serving as a primary case study for post-hoc transformation techniques [7]. Finally, the study incorporates nonlinear ensemble learning through Gradient Boosted Decision Trees (XGBoost). These models represent the latest stage in predicting chemical data, but they are often prone to overconfidence, making them an essential testing ground for rigorous reliability monitoring [8].

To ensure that any observed differences in performance are due to the differences in learning architectures, we used a fixed molecular representation across all experiments. We converted SMILES strings to Morgan (ECFP) fingerprints, which are a commonly used representation to translate the complex topological structures of molecules into high-dimensional, fixed length vectors that can be used in conjunction with classical machine

learning algorithms [9]. The resulting features were used to train models on three independent datasets: Tox21/NR-AR, BBB_Martins, and DILI from the Therapeutics Data Commons (TDC) [10] [9], [10].

This methodology is structured around two primary research objectives. Beyond establishing standard performance baselines, the core of this thesis focuses on a threefold evaluation of model reliability. First, the model's outputs are analyzed into mathematically consistent probabilities, using Platt scaling and Isotonic regression, with their success measured through the Expected Calibration Error (ECE) [11]. Second, an Out-Of-Distribution (OOD) detection module is implemented to estimate the "applicability domain" of the models, effectively flagging high-risk inferences on unfamiliar chemical structures. Finally, a batch-based monitoring procedure is introduced to detect data drift, providing an early warning system for cases where changes in chemical distributions may require model upgrades.

In summary, the main goal of this research is to develop a multifaceted framework for evaluating and analyzing ADME-Tox models and to provide a detailed comparison of how different classical ML algorithms manage uncertainty. In this way, this work aims to bridge the persistent gap between computational metrics and the demands of clinical translation, thereby offering a more reliable basis for the integration of AI in the early stages of drug discovery and development.

Chapter 2. Theoretical Background

2.1. Fundamentals of ADME-Tox in Drug Discovery

2.1.1. Historical Evolution

Some chemoinformatics applications have long history of over 70 years. Although they are relatively old, they have evolved significantly over and are currently employed in the field of predictive toxicology to assess drug safety and drug efficacy. The first structure searching program was reported by Ray and Kirsch in 1957, laying the groundwork for chemoinformatics. About a decade later, in 1962, Hansch and his coworkers introduced for the first time the Quantitative Structure-Activity Relationships (QSAR) as a formal quantitative relationship between the physicochemical properties of molecules and their biological activity [12]. For many years, drug synthesis and design were based on experience, the judgment of the scientist involved in the synthesis and the availability of raw materials. However, in 1972, John G. Topliss changed this by introducing the Topliss approach, an operational scheme based on QSAR but more oriented towards a decision-making procedures. This non-mathematical, functional decision-making model concerning the order in which substituents can be tested in a complex compound, enabled rapid exploration of candidate drugs, reducing the number of compounds that needed to be synthesized and tested. This model has brought the way to modern high performance prediction models that are in the ADME-Tox field [13]. Other applications, such as the distribution of molecules through the biological membranes were described for the first time by Timmermans and coworkers in1977 [12]. In addition, in 1997, Christopher Lipinski proposed in a seminal work the “Rule of Five” correlating molecular weight, hydrogen bond donors and acceptors with drug oral bioavailability. This approach had a great impact on the pharmaceutical industry by increasing interest in absorption and solubility predictability [14]. The discovery of High-Throughput Screening (HTS) technologies at the beginning of 21st century produced large amounts of biological data by exploiting both old and new technologies [12]. Classical statistical models are used to deal with the high number of variables involved in molecular data and to correlate structural features of the molecules with the biological activity of a compound. Therefore,

the use of computational intelligence techniques allowed for drug design has enabled a more holistic approach to drug design by considering all aspects of drug behavior in humans [15]. The drug design field is currently evolving to the ADME-Tox field of research, that deals with modeling absorption, distribution, metabolism, excretion and toxicity of drugs to assess the pharmacokinetic behavior of drugs in humans and to reduce the number of drugs that fail in clinical trials [2].

2.1.2. The ADME-Tox Journey

"A pharmacokinetic study is a fundamental component of the drug development process, in which the pharmacokinetic parameters of a test molecule are evaluated to determine the way it behaves on human body" [16]. ADME was proposed by the American pharmaceutical scientist Nelson in 1961. However, the term for Absorption, Distribution, Metabolism and Excretion is rooted back to the 1930's paper by Teorell and to a 1919 paper by Widmark [17]. Today, the addition of T or Tox for Toxicity in the acronym is also included in its definition. In 40% of the cases of candidate drugs that fail during the preclinical or clinical trials, adverse ADME-Tox properties are mainly responsible. In addition, toxicities are responsible for 30% of drug recalls [5]. The study of pharmacokinetics describes what the body does to the drug (absorption, distribution, metabolism, excretion), whereas pharmacodynamics describes what the drug does to the body (mechanism and magnitude of pharmacological effect).Before providing a detailed description of each process, it is important to describe the general stages through which a drug makes its way through the body, as illustrated in Figure 1 [16].

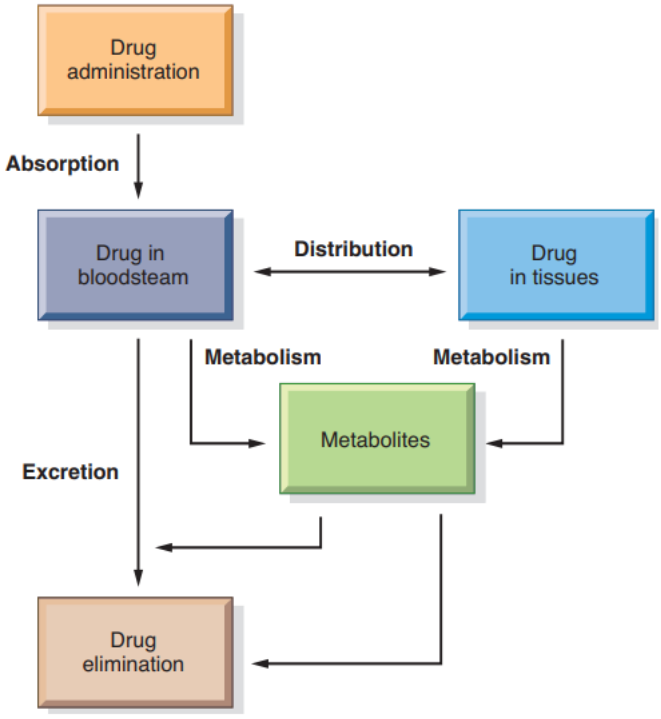


Figure 1. *Representation of the four pathways of drug movement and modification in the body*

As shown in Figure 1, the ADME processes follow the drug's administration and absorption of a drug into the body and describe how the drug is distributed throughout the organism and subsequently metabolized and excreted from the body. The first two processes, absorption and distribution, refer to the drug's entry into the body and its transport to tissues and organs through the systemic circulation. Another dimension of drug interaction is how the drug is altered chemically during the period before it is eliminated. Primarily the liver is responsible for this chemical alteration, or metabolism. The four processes that are involved in the complete picture of drug action are absorption of the drug from its site of administration, distribution of the drug throughout the body tissues, metabolism of the drug and finally excretion of the drug [16], [17].

Building upon the foundational understanding of the ADME framework, the following paragraphs provide an extensive analysis of the biological, chemical and mathematical details of each process, and cover toxicology issues.

Absorption:

The first step that a chemical takes on its path to a biological effect in the body is the process of being absorbed into the body. "Absorption refers to the transfer of a toxic chemical from its point and method of contact with humans into the bloodstream." There

are four primary routes of administration: ingestion through the digestive tract, inhalation via the respiratory system, dermal application to the skin or eye and injection through direct administration into the bloodstream [18]. For oral medications, it is necessary to have the drug product go through the process of dissolution, which typically requires a couple of minutes. This involves the transformation of a solid tablet or capsule into a solution, where most drugs require the gastric or intestinal fluids to dissolve. The dissolution of a drug into a solution can often be the rate limiting step to its eventual uptake into the circulation. In general, the ease of dissolution of drugs is dependent upon physical characteristics of the chemical. Highly soluble compounds, regardless of their lipophilicity, dissolve easily and quickly in aqueous solutions. In contrast, poorly soluble, highly lipophilic drugs do not dissolve well in aqueous body fluids of the gastrointestinal tract. Following dissolution, the next step in drug absorption is the passage of the drug through the selectively permeable membranes of gastrointestinal tract. Most drugs currently available are absorbed via simple passive transmembrane diffusion. Therefore, in the gastrointestinal tract drug is absorbed by simple diffusion from high concentration in the intestinal lumen to lower concentration in the blood circulated through the capillaries of the intestinal wall (Figure 2) [16].

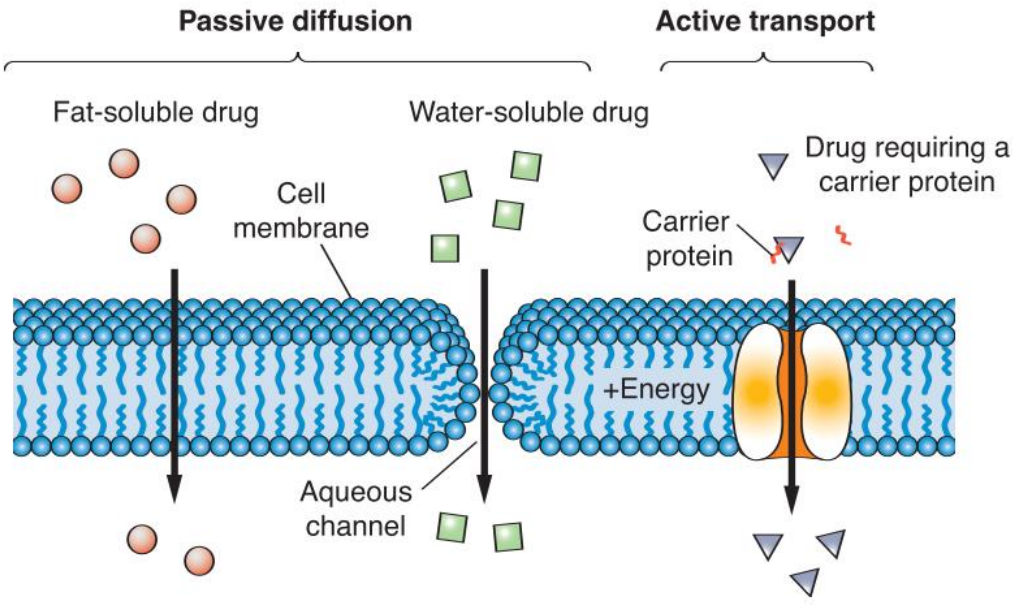


Figure 2. *Mechanisms of drug absorption through the epithelium of the gastrointestinal tract*

Migration of lipids into the cells is by movement down a concentration gradient to approximation to a thermodynamic equilibrium. This is different from transport of drugs similar in structure to the natural lipids which can occur by carrier mediated transport.

Unlike diffusion, this is a saturable process limited by the finite availability of transporter proteins, and it can move molecules against a gradient using cellular energy (ATP) [16].

Bioavailability (F) of a pharmaceutical substance in the pharmaceutical sciences is defined as "the fraction of an administered dose of the drug that reaches systemic circulation in its active form". This parameter can be determined quantitatively using the Area Under the Curve (AUC) of the concentration in serum vs time for both the oral and intravenous (IV) administration. The mathematical relationship is expressed as follows:

$$(1). \quad Bioavailability = \frac{AUC\ oral}{AUC\ injected} * \frac{Dose_{injected}}{Dose_{oral}}$$

This calculation, as shown in Figure 3 through the sources [16]highlights that the oral bioavailability of a drug is rarely 100% due to first-pass metabolism. Most of an orally administered dose is absorbed from the gastrointestinal tract into the blood, and it is then rapidly transported to the liver through the portal circulation, where the drug undergoes extensive metabolism. Other factors which might be expected to affect absorption include the perfusion rate of the gut, the gastric transit time and the large intestinal transit time. Many drugs are affected by physiological changes, such as achlorhydria which is common observed in elderly [16].

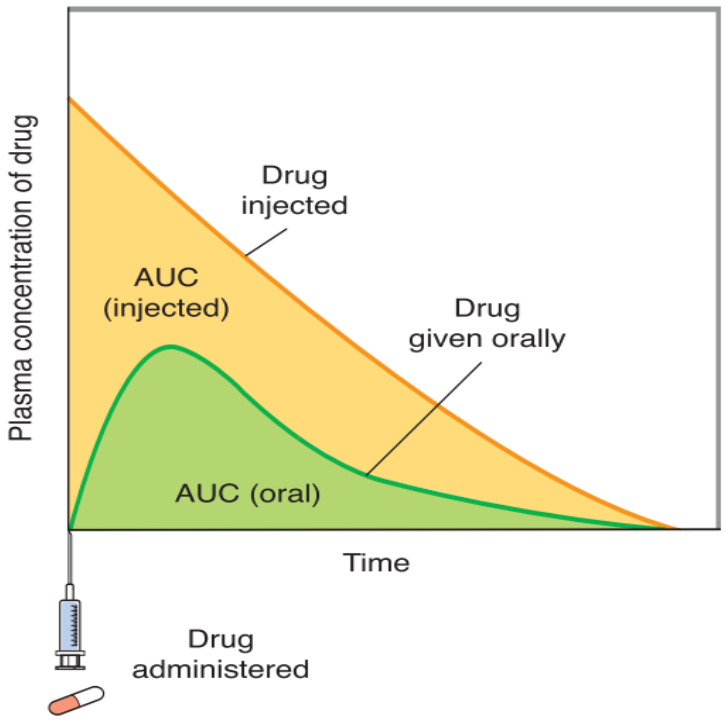


Figure 3. *Comparison of plasma concentration-time curves for oral and intravenous drug administration*

Predicting complex biological activity as early as possible in the drug discovery process has become an important function of the medicinal chemist. In addition to computer modeling of active site interactions, alerts such as the Rule of 5 can be used to predict properties like poor absorption or permeation [14]. The Rule of 5 states that such poor properties are highly likely if a molecule violates two or more of the following: Molecular Weight (MWT)>500, Log P >5, more than 5 hydrogen-bond donors or more than 10 hydrogen-bond acceptors. For many years, lipophilicity has been known to be an important property for membrane passage, but interestingly, high MWT and high LogP are generally inversely correlated with oral activity in successful drugs [14]. Modern platforms increasingly use these filters alongside machine learning models to classify Human Intestinal Absorption potential [14]. Once a molecule is successfully absorbed and enters the bloodstream, the Distribution phase begins. During this stage, the drug is carried throughout the body and dispersed into all tissues, including those where the compound will interact with its intended intracellular therapeutic target [16].

Distribution:

"Distribution is the reversible process by which a drug is transferred from sites of initial circulation to other compartments and tissues of the body including plasma, interstitial fluid and inside cells" [10]. The rate of transfer from circulation to tissues is influenced by several factors such as perfusion of various organs, capillary permeability, and tissue and plasma binding of drug [14]. Tissues that are highly perfused, such as the liver and kidneys, reach the equilibrium more rapidly than tissues that are poorly perfused. Fatty tissues, skin and bone achieve a steady state slowly. The capillary bed of the tissue is another important factor in the distribution of drugs [10]. The blood brain barrier, for example, is formed by a tight grouping of endothelial cells lining capillaries that function to restrict the transfer of most substances to the brain and spinal cord to protect these tissues from neurotoxic drugs [16] (Figure 4). Unlike the small slit junctions found between vascular endothelial cells, the spaces between liver capillary cells, known as sinusoids, are arranged to allow large molecules to be freely exchanged for metabolism [10].

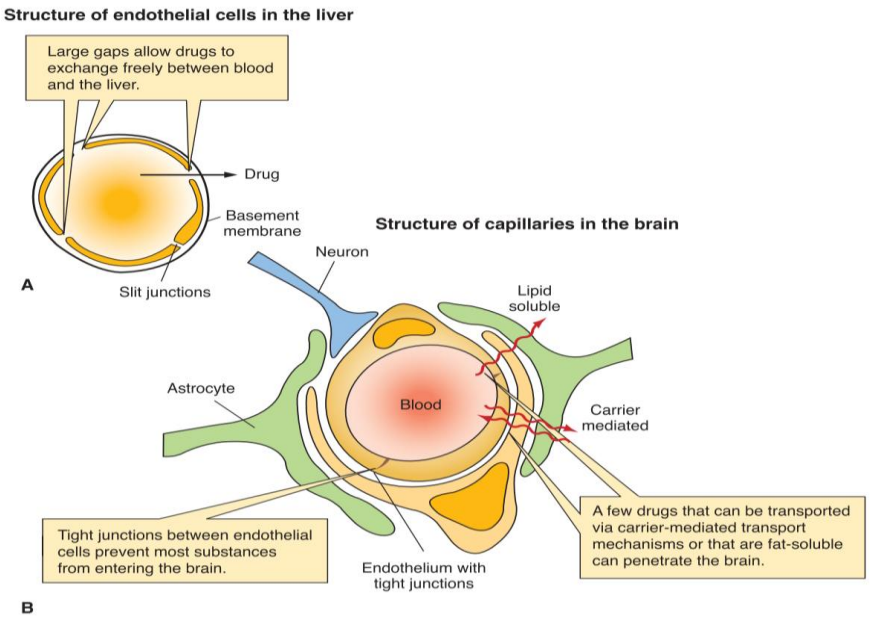


Figure 4. *Anatomical comparison of brain capillaries vs liver capillaries showing restricted vs free molecular exchange*

The distribution of a drug in the body is largely determined by its physicochemical properties, in particular lipophilicity (LogP) and molecular weight (MWT). High MWT and overly lipophilic molecules (LogP>5) generally do not distribute well to the sites and activities of pharmaceutical interest. Ideally a drug must be sufficiently soluble in nonpolar media and not overly sequestered in nonspecific, lipid environments [14]. Topliss (1972) provided an operational strategy for designing optimum substituents on a parent structure. By replacing a hydrogen atom with substituents, such as 4-chloro or 4-methoxy, the overall hydrophobic (π) and electronic (σ) effects can be optimized to improve penetration of biological barriers, or to reduce the tendency for the drug to bind to nonspecific sites [13]. In recent years, the structure based approach has increasingly been combined with insights obtained from large datasets of active compounds within Medicinal Chemistry databases using Machine Learning (ML) models. Predicting penetration across the blood-brain barrier (BBB) and binding to P-glycoprotein (P-gp) represents a major challenge which in silico approaches try to address. Machine learning models like Random Forest (RF) and neural network models such as Graph Convolutional Networks (GCNs) reached up to 80% prediction accuracy and virtual screening of large compound libraries before synthesis is of huge advantage [10] [9].

The extent of distribution of a drug is expressed as the Apparent Volume of Distribution (V_d). V_d is defined as the volume that would contain a given amount of drug, assuming that the concentration of the drug in this volume is equal to the concentration of the drug

in the plasma [16]. The Apparent Volume of Distribution is calculated using the following equation:

$$(2). \quad Vd = \frac{Total\ amount\ of\ drug\ in\ the\ body}{Plasma\ drug\ concentration\ (C_p)}$$

V_d of a drug in the body is a very important pharmacokinetic parameter, which is significantly affected by the Plasma Protein Binding. The "free drug hypothesis" describes that only the unbound fraction of a drug is active and can permeate biological membranes to interact with intracellular targets [14] [10]. Drugs that are highly protein bound will have a small V_d and remain predominantly in the plasma water, whereas drugs that are highly lipophilic will have a V_d greater than total body water and have tissue distribution. The change in V_d upon administration can be predicted by important physiological parameters like age and body weight. Therefore, dosing has to be done on an individual basis to avoid systemic toxicity [6].

Metabolism:

Once a chemical compound has arrived at its site of action, it is metabolized (biotransformed). This process serves the purpose of solubilizing lipophilic, nonpolar xenobiotics and making them amenable for urinary excretion by the kidneys through conversion into more polar metabolites. Metabolism is carried out in a large number of tissues, but the bulk of drug metabolism occurs in the liver where the necessary physical structure is provided, and the required number of biocatalysts are available [16]. Many drugs are even metabolized prior to first entering the general circulation after having been absorbed from the gastrointestinal tract, a process known as first-pass effect. Some drugs are designed as prodrugs in the first place, which requires metabolism into the active drug before any therapeutic effect can be observed [16].

The metabolic pathways of active pharmaceutical ingredients (APIs) often include interactions of many enzymes, and are typically considered to be predominantly oxidative reactions, which are frequently catalyzed by members of the Cytochrome P450 (CYP450) superfamily of enzymes, many of which exhibit saturation kinetics thereby limiting metabolic rates through so-called ceiling effects [16]. A good example of the importance of a single CYP enzyme in the metabolism of drugs is CYP3A4, which is responsible for almost 50% of all prescribed medications, in contrast CYP2D6 accounts for almost 25% via modifications to other functional groups [10]. This knowledge is routinely considered

in the process of lead optimization using tools such as the Topliss operational schemes. The modification of a single hydrogen atom to a fluorine atom, as seen in many drugs with an aromatic ring, results in the suppression of a rapid, first pass metabolizing 4-hydroxylation reaction, producing a metabolically stable, longer duration of action analogue [13].

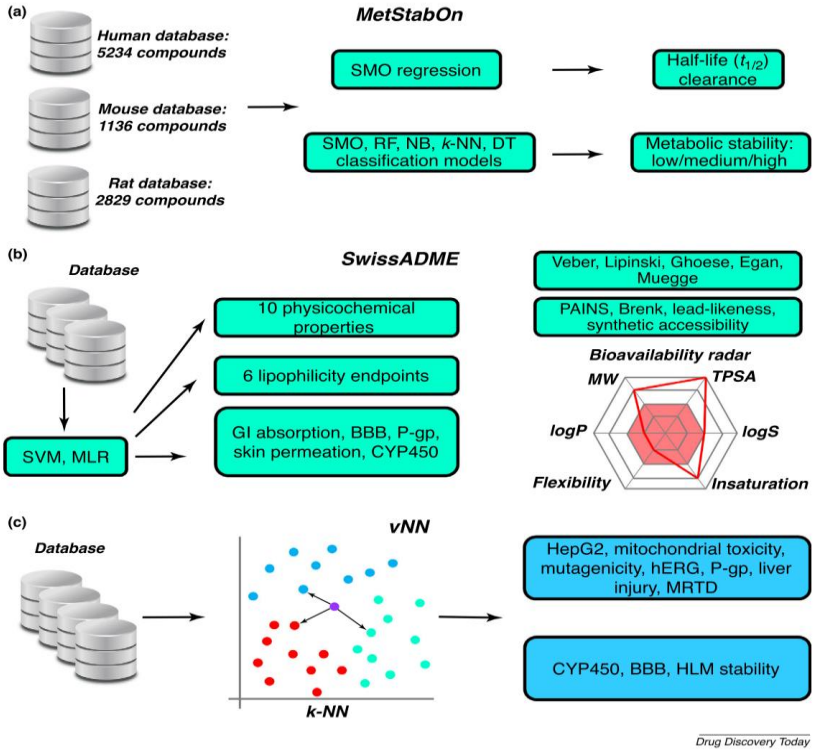


Figure 5. *This visual details recently implemented web servers (MetStabOn and SwissADME), which utilize ML models for prediction of metabolic stability (by HLM) and CYP450 inhibition. Stability predictions are categorized as "Low", "Medium" and "High" for each individual compound and are obtained using molecular descriptors [12]*

The efficiency of drug elimination can be expressed by the elimination rate, expressed as clearance (CL) in mL/min, which corresponds to the plasma volume cleared of the drug in a certain period of time. Clearance encompasses all processes such as glomerular filtration, metabolic transformation by the liver, distribution into tissues and nonspecific binding, that affect the concentration of the substance in the body. The mathematical formula for clearance is:

$$(3). \quad CL_{total} = CL_{renal} + CL_{hepatic} + CL_{other}$$

,where CL_{total}**:** The volume of plasma completely cleared of a drug per unit of time (e.g., L/h or mL/min), CL_{renal}**:** Removal of the drug through the kidneys into urine, $CL_{hepatic}$: Metabolism of the drug by liver enzymes or excretion via bile and CL_{other}**:** Minor

elimination routes, which may include pulmonary excretion (lungs), sweat, saliva, or breast milk.

Half-life ($t_{1/2}$) and other pharmacokinetic parameters are important drug characteristics, that have traditionally been determined experimentally. However, with the advent of high-throughput computing, in silico predictions have become a relevant alternative in drug discovery. Databases such as the Therapeutics Data Commons (TDC) provide extensive datasets, including experimental half-life and clearance values for over 12000 drugs [10]. Architectures such as GCNs and XGBoost demonstrate high accuracy in identifying molecules that act as inhibitors of specific CYP450 isoforms, thereby preventing adverse drug interactions [12].

Human specific metabolism has become tractable to bioengineering studies. While traditional rodent models are insufficient for metabolism prediction, liver on a chip and organoid models have been increasingly used in industry and academia to improve in vitro toxicity studies [19]. 3D human liver cell culture models that include hepatocytes and non-parenchymal cells recapitulate the complexity of human liver tissue, making them suitable for preclinical toxicity studies [3].

Excretion:

“Excretion is the final stage of the pharmacokinetic process. This is where the body gets rid of the drugs and their active metabolites. It is important to note that elimination and excretion are often confused and are frequently used interchangeably in the literature. However, they have very different meanings within the sphere of pharmaceutical analysis” [16]. Elimination refers to “the decrease in concentration of a drug at the site of measurement”, whereas excretion refers to “the irreversible removal of the chemically unchanged drug from the body, i.e. the majority of a drug is excreted into the urine via the kidneys” [17]. Small amounts may be excreted via the biliary system and volatizes such as anesthetic vapors can be exhaled from the lungs [16]. Other routes of excretion are found in very minor quantities, for example via sweat, saliva and in the case of women who are breastfeeding into milk. The kidneys play a major role in drug excretion, and the process occurs in the nephron, a functional unit of the kidney [16]. Blood enters the glomerulus, a section of capillary where water and solutes that are not large enough to be retained in the solution by proteins such as albumin can pass through the capillary membrane into the center of the nephron, becoming part of the filtrate [16]. The filtrate

then moves through the nephron where some substances are reabsorbed into the bloodstream while others are excreted into the urine [16]. Factors such as the solubility of a drug in water and the properties of a drug's metabolites determine their rate of removal from the body. Very polar and water soluble compounds are removed from the body within hours, lipid soluble compounds will tend to re-enter membranes within the body before excretion from the kidneys and may be reabsorbed into the bloodstream [18].

Clearance (CL) together with half-life ($t_{1/2}$) of a drug (i.e., time required for its concentration to decrease to half of the administered dose) in the body, dose intervals (i.e., time between drug administration) need to be determined. From a clinical perspective, it is always desirable to attain the Steady State as early as possible, and the rate of drug input equals the rate of drug output and forms a plateau or constant concentration in the body [16]. As illustrated in figure 6, the decreased function of excretory organs with age, or as a result of renal disease, for example, increases half-life of a drug, thus increasing the risk of toxicity. Therefore, a critical assessment of dose is important [16].

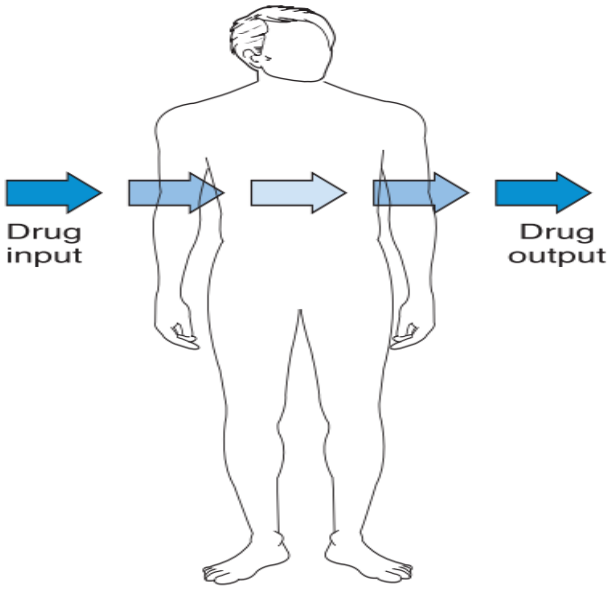


Figure 6. *Drug equilibrium at steady state, where drug input equals drug elimination [16]*

Toxicity:

Toxicity is "the final piece of the ADME-Tox puzzle that predicts the clinical success and market value of lead compounds" [2]. Toxicology is "the branch of science that assesses the adverse or toxic effects (injury or death) that substances, forces, products, etc. may produce on living organisms" [5]. Early toxicity prediction is a key component of drug

development process. Statistically it has been observed that approximately 30 or 40% of preclinical candidates fail during development for "unanticipated toxicity" [5]. Furthermore, adverse drug reactions have been the most common cause of removal of pharmaceuticals from the market over the last few decades [2]. Toxicological assessment can cover a variety of endpoints. These are grouped into several categories [2]:

1. Acute toxicity: refers to the toxicity aspect that concerns the effects of toxic substances on a living Organism after a short period of exposure to high doses of toxic substances. Acute toxicity is mostly expressed as the lethal dose for 50% of test animals. It is commonly reported as LD50 for oral or dermal exposure (mg/kg) and LC50 for inhalation exposure (mg/L or ppm) [2].
2. Organ specific toxicity: includes Hepatotoxicity (DILI), elevated transaminases (ALT/AST), Nephrotoxicity (elevated levels of serum creatinine and BUN), and Cardiotoxicity (hERG channel inhibition and risk of fatal arrhythmias) [2].
3. Chronic and genetic toxicity: concerns Carcinogenicity, Genotoxicity (DNA damage), and reproductive toxicity, which may lead to permanent developmental abnormalities or endocrine disruption [20].

Safety assessment has traditionally relied on in vivo animal experiments and in vitro bioassays (Figure 7).

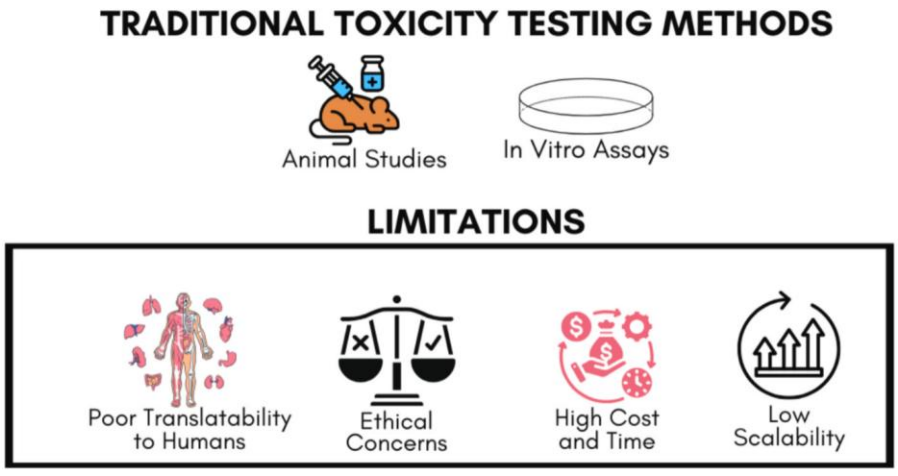


Figure 7. *Limitations of traditional toxicity testing methods [19]*

Such approaches are, however, expensive and time consuming. There is a translational gap due to species specific differences in biology [19]. Notably, more than 90% of compounds assessed for safety in animal tests for human use failed when tested clinically [3]. Due to the high costs involved and the increasingly stringent adherence to the

replacement, reduction and refinement of animal use in scientific research, there is a rapidly emerging field of computational toxicology [19].

AI and ML technologies have started to revolutionize toxicology, transforming art into a data-driven science. Deep learning networks such as graph neural networks (GCNs) and Transformers are capable of predicting toxicity and inferring structure activity relationships that surpass the current state-of-the-art [2]. Additionally, innovative Microphysiological systems such as the liver-on-a-chip and organoids provide alternative human relevant systems and models to study toxicities reducing their reliance on animal models [3]. This virtual issue offers a first snapshot of current developments and capabilities in the field.

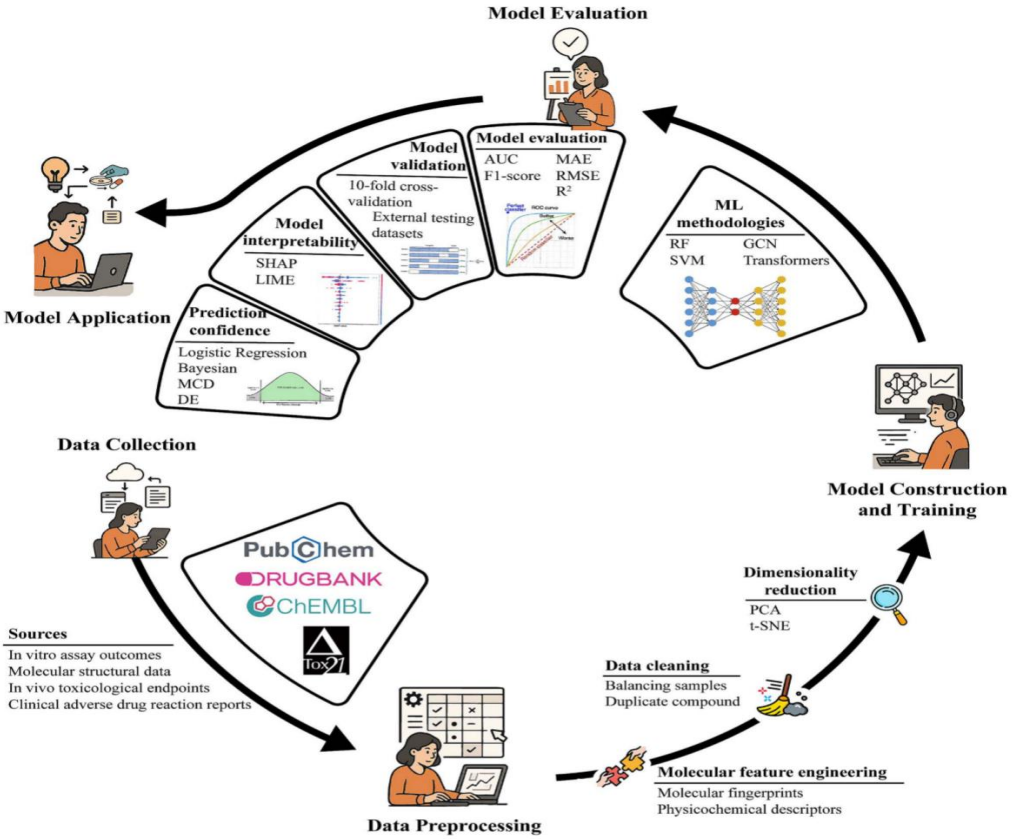


Figure 8. *Workflow for AI/machine learning–based drug toxicity prediction [2]*

2.2. Machine Learning Architectures

2.2.1. Logistic Regression

Logistic Regression (LR), as a linear model, is the basic model for many tasks in computational toxicology and pharmacokinetic modelling. Although numerous deep learning models have been developed and reported, in the context of an ADME-Tox research workflow, LR is an essential component due to its balance of simplicity and interpretability. As a regular linear model, ordinary least squares (OLS) regression is typically used for predicting a continuous outcome, however, for classification tasks, OLS is not sufficient [20]. Linear models typically generate predictions outside of the range of the outcome variable of interest. Moreover, for many tasks, it is preferred to obtain a probability of classification (toxic (1) or non-toxic (0)) [21]. For low-risk compounds, negative predictions are not meaningful, and for high-risk compounds, predictions greater than 1 are not relevant. Although negative values or values greater than 1 are not impermissible for linear models, for probability predictions, these values are not meaningful [19]. Therefore, a logistic (sigmoid) function is applied to the regular linear model to map any real value of the input to a value between 0 and 1, i.e., a value that is meaningful for probability predictions of the likelihood of a toxicological effect given a set of molecular fingerprints [21].

A heuristic model for predicting toxicity is presented. The model is based on a logit transformation of the logarithm of the Odds Ratio (OD) of a property. An example is given using the property of hepatotoxicity, where the logit of the probability of hepatotoxicity (p) is modelled as a function of molecular properties (X) [22]:

$$(4). \quad p(X) = \frac{e^{\beta_0+\beta_1 X_1}}{1+e^{\beta_0+\beta_1 X_1}}$$

The linear form of the model is expressed through the log-odds (logit):

$$(5). \quad logit(X) = ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1$$

It is important to note that whereas linear regression involves minimizing the mean squared difference between predicted and observed outcomes (least squares) to find the coefficients (β), LR uses Maximum Likelihood Estimation (MLE) instead [23]. In MLE, the objective is to estimate the model parameters that most probably predict the observed (training set) outcomes toxic or non-toxic [23]. The likelihood function (L) encodes the probability of the observed data given unknown parameters. The log-Likelihood for LR is [23]:

$$(6). \quad L(\beta) = \sum_{i=1}^{n}[y_i log p_i + (1-y_i)\, log(1-p_i)]$$

Due to the nonlinear relationship between the score equations and the parameters β, a closed form solution is not possible, and maximization must be performed numerically. Typically, the Newton-Raphson method is used to update the β values in each iteration until the maximum likelihood is reached [6].

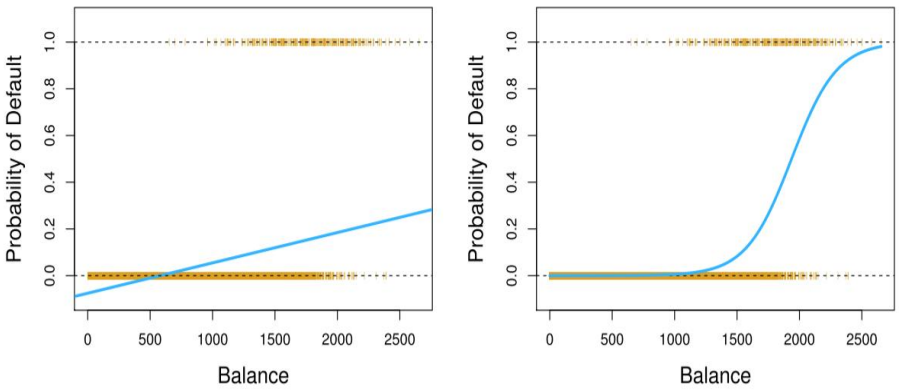


Figure 9. *Comparison of Linear (left) and Logistic (right) Regression fits on binary data. Linear regression tends to assign negative probabilities to individuals of low risk, whereas LR fit produces an S-shaped sigmoid that satisfies physical constraints and never yields a probability less than 0 or greater than 1 [23]*

LR provides transparent predictions of missing ADME-Tox properties [20]. Each coefficient can be transformed into Odds Ratio (OR) by exponentiation (e^{β_i}) and easily interpreted in terms of influence of variations of different physicochemical properties on toxicological risk for toxicologies like Drug-Induced Liver Injury (DILI) or hERG channel inhibition [23]. By comparing the LR-based "white box" approach to deep learning-based "black box" models structural alerts for severe DILI or hERG channel inhibition can be identified [21].

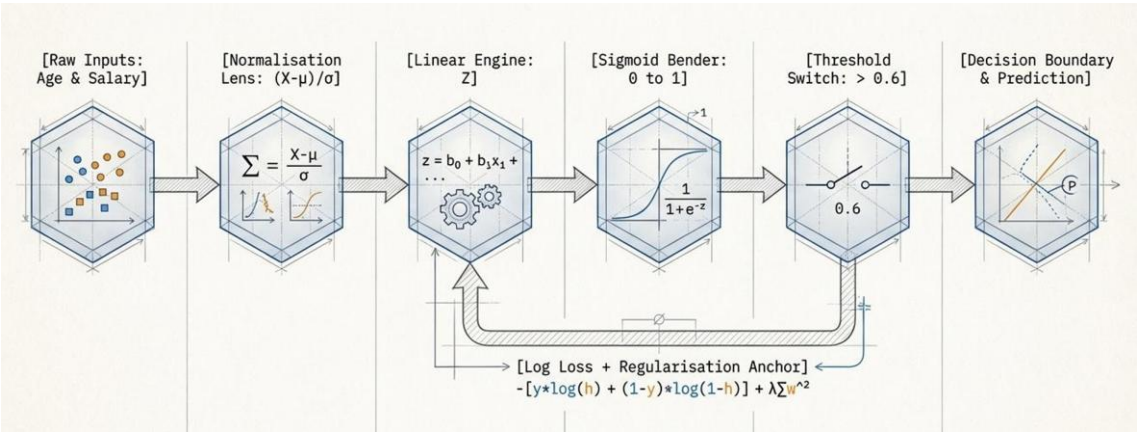


Figure 10. *The complete LR architecture [21]. The architectural flow of LR shows the transformation of raw molecular inputs through a "linear engine" and a "sigmoid bender", resulting in a final probabilistic output that can be used for binary decision-making*

As a reference method for the comparison of new approaches and as a basis for the assessment of their quality, the LR model serves as an effective baseline model[20]. Using large numbers of benchmark datasets, it is demonstrated that simple LR can reach high performance levels for toxicological prediction tasks [15]. It is robust for small training set sizes and imbalanced classes, thus it can serve as a prefilter to select promising compounds that can later be analyzed with more time consuming methods [20].

Importantly, LR is inexpensive to compute, and overfitting is mitigated, especially when using L1 and L2 regularization (Lasso and Ridge) [20]. Therefore, it is the most suitable method to use as a control for comparative performance studies. Moreover, its application to high dimensional molecular fingerprints (p>>N) enables the use of lasso regularization for feature selection [6]. Also, it is possible to use analytical methods to determine the most important chemical substructures, fragments or features responsible for adverse effects, such as carcinogenicity or genotoxicity [2].

Simple yet robust and interpretable, LR is an essential tool in pharmaceutical research and provides a solid basis for benchmarking studies aiming to perform toxicological predictions in big data environments [12].

2.2.2. Support Vector Machines (SVM)

Support Vector Machines (SVM) have become “one of the most important and widely used supervised learning methods in computational toxicology” [24]. First implemented by Vapnik and his team at AT&T Bell Laboratories in the mid-1990s for binary classification tasks, recently successful applications have been reported for regression as well as multi-class classification problems [15] . Within the ADME-Tox framework SVMs are used for toxic and non-toxic classification as well as for prediction of several important physicochemical properties [10].

SVM are learned, so-called, by finding the best hyperplane in the p dimensional space of the observed variables to separate the two classes of the target variable [6]. A hyperplane is defined by the equation:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p = 0 \quad (7).$$

, where β are the coefficients of the hyperplane and determine the orientation of the plane within the chemical descriptor space [6].

The maximal margin classifier is a classification method which attempts to identify the hyperplane with the largest possible distance to the training data of each class. This distance is called the margin of the classifier [20]. The points on the decision boundary are called support vectors. It has been theoretically proven that large margins on the training set lead to a superior generalization performance and high accuracy on unseen data [20].

While the linear support vector machine is guaranteed to perfectly classify linearly separable data, real world problems like predicting the effective dosage of a drug for treating cancer are inherently noisy [20]. The soft margin generalizes by allowing vectors to be on the wrong side of the margin or even on the wrong side of the hyperplane altogether, and by incorporating slack variables for its training instance. The optimization process is controlled by the tuning parameter C, which defines the "budget" for violation. For large values of C, misclassifications are penalized more heavily, typically resulting in a narrower margin and lower tolerance to violations. In contrast, smaller values of C allow a wider margin and greater tolerance to margin violations, which can improve generalization in noisy datasets. [20].

To address toxicological problems of nonlinear relations SVMs transform the data into a higher dimensional space by using kernel functions, where the relations can be solved by linear separation [5]. In computational pharmacology, all calculations are performed in the original space of the objects of study [15]. Only the inner products of the vectors are computed without explicit calculation of the coordinates in the extended feature space [6]. The most commonly used kernels include:

- Linear kernel: Suitable for scenarios with a very high number of descriptors [15].
- Radial basis function (RBF/Gaussian): the most commonly used kernel which is powerful enough to model very complex relationships [15].
- Tanimoto kernel: a specialized kernel designed for comparing molecular fingerprints based on structural similarity [9].

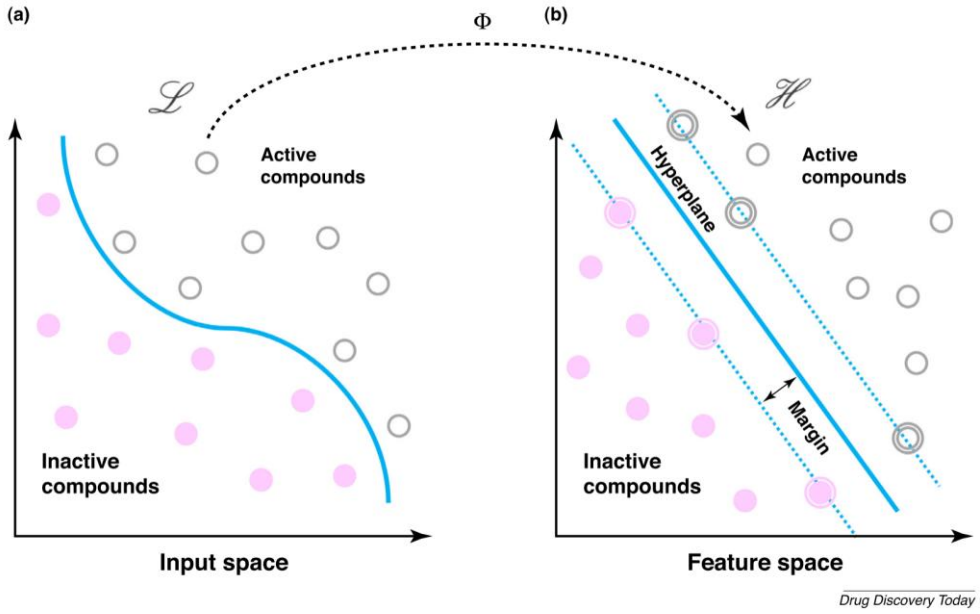


Figure 11. *Projection into high-dimensional feature space. Using mapping function F, active (empty gray points) and inactive (filled pink points) compounds that are not linearly separable in low-dimensional input space L (a) are projected into high-dimensional feature space * (b) and separated by the maximum-margin hyperplane. Points intercepted by the dotted line are called 'support vectors' (circled points) [15]*

Although the SVM is a binary classifier that only distinguishes between two predefined labels (0 and 1), in the context of ADME-Tox, risk assessment 1 is more interested in knowing the probability of the material being toxic [2]. In order to obtain this, the output of the SVM classification can be converted to probability using a process known as Platt Scaling [11]. This involves a sigmoid function which transforms the SVM score into a probability [11]:

$$(8). \quad P(y = 1|f) = 1/(1 + exp\,(Af + B))$$

,where:

y: The true class of a sample, in binary classification this is equal to 1 or 0.

$P(y = 1|f)$: The posterior probability that a new sample belongs to the positive class (e.g. Toxic) given the output of the learning algorithm, f.

f: the output of the SVM algorithm (the decision value) before the thresholding. This is the signed distance of the sample in feature space to the separating hyperplane. In other words, it is the margin distance of the opposite side of the separating hyperplane from the training samples.

A and B: These are two tunable parameters and can be optimized to fit experimental data using a validation set and a procedure for Maximum Likelihood Estimation.

The raw output of the SVM is passed through a sigmoid function. The reason for this is that the raw output of an SVM has an unlimited range of -∞ to +∞ and encodes distance to decision boundary information which is not calibrated to any absolute scale and hence yields non quantitative probabilities which are not important for risk assessment in toxicological studies and so are mapped to the range [0,1] [19].

SVMs have achieved exceptional performance across a wide range of tasks and are competitive to Artificial Neural Networks and Random Forests. Especially when training data are limited, SVMs often yield the best results. Some significant applications of SVMs include [12]:

- Cardiotoxicity (hERG): The SVM model developed by Ylipää et. al. achieved an AUC ROC of 0.95 on a dataset of 291.219 compounds, representing the most reliable potassium channel inhibition predictor reported so far [19].
- Blood-Brain Barrier Permeability: Studies have shown that compounds can be classified as CNS-active or inactive with greater than 80% accuracy [12].
- Hepatotoxicity (DILI) and Carcinogenicity: SVM serves as a core algorithm in platforms such as ADMETlab 3.0 and SwissADME for Predicting liver injury and carcinogenic potential, offering robust results on benchmarks like the TDC (Therapeutic Data Commons) [12].
- Metabolic Stability: Classification of compounds into stable or unstable compounds within liver microsomes with accuracy greater than 85% [12].

SVMs offer much robustness against the “curse of dimensionality” since the complexity of the SVM model does not significantly increase with the number of predictors (features) if the number of support vectors remains small. In addition, the SVM optimization problem is stated as a convex optimization problem which means that a global minimum of the function is guaranteed to be found and not a local minimum as opposed to neural networks [24]. However, unlike standard LR, little can be said about the variables that contribute to the prediction of toxicity [24]. The choice of the kernel function as well as the parameters (C and γ) requires extensive cross validation studies [2].

2.2.3. Gradient Boosted Decision Trees (XGBoost)

“XGBoost (Extreme Gradient Boosting), an extremely fast, scalable and accurate implementation of Gradient Boosting algorithms has gained immense popularity in recent

years" [8]. The improvement over traditional GBM algorithm lies in the enhanced tree based modeling approach, known as Greedy Function Approximation, proposed by Friedman in 2001 [25]. XGBoost Is one of the most powerful tools to handle complex, structured chemical data in computational toxicology and it has achieved state-of-the-art performance in predicting several toxicity endpoints, including hepatotoxicity and nephrotoxicity [9].

Unlike training model from start to finish, XGBoost follows on additive approach. Given n examples and m features per example, XGBoost builds an additive ensemble of (K) trees sequentially. At each boosting iteration, a new tree is fitted to the current residual errors (or gradients) of the training set, and predictions are updated as the sum of all trees [8]:

$$\widehat{y_i} = \sum_{k=1}^{K} f_k(x_i), f_k \in F \quad (9).$$

, where F is the space of all possible regression trees. Each regression tree f_k is defined by a structure q and a leaf weight w. The structure q encodes for its compound identifier a leaf node it is mapped to. The leaf weight w states the regression value for this leaf node [8].

XGBoost optimizes the given data and output combinations by attempting to minimize a Regularized Objective function [8]:

$$L(\varphi) = \sum_i l(\widehat{y_i}, y_i) + \sum_k \Omega(f_k) \quad (10).$$

The term Ω(f) represents the model complexity penalty, acting as a regularizer that controls the trade-off between fitting the training data and maintaining model generalizability. In this function, l is a differentiable convex loss function which measures the error between predicted toxicity and the corresponding toxicological outcome [8].

Ω represents the model complexity, defined as [8]:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2 \quad (11).$$

,where T is the number of leaves and λ is L2 regularization parameter. This is crucial in preventing overfitting in the small ADME datasets and at the same time keeping the models simple and more generalizable [8].

In XGBoost, an innovative technique is incorporated into the tree based boosting algorithm, by expanding the loss function using a second order Taylor expansion.

Contrary to standard methods that utilize gradient boosting which relies on the first derivative of the loss function (the gradient), XGBoost also considers the second moment of the gradient, specifically the first order g_i and the second order h_i gradient statistics on the loss function [8]:

$$(12). \quad g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$$

and

$$(13). \quad h_i = \partial^2{}_{\hat{y}^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$$

For each leaf j, we compute the optimal weight $w^*{}_j$ using the second derivative. The quality of a tree structure is measured by the so-called Structure Score, also referred to as Gain, of the best split point that results in the largest decrease in loss [8].

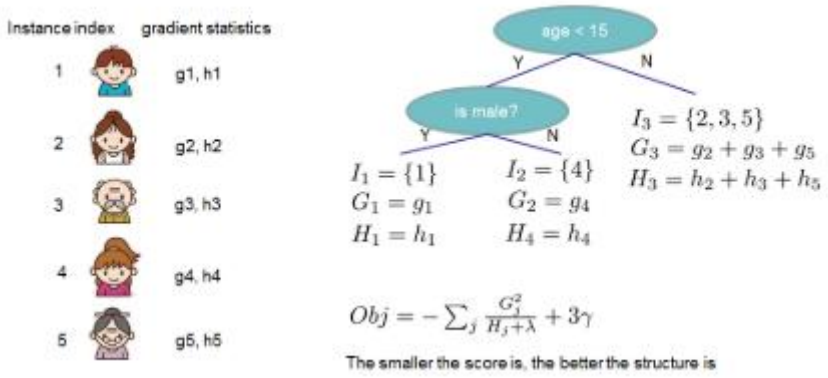


Figure 12. *XGBoost uses the structure score, which is a function of both g_i and h_i, to select the best split for chemical features, and the score at each leaf node is the sum of g_i and h_i [8]*

Sparsity-Aware Learning real-world drug design problems encounter "sparse" data where a binary molecular fingerprint describes a molecule with many zeros. For efficient split finding, XGBoost implemented a Sparsity-aware Split Finding algorithm. This technique typically boosts the performance by up to 50 times compared to standard methods due to the "default direction" which XGBoost learns for each node [8].

XGBoost performance has been demonstrated across many different types of problems beyond regression and classification, including many of the ADME-Tox prediction tasks in MoleculeNet [9]. Examples of applications in ADME-Tox prediction of XGBoost:

- Physicochemical Properties: For the tasks of predicting aqueous solubility (ESOL) and hydration free energy (FreeSolv), XGBoost achieved an RMSE of

0.99 and 1.74, respectively, matching or exceeding the performance of ab initio methods in some cases [9].

- Toxicological Assessment: XGBoost performed well for toxicity assessment. In the study by Zhang et al. 2025, XGBoost was shown to be the central algorithm of the ADMETboost platform, demonstrating robust performance for DILI and CYP450 predictions [2].

Overall, XGBoost is an optimized solution based on using the second-order optimization techniques together with novel optimization methods for regularization and efficient information transfer. This solution makes it possible to process datasets with millions of substances consisting of only non-standard fragments, using standard IT infrastructure. Many current problems in the pharmaceutical industry, particularly in terms of high-throughput virtual screening, require such a solution [8].

2.3. Reliability and Uncertainty Quantification

2.3.1. The Calibration Gap

The Calibration Gap in machine learning remains a major challenge in state-of-the-art applications, including high stakes domains such as pharmaceutical toxicology and clinical diagnostics [26]. While current models strive to improve accuracy and discrimination, they often fail in providing reliable probabilities [27].

While classification accuracy is a key criterion for selecting models to be used in computational toxicology for distinguishing toxic from non-toxic chemicals, another important consideration is whether the probabilities assigned to classifications (i.e., prediction confidence) such as “toxic” (0.85) reflect actual empirical frequencies with which particular toxic effects occur [27]. The issue of how well predictions reflect true underlying risks Is addressed by the concept of calibration, which refers to the degree to which a model is internally consistent (i.e., well calibrated) and accurately conveys the true risk of an effect by estimating the probability of observed empirical outcomes [26].

For a given classifier h, that classifies input X (molecular descriptors) into a class prediction Y, assigns a class probability P to this class, a perfectly calibrated model has the property that the predicted probability P equals the true probability that h correctly classifies an input X [11]:

$$(14). \quad P\left(\hat{Y} = Y \mid \hat{P} = p\right) = p, \ \forall p \in [0,1]$$

The ideal scenario (1) in practical applications shown above is not attainable due to limitations with the sample size. The Calibration Gap is defined as the difference between actual accuracy and mean confidence for particular ranges or bins of values [11].

The error of this prediction can then be quantified by the prediction error gap (ECE) which is the weighted average of the absolute difference between accuracy and confidence across all bins. The accuracy, acc(B_m) and the confidence conf(B_m) of the partition into the bins are defined as [11]:

$$(15). \quad acc(Bm) = \frac{1}{|B_m|} \sum_{i \in B_m} 1(\hat{y}_i = y_i)$$

,

$$(16). \quad conf(Bm) = \frac{1}{|Bm|} \sum_{i \in Bm} \hat{p}_i$$

and

$$(17). \quad ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} |acc(Bm) - conf(Bm)|$$

The calibration gap for each bin is defined as the absolute difference between the observed proportion of the study sample falling into the given category and the proportion of the population that actually belongs to that category: | acc(Bm) - $conf$(Bm) | [11].

State-of-the-art deep neural networks (CNNs) such as OxfordNet, Inception-ResNet-V2 and others achieve record classification accuracy on image recognition benchmark tasks like ImageNet (Guo et al. 2017) [11]. However, unlike LeNet [11] the recently developed models are less well calibrated. This is attributed to several factors:

- Model Capacity: Increasing the depth and width of such models leads to a decrease in training error but an increase in overconfidence as deeper models learn to reduce the Negative Log Likelihood (NLL) loss by increasing the output probabilities slightly [11].
- Batch Normalization: Training is stable, but calibration does not improve [11].
- Reduced Weight Decay: The modern deep learning of using less L2 regularization negatively impacts calibration. As increased regularization tends to improve the reliability of probabilities [11].

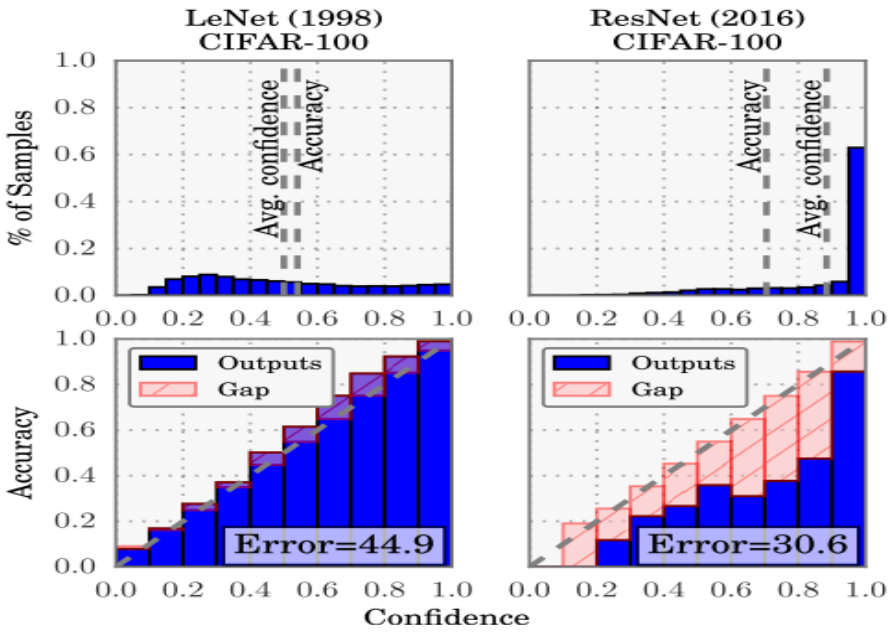


Figure 13. *Reliability Diagrams show that in modern models (e.g., ResNet), the accuracy bars fall significantly short of the diagonal (perfect calibration), indicating a substantial Calibration Gap where the model is "overconfident" in its incorrect predictions [11]*

Calibration is often considered the "Achille's heel" of prediction models in pharmaceuticals and ADME-Tox in general [26]. While an overestimation of the probability of Drug-Induced Liver Injury (DILI) for a compound may lead to non-selection of a potent drug (false negative in development), an underestimation of risk can have grave consequences for human health (serious health hazards during clinical trials) [26].

ADME-Tox prediction platforms such as ADMETboost that have been integrated with uncertainty quantification methods provide "confidence scores" for predictions that are essential for making proper decisions in high risk situations [2]. While a good discriminative power (AUC) is important, a model with large calibration gap is providing "deceptive" probabilities that can create unrealistic expectation and direct limited resources to further development of unsatisfactory compounds [26]. Accordingly, in addition to ability to predict well (predictability), a model reliability that is translated to actual clinical decisions (calibration) is an equally important criterion [26].

2.3.2. Post-hoc Calibration (CalibratedClassifierCV)

This section covers post-hoc calibration methods, i.e. methods used to adjust the predicted probabilities of toxicity in order to obtain the correct fraction of molecules that are

toxicologically adverse [19]. Calibration of machine learning models is a critical task in the pharmaceutical industry in order to turn the scores into a meaningful risk assessment [28].

While machine learning methods like SVM or Neural Networks rank outputs correctly for a set of compounds in a virtual screen, the obtained scores do not necessarily correspond to true empirical probabilities [28]. Post-hoc calibration serves as an additional computational mapping function, that transforms the score f(x) into a calibrated probability P(y=1|x) [28].

Originally developed for SVMs, Platt Scaling proposes the use of a sigmoid function for calibration. The mathematical relationship is defined as [7]:

$$(18). \qquad P(y = 1|f) = \frac{1}{1+exp\,(Af+B)}$$

,where: f: is the classifier raw output (distance from decision boundary) and A,B: are scalar parameters obtained by Maximum Likelihood Estimation (MLE) on an independent validation dataset [7].

When the score-to-probability relationship is approximately sigmoidal, this method can be particularly effective at detecting and fixing errors. Such distributional properties are commonly observed in calibration errors and are particularly pronounced in models trained with cross-entropy loss [11].

When the relationship between scores and probabilities is non-linear then Isotonic Regression can be used to fit the relationship. This is a non-parametric regression method in which no assumptions are made about the form of the regression function. A Pair-Adjacent Violators (PAV) algorithm is used to build a stepwise non-decreasing regression function that minimizes the mean squared error [28]:

$$(19). \qquad min \sum_{i=1}^{n} (\widehat{p_i} - y_i)^2$$

Subject to $\widehat{p_i} \leq \widehat{p_j}$ if $f_i \leq f_j$

Isotonic regression performs well with sufficiently large calibration datasets but may overfit on small datasets [19].

Most of these methods can be implemented using the CalibratedClassifierCV class in the scikit-learn package [29]. Calibration itself is challenging because it is easy to overfit the classification. If calibration is performed on the same data used for fitting, probability

estimates become over-optimistic. The CalibratedClassifierCV class avoids overfitting by performing cross-validation methods. During these methods, the dataset is split into k folds. For each fold, the classifier is trained on all k-1 segments, and a calibration (Platt/Isotonic) is performed on the remaining segment. For final inference, probabilities from the calibrated fold-specific models are averaged [29], [30].

In the ADME-Tox sector, post-hoc calibration is a necessity for informed decision-making rather than a mere statistical refinement. Regulatory authorities, such as FDA/EMA, emphasize robust validation, transparency, and reliable risk estimation. Safety limits could then be determined by using the tools from the calibrated classification module. For example, knowing that the probability of hepatotoxicity for a given drug is 15% (from a calibrated model) may be safe for development by a company. However, this value could actually represent a 40% risk, leading to significant failures in final stage of clinical trials. As a result, Calibration is considered the bridge between computational prediction and clinical safety [19].

2.3.3. Metrics of Reliability

Reliability of machine learning models applied to ADME-Tox prediction tasks needs to be evaluated in addition to the discrimination performance . Standard discrimination measures, such as AUC, evaluate the ranking performance of a model, i.e. how well the model sorts instances in accordance with their labels. Reliability metrics, in contrast, evaluate the quality of the predicted probabilities in relation to the actual frequency of toxic events in the data [31].

A commonly used criterion for evaluating the performance of a classification model is that the model should be perfectly calibrated. Intuitively, perfect calibration means that for any predicted probability p, approximately p% of the corresponding predictions should be correct. However, exact perfect calibration is unattainable from a finite sample [31]. Due to that, the following statistical approximations are utilized:

Expected Calibration Error (ECE) and Maximum Calibration Error (MCE).

The ECE measures the extent to which a given model deviates from perfectly calibrated probabilities. For models that are perfectly reliable, the standard metric to measure their error is the Expected Calibration Error (ECE). To compute the ECE of a probabilistic

classifier, one first partitions the output range of the model's distribution into M evenly spaced intervals (bins). Then, for each bin, the error is computed as the absolute difference between the mean predicted confidence and the observed accuracy within that bin. The ECE is the weighted average of these errors over all bins [11]:

$$(20). \quad ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} |acc(B_m) - conf(B_m)|$$

,where:

M: the number of the bins in total [31]

n: the total number of samples in the test set

B_m: The set of samples whose confidence falls within bin m

acc(B_m): the accuracy of bin m. It is the ratio of correctly classified items within bin m

conf(B_m): the mean predicted probability (confidence) of the samples in bin m [31]

Maximum Calibration Error (MCE) is especially critical for high-risk applications such as pharmaceuticals. Maximum Calibration Error focuses on the Worst Case Error i.e. representing the worst-case calibration error, making it particularly conservative and suitable for high-risk applications [31]:

$$(21). \quad MCE = \max_{m \in \{1,\dots,M\}} |acc(B_m) - conf(B_m)|$$

Brier Score: The "Proper" Scoring Rule

A Proper Scoring Rule is minimized in expectation only when predicted probabilities equal the true probabilities, thus incentivizing honest probability reporting. The Brier Score is defined by the mean squared error between the predicted probabilities (s_i) and the actual outcome (y_i) [32]:

$$(22). \quad BS = \frac{1}{N} \sum_{i=1}^{N} (s_i - y_i)^2$$

The strength of the Brier Score lies in its decomposition into three measures that evaluate the accuracy of probability forecasts:

1) Calibration Loss (CL): This loss function measures how well the probabilities assigned by a learning machine correspond to the true frequencies of classes [32].

2) Refinement Loss (Resolution): Resolution measures how well the model distinguishes between classes. Higher resolution indicates better discrimination [32].
3) Uncertainty: The variance within the dataset [32].

Negative Log-Likelihood (NLL)

NLL or Cross-Entropy Loss is the common quality metric for probabilistic models [11]. NLL is also a common objective function for deep learning that gets minimized during training [11].

$$L = -\sum_{i=1}^{n} log(\hat{\pi}(y_i|x_i)) \quad (23).$$

, where $\hat{\pi}$ is the predicted probability for the actual class y_i [31]. Despite its popularity, NLL can be prone to overfitting, where a model improves its classification accuracy while simultaneously degrading its probability calibration [11].

Modern Variants: Adaptive and Class-wise ECE

The traditional Expected Calibration Error (ECE) is affected by several well-documented limitations. First, fixed-width binning can produce unreliable estimates when confidence scores are not uniformly distributed across bins, since some bins may contain very few samples. Second, standard ECE evaluates only the top-predicted class, making it less suitable for multi-class settings where calibration errors across all output classes should be assessed [27].

- ACE (Adaptive Calibration Error): Instead of using equal-width intervals, equal-frequency intervals (adaptive binning) are used, ensuring that every bin has the same number of samples. This prevents the metric from being dominated by sparsely populated confidence regions and ensures more reliable estimates across the full probability range [27].
- Static/Class-wise ECE-Traditional ECE computes the calibration error for the class with the highest probability (top 1). Class-wise ECE extends this to all K classes and instead calculates and sums the calibration error for every value in the output vector.
- Thresholded ACE (TACE): This measure does not allow probabilities of unlikely classes to "drown" the measure by only including probabilities greater than a threshold, ε [27].

In pharmaceutical research, the selection of the appropriate metric depends on whether the objective is global optimization (Brier Score) or ensuring the accuracy of specific risk levels (ECE/ACE). The use of adaptive metrics is important to avoid biased estimates in datasets with significant class imbalance [33].

2.4. Out-of-Distribution Detection (OOD)

Out-of-Distribution Detection (OOD) is an important aspect of developing robust ADME-Tox models [34]. OOD test compounds (which fall outside of the chemical space of the training sett) such as those containing novel scaffolds, rare functional groups or even those with less obvious but distinct patterns are difficult to identify reliably with standard predictive pipelines. Failure to identify such compounds can result in inaccurate safety assessment and significant risk in early drug development processes [2]

To improve OOD detection methods, it is useful to study the role of confidence. Confidence is typically defined as a measure of how certain a network is of a particular classification [35]. The Maximum SoftMax Probability (MSP) of the classification outcome of a network has been employed as a simple confidence measure for classification error in deep networks. Intuitively, in-distribution data that have been correctly classified, including OOD data that have been misclassified, should be separated well by a confidence measure [35]. For classical machine learning methods, the class probabilities estimated by a predictor are a common surrogate for confidence. However, for classifiers trained on small datasets, the predicted probabilities in the vicinity of the decision boundary can become overly confident [35].

For some classifiers such as LR, methods to calibrate the probability output (e.g. Platt Scaling, Isotonic Regression) can make more sense as they can increase the interpretability of the confidence scores assigned by the classifier [34]. However, calibrated probabilities alone are not sufficient for reliable OOD detection. For a model such as XGBoost, one possible method for performing OOD detection is to use the variability of predictions on split decision trees as a proxy for uncertainty [34].

For certain models (e.g. SVM) and representations (e.g. ECFP4 fingerprints), distance-based methods may be more suitable for defining the Applicability Domain (AD) [36]. The Mahalanobis distance, with respect to the training distribution, is often more informative than Euclidean distance, especially for high dimensional data. For advanced

user control, a threshold can be pre-specified, and samples predicted above this threshold flagged as probably OOD [34].

In addition to the already mentioned approaches, structural similarity analysis can also be used in chemoinformatics. In ADME-Tox tasks, OOD status can be approximated through Tanimoto similarity to compounds in the training set. Thus, for a given molecule, which is to be predicted, low similarity to compounds in the training set would suggest a region outside the model's applicability domain [36].

We note that energy-based scores provide non-probabilistic confidence scores and, for learning, lower energy for in-distribution samples and higher energy for OOD samples can be used as an additional signal to distinguish between known (in distribution) and unknown compounds [36].

In addition to providing more accurate toxicity predictions, detecting OOD instances can enable virtual triage such that most of the chemical space can be screened to identify compounds for which models are most reliable [2]. Rather than flagging such instances for automatic rejection as inactive, we suggest that they be viewed with suspicion, verified manually or, where appropriate, tested in vitro in order to increase confidence in the prediction [2]. Overall, the detection of OOD instances serves to increase the transparency of model limitations, mitigate the silent failure risk, and to enhance the trustworthiness of the computational toxicology approach for pharmaceutical development [2].

2.5. Data Drift Monitoring and Applicability Domain

2.5.1 The Phenomenon of Dataset Shift in Machine Learning

When applying a supervised statistical learning model in a deployment, it is typically assumed that the training data and test data in the deployment are independent and identically distributed [37]. In terms of the data that were used to generate a set of features X and corresponding target variables Y, it is typically assumed that the joint probability distribution P(X,Y) over the training data $P_{train}(X,Y)$ equals the joint probability distribution P(X,Y) over test data $P_{test}(X,Y)$ [38]. However, in many pipelines for computational toxicology, the data do not satisfy this stationarity assumption, and the problem of dataset shift needs to be addressed [37].

The problem of dataset shift can be decomposed by the chain rule of probability into the two main causes for this phenomenon: covariate shift and concept drift [38]. Covariate shift occurs when there is a modification in the marginal distribution of the input features ($P_{train}(X) \neq P_{test}(X)$), while the underlying true conditional mapping P(Y|X) remains invariant [37]. For cheminformatics and QSAR modeling the problem of covariate shift is directly related to the problem of covering the chemical space not yet covered by a model in training in a large database by this model [39]. For models which are unable to perform an adequate statistical extrapolation of unrepresented structure or of function groups not included in the data used for the model development in the new chemical space explored by extend the modeling using new data set, a severe loss of prediction accuracy is observed [37].

Conversely, concept drift arises when the conditional probability mapping changes ($P_{train}(Y|X) \neq P_{test}(Y|X)$), indicating that the underlying biological relationship or "concept" defining the target property has mutated [37]. Misdiagnosing such changes leads to over-optimistic predictions made by a QSAR model. The validation strategy of time-split cross-validation is more suitable for measuring the predictive performance of a model that faces distributional changes [40].

2.5.2 The Concept of Applicability Domain

Within QSAR modeling the Applicability Domain (AD) of a model refers to the 3rd OECD Principle for Validation and is defined as the response space and the chemical structure space in which a model is able to make reliable predictions [41]. “The Applicability Domain is an important aspect in QSAR modeling since most of the QSAR models are based on machine learning algorithms”. These models are inherently reductionist and thus limited to the chemical structures, the set of physicochemical properties, and the mechanisms of action that have been included in the training set of data [41].

The Applicability Domain (AD) defines the interpolation space of a model, establishing a boundary that separates it from the extrapolation space [41]. Within the Applicability Domain of a QSAR model, the predictions correspond to an interpolation in the response space (interpolation), while outside the Applicability Domain, the predictions are associated with high uncertainty and correspond to a region outside the model’s validation

space or an extrapolation [42]. The safeguarding of such large collections of compounds for corporate use, stored in remote areas of chemical space, is critical. For new potential drugs as well as their predicted toxicity, an explicit evaluation of their AD properties is required to prevent 'silent' prediction errors. Standard machine learning models will return a confident numeric answer for any input molecule within the domain of trustworthy prediction of the training data set, even when that domain has not been explored in sufficient detail for new compounds of interest [42].

There is an inherent trade-off between the breadth of the applicability domain and the model's predictive power: a broad AD (global model) often suffers from lower precision, while a restricted AD (local model) provides higher reliability for specific chemical classes [41]. Consequently, characterizing the AD enables researchers to perform virtual triage, effectively prioritizing compounds for which the computational predictions can be trusted for clinical decision-making [39].

2.5.3 Feature Conditioning and Dimensionality Reduction for Molecular Fingerprints

For computational toxicology models, delimiting the Applicability Domain (AD) and performing Uncertainty Quantification (UQ) of predictions essentially depends on a correct mathematical representation of the molecules under study [37]. Molecular fingerprints (MP) such as Extended-Connectivity Fingerprints (ECFP) convert given chemical information into computable variables leading to a set of problems in feature space [37].

In computational toxicology molecules are treated as single points in a n-dimensional space where n corresponds to the length of a fingerprint, for example 1024 or 2048 bit for Extended-Connectivity Fingerprints (ECFPs). Circular so-called ECFPs [37] (Extended-Connectivity Fingerprints) encode the local topological environment of all atoms in a molecule by numbering them in increasing order and assigning them an integer identifier which is also used as a feature in the fingerprint [37]. These molecular identifier (Mol ID) values are mapped into a feature vector x:

$$x = [b_1, b_2, \dots, b_n]^T \quad (24).$$

where $b_i \in \{0,1\}$ for binary fingerprints or $b_i \in N$ for count-based fingerprints [37]. The geometry of the molecular representation in feature space is highly sparse. Only a small

fraction of the thousands of possible structural features of a molecule is “active” and therefore included in the representation of a single chemical compound [37].

To calculate the AD the Mahalanobis distance, taking into account the correlations between the descriptors, is used [37]. The formula for the distance d_M of a sample x from the mean μ of the training set is:

$$d_M(x,\mu) = \sqrt{(x-\mu)^T S^{-1}(x-\mu)} \quad (25).$$

High-dimensional feature spaces (high n) in combination with a low number of training samples (N) for the calculation of the inverse S^{-1} leads to mathematical failures as n is greater than N [37]. If the number of features (n) is greater than the number of training samples (N), the matrix S is singular and det(S)=0, therefore S cannot be inverted [37].

There is also the fact that the fingerprint itself is already highly dimensional, i.e. of very high dimensionality, and many of the descriptors obtained for a single molecule (or image) have features that are either highly multi-collinear or even have zero variance for a given set of samples [37]. This makes the features themselves very unreliable for many subsequent processing stages, but also results in the features’ covariance, i.e. the covariance matrix S of the features, being ill-conditioned and thus statistically unstable on many cases [37].

PCA (Principal Component Analysis) is a crucial component in our pipeline, to transform the high dimensional input space of the fingerprints to a more suitable space for the mathematical computations for the distances in the classification space [37]. Therefore, by applying a PCA to the descriptors of the input samples, the high dimensional input space is mapped to a new space of lower dimensionality, where the axes of this new space correspond to the principal components or the directions of the highest variance of the input data in the original space after an orthogonal transformation, i.e. the axes of the input space were rotated [37].

The process involves:

- Preprocessing and Centering: In a general PCA pipeline, input features are typically centered before dimensionality reduction. For continuous descriptors, standardization may also be applied so that all variables contribute on a comparable scale. In the case of binary molecular fingerprints, all dimensions

already share the same binary scale, while the PCA transformation centers the data internally before extracting principal components [37].

- Eigenvalue Decomposition: PCA identifies orthogonal directions of maximum variance by decomposing the covariance structure of the transformed data. These directions correspond to the principal components and provide a new coordinate system in which correlated high-dimensional features are represented more compactly [37].
- Dimensionality Reduction: By retaining only the first k principal components, the original high-dimensional fingerprint space can be projected into a lower-dimensional representation. This reduces redundancy and improves the numerical stability of subsequent distance-based calculations [37].
- Regularized Covariance Estimation: After dimensionality reduction, the covariance matrix is estimated in the reduced principal-component space. A small diagonal regularization term can be added before matrix inversion to avoid numerical instability and to make Mahalanobis-distance computation more robust [37].

The new covariance matrix in the feature space of the Principal Components is a diagonal and full rank matrix, which can be then inverted to compute the distance in the feature space and to use it as a tool for OOD detection [37].

2.5.4 Statistical and Information-Theoretic Drift Metrics

To quantify dataset shift and precisely delineate the Applicability Domain (AD), the deployment of rigorous mathematical metrics capable of comparing a reference distribution (training data) against a newly observed distribution (test or production data) is indispensable. These metrics enable the system to "fail loudly," issuing systematic alerts whenever the independent and identically distributed (i.i.d.) assumptions are violated.[37].

Two-Sample Kolmogorov-Smirnov (KS) Test

The two-sample Kolmogorov-Smirnov (KS) test is a non-parametric statistical hypothesis test used to determine whether two continuous distributions differ significantly [37]. In the context of data drift detection, the KS test quantifies the maximum vertical distance

between the empirical cumulative distribution functions (CDFs) of the two underlying samples [43].

The test statistic Z is formally defined as:

$$(26). \quad Z = \sup_{z} \left| F_p(z) - F_q(z) \right|$$

where F_p and F_q denote the empirical cumulative distributions of the training and test datasets, respectively. If the calculated statistic Z exceeds a critical threshold corresponding to a predefined significance level, the null hypothesis, which posits that both samples originate from the identical distribution, is rejected [37]. In computational toxicology, the KS test is frequently executed in a feature-wise manner to identify specific molecular descriptors that exhibit significant statistical divergence [43].

Population Stability Index (PSI)

The Population Stability Index (PSI) is a well-established measure to quantify the change in a time-dependent variable. The symmetric Kullback-Leibler (KL) divergence, often used in information theory, is the base for the symmetric formulation of the PSI. It calculates the difference of the frequencies of a time-series by dividing it into bins [37].

The PSI is mathematically formulated as:

$$(27). \quad PSI = \sum_{i=1}^{B} (\widehat{p_i} - \widehat{q_i}) * (ln\widehat{p_i} - ln\widehat{q_i})$$

where:

- B represents the total number of bins.
- $\widehat{p_i}$ denotes the relative frequency of observations within the i-th bin for the baseline (training) distribution.
- $\widehat{q_i}$ signifies the corresponding relative frequency within the same bin for the target (test) distribution.

In empirical practice, the interpretation of the resulting PSI score adheres to the following standardized thresholds [37]:

- PSI < 0.10 : Minimal or no significant distribution shift.
- 0.10 ≤ PSI < 0.25 : Moderate drift, indicating a variation that warrants continuous monitoring.

- PSI ≥ 0.25 : Severe distribution shift, compromising the model's predictive reliability and necessitating model retraining.

Wasserstein Distance (Earth Mover's Distance)

The Wasserstein distance or Earth Mover's Distance (EMD) solves the classical transportation problem [37]. Given a "supply" (or "earth") probability distribution and a "demand" (or "holes") probability distribution, the EMD will be the minimum amount of work that has to be done in order to move the earth into the holes. The amount of work that is needed for moving a given amount of ground distance between two points in space (i.e. between two features in a molecular signature) is defined to be 1. The formula for the EMD of two discrete molecular signatures P and Q is given by the formula for the mass flow matrix f_{ij} that minimizes the total transportation cost [37]:

$$EMD\ (P,Q) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}\, d_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}} \quad (28).$$

Where d_{ij} represents the ground distance between the structural elements p_i and p_j and f_{ij} signifies the optimal mass flow directed between them [37].

The primary advantage of the EMD over classical histogram-based metrics (such as the PSI) lies in its inherent robustness against quantization artifacts. While standard bin-by-bin comparisons completely collapse if empirical values shift marginally across a bin boundary, the EMD explicitly accounts for the geometric proximity of adjacent regions. Consequently, it yields a more natural, stable, and mathematically continuous measure of structural similarity within the defined chemical space [37].

Chapter 3. Experimental Methodology

The experimental methodology employed in this work was developed on the basis of the theoretical framework presented in chapter 2. The subsection on ADME-Tox justified the selection of some biologically relevant ADME and toxicity endpoints as prediction targets, the subsection on machine learning justified the choice of some machine learning algorithms for model development, The subsection on reliability and uncertainty justified the probability calibration requirement for models developed in this work, and the subsection on Applicability Domain justified the need for OOD samples detection mechanisms in this work. This chapter outlines the practical implementation of the theoretical concepts introduced in chapter 2 and the resulting computational system developed in this work.

3.1.Methodology Overview

The methodology adopted in this paper is based on four independent steps designed to make the best use of the available data and to assess the performance in a realistic way. These steps are naturally divided into the steps of training, calibration and validation. The steps followed in real life work are exactly reproduced and are clearly separate and reproducible.

Step 1: The initial stage of this process starts by downloading three publicly accessible ADME-Tox datasets from the Therapeutics Data Commons (TDC) website. These were then validated with the public Version of the RDKit tool. Some key physicochemical properties were also calculated based on Lipinski's Rule of Five, and the degree of class imbalance was quantified for each dataset.

Step 2: In order to make use of the chemical information in the SMILES string representing a molecule, it is necessary to transform this string into a vector of numbers which can be processed by machine learning algorithms. For this purpose, in this work the chemical information has been transformed into extended connectivity fingerprints (ECFP4-type), with a radius of 2 and 2048 features (bits).Each dataset was divided into training (70%), validation (15%) and test (15%) sets using stratified sampling to preserve class proportions across all subsets. The three different classifiers used are LR model (a

linear classifier), a Support Vector Machine model (a kernel based classifier) and an XGBoost model (an ensemble of decision trees). Class imbalance was addressed via class_weight='balanced' for LR and SVM, and scale_pos_weight for XGBoost. Their performance has been measured by four different metrics, adequate for the assessment of the quality of a classification system for imbalanced classes. The metrics used are ROC-AUC, PR-AUC, F1-score and Balanced Accuracy.

Step 3: Probability Calibration: Even high ROC-AUC values do not imply that the model's quality at predicting the output quality in terms of the toxicity predictions probability is also reliable. In this step we compare two very popular post hoc methods for calibrating a model's output quality in terms of the output probability's frequency of true toxicity. The first one is the often used Platt Scaling, a very simple method that trains a logistic function over a model's raw output to improve the output's probability quality. The second approach is the nonparametric Isotonic Regression method, that fits a monotonically non decreasing step function to map uncalibrated probabilities to better calibrated values. As before, both of the methods are trained on the validation data and are then applied to the test data, which the model has not seen before in both cases. The model's output is then compared to the output of the both post hoc methods with the help of five quality measures, the Expected Calibration Error (ECE), the Adaptive Calibration Error (ACE), the Maximum Calibration Error (MCE), the Brier score as well as the Negative Log-Likelihood (NLL) score. Furthermore, also so-called reliability diagrams are created, and it is made visible how both methods improve the model's output quality.

Step 4: OOD Detection: Even when a model produces confident predictions, it may fail to correctly predict very novel, outside the chemical space of the training compounds [35]. In order to identify such compounds within the applicability domain of a model, the applicability domain is defined by two OOD detection methods commonly used in the literature for OOD detection in chemical space [12]. Mahalanobis distance OOD and k-NN distance OOD detectors are first fitted on the training data and then used to compute a score for each compound in the test and synthetic OOD datasets [34] [12]. Compounds scoring above a threshold set at the 95th percentile of the scores for the training compounds in the training datasets are flagged as OOD [34]. For the Mahalanobis distance OOD detector, the feature space is first reduced to 50 principal components via PCA to ensure a well-conditioned covariance matrix given the small dataset sizes [5]. The results are presented through score distribution histograms, OOD detection ROC curves,

plot of OOD score versus predicted confidence (XGBoost), a PCA chemical space visualization, and a summary table reporting threshold values, flagging percentages, and detection AUROCs for both methods.

3.2.Datasets

This work utilizes three different binary classification datasets from the Therapeutics Data Commons (TDC) platform [44]. The TDC is an open access repository containing curated, ready to use data for machine learning in drug discovery. Three fixed versions of the datasets were used, the molecular structures were in smiles format, directly compatible with the RDKit library and existing fingerprint generation pipelines. The TDC already contains a wide variety of datasets for different ADME-Tox endpoints of established biological relevance. In this work, three datasets selected to study Distribution for blood brain barrier (BBB) permeability and toxicity for two different toxicological endpoints of clinical relevance each. Together, these datasets cover both the ADME (Distribution) and Toxicity properties of the ADME-Tox spectrum.

3.2.1. Tox21 / NR-AR

The Tox21 dataset consists of approximately 10,000 chemicals tested in 12 different nuclear receptor and stress response pathway assays using High Throughput Screening (HTS). The data were originally released as part of the Toxicology in the 21st century initiative, a multi-agency collaboration between the National Institutes of Health, the US Environmental Protection Agency (EPA) and the Food and Drug Administration [45]. The NR-AR (Nuclear Receptor – Androgen Receptor) endpoint was selected from the 12 assays conducted as part of the Tox21 effort. The NR-AR assay detects whether a chemical activates the androgen receptor, a nuclear receptor that regulates the expression of genes involved in androgen signaling. Chemicals that activate the NR-AR are classified as endocrine disruptors (EDCs) with the potential to cause hormone dependent cancers and adverse reproductive or developmental effects [46].

This dataset was obtained from the TDC platform. In this work, the NR-AR endpoint was used for toxicological activity measurement. After filtering for missing labels, the resulting dataset contains 7,265 chemical compounds, each assigned a binary label: 1 for positive activation of the NR-AR nuclear receptor by the chemical, or 0 for no activation

of the NR-AR. The dataset exhibits significant class imbalance, with a negative-to-positive ratio of approximately 22:1, reflecting the real-world reality of biological screening where active compounds constitute a small minority. This is the largest of the three datasets used in this work, providing sufficient statistical power for model training and evaluation.

The NR-AR endpoint was selected for several reasons. First, it represents a well-characterized biological target of high clinical significance. Second, its substantial class imbalance provides a realistic and challenging evaluation scenario for classifier performance and last but not least, it constitutes a standard benchmark in the computational toxicology literature, enabling comparison with published results.

3.2.2. BBB_Martins

The BBB_Martins dataset represents the Distribution component of the ADME spectrum. The Distribution dataset BBB_Martins consists of 2,030 compounds for the property of Blood-Brain Barrier Permeability. It is formed by the highly specialized endothelial cells which line the brain capillaries. These cells are supported by the foot processes of astrocytes and by pericytes [47]. The barrier function is restricted to the cerebral vessels, and its role is to control the entry of substances from the blood to the brain. Consequently, the BBB function prevents the entry of neurotoxic substances and pathogens into the CNS, while also limiting the action of most drugs to the brain.

BBB permeability is a critical pharmacokinetic property that has two opposing roles in drug development. First, for many therapeutic applications in the CNS (e.g. Alzheimer's disease, Parkinson's disease, depression) a drug must penetrate into the brain to be active. On the other hand, for many outside the CNS (e.g. anti-infectives, cancer chemotherapeutics), a drug must avoid penetration into the brain to prevent toxic side effects in this very sensitive organ.

The BBB_Martins dataset was compiled from both in vivo and in vitro experimental studies and was published by Martins et al. [48]. The dataset is part of the TDC ADME class. It contains information on aproximately 1,975 compounds, following the removal of 55 duplicate SMILES entries identified during canonical SMILES standardisation, where each compound is represented by a binary label (1 for BBB+ and 0 for BBB−). The dataset is moderately imbalanced, with approximately 76% BBB+ compounds.

This dataset was selected to represent the Distribution component of the ADME spectrum because it is publicly available, well-studied, suitable in size for classical ML and straightforwardly formulated as a binary classification task.

3.2.3. DILI

Drug Induced Liver Injury (DILI) is reported to be one of the leading causes of post-market drug withdrawal. DILI is also reported to be one of the leading causes of late-stage failure of clinical trials [49]. The liver is the primary organ involved in the metabolism of drugs. As a result, this organ is particularly sensitive to toxic substances, either in their original form or as toxic metabolites produced by biotransformation reactions. The forms of DILI include acute hepatitis, cholestatic hepatitis and even hepatocellular necrosis. In severe cases of DILI, patients may need to have a liver transplant or even die as a result. The prediction of DILI using in silico models is a very active area of research, since the early identification of potential hepatotoxic compounds could prevent the high costs of late-stage clinical trials and their frequent failures. The DILI dataset from TDC was created from adverse reaction information retrieved from publicly available FDA databases as well as from the scientific peer-reviewed literature [50].

The DILI dataset is the smallest of the three datasets used in this study, comprising 475 compounds, each assigned a binary label: DILI-positive (hepatotoxic, label = 1) or DILI-negative (label = 0). The dataset exhibits a moderate class balance, with an approximately balanced class distribution (239 DILI-negative vs. 236 DILI-positive, ratio ~ 1:1). The relatively small size of this dataset makes it particularly well-suited for assessing the reliability of predicted probabilities (calibration) and the applicability domain of the trained models (OOD detection), both of which constitute central objectives of this study.

Dataset	Endpoint Type	N	Class Imbalance (neg/pos)	Learning Task	Task	Positive Label (1)
TDC.Tox21 / NR-AR	Toxicity	7,265	22.51	single_pred.Tox	Binary	Androgen receptor activation
TDC.BBB_Martins	ADME (D)	1,975	0.32	single_pred.ADME	Binary	BBB+

TDC.DILI	Toxicity	475	1.01	single_pred.Tox	Binary	DILI

Table 1. *Dataset Summary Table*

3.3.SMILES Validation & Class Balance Analysis

Before performing any further processing of the TDC datasets, all compounds were first validated in the datasets for chemical accuracy. Each compound is given as a SMILES string (Simplified Molecular Input Line Entry System) and as the TDC datasets are curated, there are only a few errors with the SMILES strings for the compounds, i.e. the SMILES string for a compound does not form a valid chemical structure.

The SMILES strings representing the chemical structures of the compounds in the three datasets were analyzed to assess their chemical validity. For this purpose, the Chem.MolFromSmiles function of the RDKit cheminformatics library was used. SMILES validation showed that all three datasets contained chemically parseable SMILES strings. Specifically, no invalid SMILES were detected in Tox21/NR-AR, BBB_Martins, or DILI. However, canonical SMILES standardization identified 55 duplicate molecular entries in BBB_Martins, which were removed before model training and analysis. No duplicate canonical SMILES were detected in Tox21/NR-AR or DILI. The remaining compounds were then used to generate ECFP4 Morgan fingerprints, which served as the common molecular representation for all machine learning models. As a result, the final datasets used in the experiments contained 7,265 Tox21/NR-AR compounds, 1,975 BBB_Martins compounds, and 475 DILI compounds.

Following SMILES validation, the class distribution of each dataset was examined. For each dataset, the number of positive and negative instances was computed along with the negative-to-positive ratio (neg/pos), which serves as a measure of class imbalance.

Class imbalance is a critical parameter that influences both algorithm selection and the choice of evaluation metrics. Accordingly, all models were trained with explicit imbalance correction: class_weight='balanced' for LR and SVM, and scale_pos_weight for XGBoost.

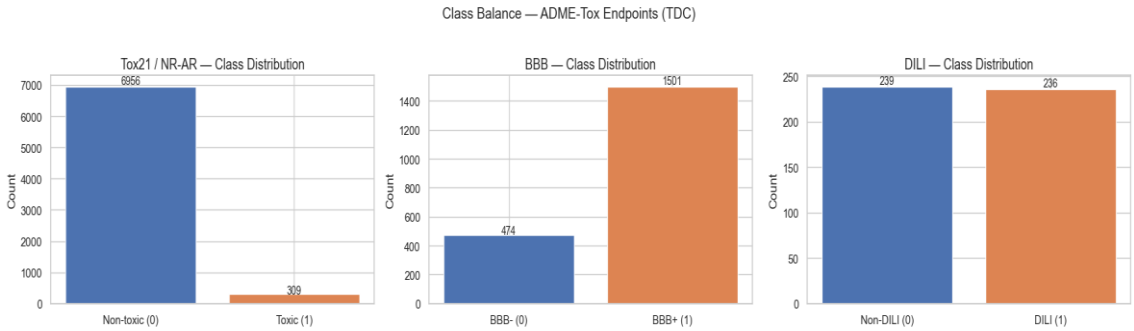


Figure 14. *Class distributions of the three datasets after SMILES validation (TDC). Tox21/NR-AR exhibits severe class imbalance, BBB_Martins a majority-positive distribution, and DILI a near-balanced distribution (~1:1)*

3.4.Physicochemical Properties & Lipinski Rule of Five

To characterise the chemical space of the datasets, six physicochemical descriptors were computed using RDKit for the BBB_Martins and DILI datasets: molecular weight (MolWt), lipophilicity (LogP), topological polar surface area (TPSA), number of hydrogen bond donors (NumHDonors), acceptors (NumHAcceptors), and number of rotatable bonds (NumRotatableBonds). The Tox21 dataset was excluded from this analysis due to the nature of its endpoint, which is not directly related to absorption-driven physicochemical properties. A total of 2,450 molecules were analysed (1,975 from BBB_Martins and 475 from DILI).

Lipinski's Rule of Five defines four upper-bound criteria for oral bioavailability: MolWt $\leq$ 500 Da, LogP $\leq$ 5, NumHDonors $\leq$ 5, and NumHAcceptors $\leq$ 10. A compound with two or more violations is considered to face a significant barrier to passive oral absorption [14].

	MolWt	**LogP**	**NumH Donors**	**NumH Acceptors**	**TPSA**	**NumRotatable Bonds**	**Label**
count	2450.00	2450.00	2450.00	2450.00	2450.00	2450.00	2450.00
mean	346.36	2.24	1.68	4.68	74.21	4.34	0.71
std	157.49	2.25	2.03	3.35	61.91	3.35	0.45
min	6.94	-15.44	0.00	0.00	0.00	0.00	0.00
25%	254.28	1.14	1.00	3.00	35.85	2.00	0.00

50%	322.36	2.45	1.00	4.00	59.11	4.00	1.00
75%	411.45	3.72	2.00	6.00	94.83	6.00	1.00
max	1879.68	10.81	27.00	33.00	736.61	35.00	1.00

***Table 2**. Descriptive Statistics of Physicochemical properties.*

The median molecular weight (322.36 Da) and median LogP value (2.45) fall within Lipinski's Rule of Five limits, indicating that most compounds occupy drug-like chemical space. However, the extreme values, particularly MolWtmax = 1,879.68 Da and LogPmin = −15.44, indicate that the datasets also contain compounds outside the classical drug-like space, such as highly polar or large biologically active molecules. This broad chemical diversity affects the sparsity and distribution of ECFP4 fingerprints and may influence model generalisation.

Figure 15 presents the distributions of each physicochemical property across datasets (BBB, DILI) and class labels (0 = negative, 1 = positive) using violin plots. In the BBB dataset, BBB+ molecules (label = 1) display lower TPSA and NumHDonors values compared to BBB− compounds, consistent with the hypothesis that blood–brain barrier permeability is favored by lower polarity and fewer hydrogen bond donors. In the DILI dataset, the distributions of the two classes show substantial overlap, confirming the more elusive nature of this endpoint.

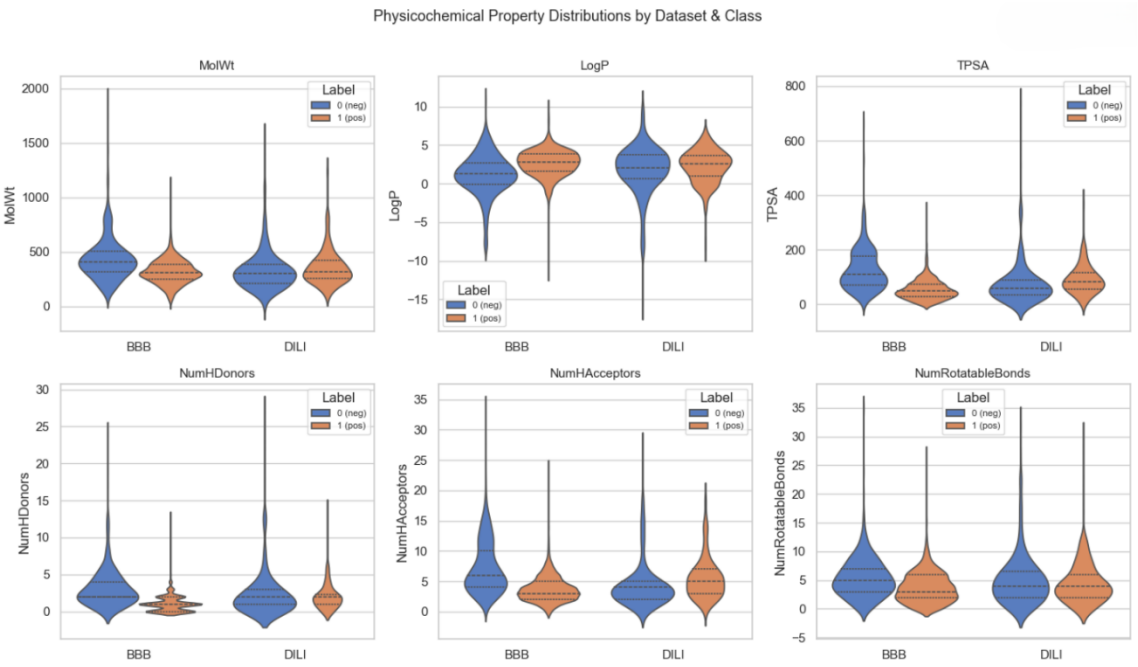


***Figure 15.** Physicochemical Property Distributions by Dataset and Class*

Rule of 5 violations	BBB (N=1,975)	%	DILI (N=475)	%
0	1,677	84.9	373	78.5
1	162	8.2	53	11.2
2	91	4.6	34	7.2
3	43	2.2	14	2.9
4	2	0.1	1	0.2

***Table 3**. Rule of Five Violation Counts by Dataset.*

In both datasets, most compounds (BBB: 84.9%, DILI: 78.5%) satisfy all Rule of five criteria. The minority of compounds with two or more violations (BBB: 6.9%, DILI: 10.3%) consists primarily of complex or macrocyclic small molecules, natural products, and glycopeptides, compounds that, despite their larger size or higher polarity, retain measurable biological activity.

3.5.Fingerprints Generation and Baseline Models

3.5.1. ECFP4 Morgan Fingerprints

For each molecule a numerical feature vector was generated by using the ECFP4 Extended-Connectivity Fingerprints (radius = 2, 2048 bits). The Morgan algorithm was used to calculate the fingerprints within RDKit. The Morgan algorithm starts from the individual atoms of a molecule and for each of them it computes the chemical neighbourhood of atoms within a radius of 2 bonds. Each unique circular substructure encountered at each radius is hashed to a bit position in the output vector. A bit is set to 1 if that substructure is present anywhere in the molecule. The final fingerprint is a 2048-dimensional binary vector. Each position of the resulting vector thus corresponds to a particular local chemical substructure pattern.

Extended-Connectivity Fingerprints (ECFP4) are by far the most common molecular representation in computational chemistry and drug discovery. In this work, they were selected for the following reasons:

- Expressiveness: ECFP4 encodes local molecular structure, functional groups, rings, and heteroatoms, without requiring a 3D geometry, allowing a fixed-size representation to be generated directly from a SMILES string.
- Compatibility with classical ML: As fixed-length binary vectors, ECFP4 features are directly compatible with LR, SVM, and XGBoost without normalization or

scaling (note that the gamma='scale' setting in SVM does implicitly handle variance).

- Benchmark consistency: By using the same ECFP4 parameters as in typical ADME-Tox studies, the results obtained can be directly compared with other published results.
- All three models use the same molecular representation: Using the same featurization for all three models ensures that observed any differences in performance reflect model architecture rather than molecular representation.

Dataset	Dimensions	Fp_density	Pos%
Tox21	(7265 , 2048)	0.014	4.3%
BBB	(1975 , 2048)	0.021	76.0%
DILI	(475 , 2048)	0.019	49.7%

***Table 4**. Featurization Results.*

The fingerprint density (fp_density), the proportion of bits set to 1, ranges from 1.4% to 2.1%, confirming that ECFP4 produces sparse vectors. The majority of the 2048 bits are 0 for any given molecule. This sparsity is expected and poses no issue, as the selected models (particularly LR with L2 regularisation and XGBoost) are robust to sparse features.

3.5.2. Train / Validation / Test Split

Each dataset was partitioned into three non-overlapping subsets with a ratio of 70/15/15 using stratified sampling to preserve the positive/negative class ratio across all splits. The 70/15/15 ratio represents a reasonable compromise between competing requirements. The 70% training portion provides sufficient data for all three models, while the equal 15% allocations for validation and test ensure that both the calibrator and the final evaluation are based on representative samples. For DILI (N=475), the test set contains only 72 compounds, small enough that results should be interpreted with caution, but this reflects the realistic challenge of limited data in computational toxicology.

3.5.3. Baseline Model Training and Evaluation

Three baseline classifiers, LR, SVM and XGBoost were trained exclusively on the training set of each dataset, using the ECFP4 fingerprints as input. Class imbalance was

addressed in all three models: LR and SVM via class_weight = 'balanced' setting and XGBoost via scale_pos_weight = N_neg / N_pos setting.

Each model was evaluated on the test set using four metrics:

- ROC-AUC: Discriminative ability regardless of threshold.
- PR-AUC: Performance under class imbalance.
- F1 Score: Mean of precision-recall at threshold 0.5.
- Balanced Accuracy: Mean per class recall.

ROC and Precision-Recall curves were generated for each dataset, along with heatmaps enabling direct comparison across the three models. These results are presented and analyzed in Chapter 4.

3.6.Probability Calibration

As discussed in Chapter 2, discriminative performance, as expressed by ROC-AUC, does not provide a sufficient guarantee of the reliability of estimated probabilities. The objective of this section is to describe the experimental post-hoc calibration procedure applied to the three baseline models for each of the three datasets.

3.6.1. Experimental Design

The calibration procedure was designed to maintain strict separation between the training, the calibration and the evaluation. In this work, all base models were trained on the training set (70%), as described in section 3.5.2. The calibrator was trained on the uncalibrated probabilities produced by the base model on the validation set (15%), using the true labels as targets, while evaluation was performed exclusively on the test set (15%), which remained held out throughout the entire development process.

The choice of a post-hoc approach over in-training calibration (e.g., CalibratedClassifierCV with cross-validation) was deliberate. The post-hoc approach preserves the base model unchanged and ensures that the validation set is used exclusively for calibration only, avoiding any form of data leakage. Using training data for calibrator training would result in artificially optimistic probability estimates, as the model has already been exposed to those samples during training.

3.6.2. Methods

Two post-hoc calibration methods were applied, the theoretical foundations of which were presented in section 2.3.2:

- Platt Scaling: A LR model was used, with the single parameter C set to 1e10, effectively disabling the L2 regularizer, as its absence allows the calibrator to fit the relationship between uncalibrated probabilities and true labels as good as possible, without any shrinkage of coefficients. The validation set served as training data, with uncalibrated probabilities val_p as a one-dimensional input (after reshape(-1, 1)) and labels y_val as the target.
- Isotonic Regression: Implemented via Isotonic Regression with parameter out_of_bounds='clip', which ensures that probabilities outside the interval [0,1] are clipped to the boundaries rather than generating errors when applied to test data that may lie outside the range observed during calibration. Training uses the one-dimensional array val_p directly without reshape, as Isotonic Regression accepts 1D input.

Both methods were applied independently to each of the 9 combinations (3 datasets × 3 models), producing 18 sets of calibrated probabilities in total, stored in the dictionary cal_probs with dataset, model, method as a key.

3.6.3. Evaluation of Metrics and Visualizations

For quantitative comparison of calibrated versus uncalibrated probabilities, five reliability metrics were computed for each dataset x model x variant combination (Raw/Platt/Isotonic):

Metric	**Focus**
Expected Calibration Error (ECE)	Mean calibration error (primary metric)
Adaptive Calibration Error (ACE)	Calibration with adaptive bins
Maximum Calibration Error (MCE)	Worst-case calibration error
Brier Score (BS)	Holistic probability quality
Negative Log-Likelihood (NLL)	Information-theoretic probability assessment

Table 5*. Calibration Evaluation Metrics and their Focus.*

Also, three types of graphical representations were produced for each dataset:

- Reliability Diagrams: One subplot per model, with overlapping curves for Raw, Platt, and Isotonic.
- ECE Bar Charts: Horizontal bar charts comparing ECE before and after calibration for each model.
- Brier Score Grouped Bar Charts: Comparison of Brier Score per model for Raw, Platt, and Isotonic variants.

A vigorous analysis and interpretation of the results are presented in Chapter 4.

3.7.Out-Of-Distribution (OOD) Detection

3.7.1. Motivation and Applicability Domain

A machine learning model is trained on a finite set of chemical compounds and is expected to generalize to new compounds with similar chemical structure. However, when the model receives as input a molecule that lies structurally outside its training space, an OOD molecule, there is no guarantee of prediction reliability. Furthermore, many models tend to assign high confidence even to structurally unfamiliar compounds, leading to silent failures cases in which the prediction is incorrect, but the model appears certain.

The Applicability Domain (AD) is defined as the chemical space within which QSAR/ADME-Tox models are considered reliable [37]. In this work, OOD detection is implemented as follows: for each new compound, an OOD score is computed reflecting its distance from the training space. Compounds with a score exceeding a predefined threshold are flagged as OOD and their predictions are considered unreliable.

3.7.2. OOD Detection Methods

Two independent OOD scorers were implemented and trained separately on the training set of each dataset, without access to labels:

- Mahalanobis Distance OOD: The Mahalanobis distance measures how far a point lies from the training distribution, while accounting for feature correlations. Due

to the high dimensionality of ECFP4 fingerprints (2048 bits), dimensionality reduction is first applied via Principal Component Analysis (PCA) to $n_{\text{components}} = 50$ principal components. In the reduced space, the scorer estimates the mean μ and inverse covariance matrix S^{-1} from the training data and computes for each new point x:

$$d_M(x) = \sqrt{(x-\mu)^T S^{-1}(x-\mu)} \quad (29).$$

The threshold is set at the 95th percentile of the Mahalanobis distances of the training data themselves in the reduced space (threshold_quantile = 0.95). Compounds with $d_M >$ threshold are flagged as OOD.

- k-Nearest Neighbors Distance OOD: The kNN OOD scorer computes for each new compound the mean Euclidean distance to its $k = 5$ nearest neighbors in the ECFP4 fingerprint space of the training data:

$$s_{kNN}(x) = \frac{1}{k}\sum_{i=1}^{k} \| x - x_i^{(k)} \|_2 \quad (30).$$

A larger mean distance indicates a structurally more distant compound from the training space. The threshold is similarly set at the 95th percentile of the kNN distances of the training data (threshold_quantile = 0.95, metric='euclidean').

The choice of the 95th percentile as threshold implies that 5% of the training data are by construction flagged as OOD, a deliberate setting that controls the conservatism level of the AD.

3.7.3. OOD Batch Simulation

For each dataset, a synthetic OOD-like batch of 500 non-molecular binary fingerprint vectors was generated. The bit activation probability was matched to the training fingerprint density of the corresponding dataset (0.0144 for Tox21/NR-AR, 0.0209 for BBB_Martins, and 0.0194 for DILI), sampled from a binomial distribution:

$$x_j \sim Bernoulli(\rho_{train}),\ j = 1, \ldots, 2048 \quad (31).$$

This procedure produces vectors with similar overall density to the training data, but without the structural coherence that characterizes real chemical molecules. The intent is to simulate compounds lying outside the training chemical space, while density matching prevents trivial detection based solely on total bit density.

3.7.4. Evaluation

The performance of the OOD scorers is assessed through four complementary analyses:

- OOD Score Distributions: Histograms of distance scores for in-distribution test compounds and synthetic OOD-like fingerprint vectors, with threshold annotation. Effective detection corresponds to clear separation between the two score distributions.
- Detection AUROC: The in-distribution/OOD problem is formulated as a binary classification task, with label 0 assigned to held-out in-distribution test compounds and label 1 assigned to synthetic OOD-like fingerprint vectors. ROC-AUC is computed to measure how well each OOD scorer separates the two groups.
- Confidence vs. OOD Score: A scatter plot correlates the OOD score with the XGBoost predicted probability. This enables identification of cases with simultaneously high model confidence and high OOD score, corresponding to the critical silent-failure scenario.
- Chemical Space Visualization (PCA): Training compounds, test compounds, and synthetic OOD-like fingerprint vectors are projected into two dimensions using PCA, allowing visual inspection of their separation in the reduced fingerprint space.

The results and discussion are presented in Chapter 4.

3.8.Drift Monitoring

Having established the theoretical framework for dataset shift, applicability domain characterisation, and the statistical metrics employed (KS test, PSI, Wasserstein distance) in Chapter 2, this section describes their concrete implementation and the experimental design adopted in the present work.

3.8.1. Experimental Design

A controlled two-batch evaluation scheme was adopted to simultaneously assess each method's sensitivity to genuine distributional shift and its specificity for in-distribution

data. For each dataset, the training set X_{train} serves as the fixed reference distribution against which all incoming batches are compared:

- Batch A (X_{test}): the held-out test compounds drawn from the same stratified split as the training data. Under the i.i.d. assumption, this batch should produce no significant drift signal, serving as a negative control for false alarm rate.
- Batch B (X_{sim}): the 500 random binary fingerprint vectors generated in Section 3.7.3 via matched bit-density sampling. These vectors are structurally non-molecular and lie well outside the training distribution, serving as a positive control for detection sensitivity.

This design enables a direct within-experiment comparison of the three methods' discriminative behavior under known conditions, without requiring external OOD datasets.

3.8.2. Covariate Drift. KS Test on PCA Components

Applying the two-sample KS test directly on the raw 2048-dimensional fingerprint vectors is computationally intractable and statistically unreliable under the curse of dimensionality. The drift test is therefore conducted in a reduced feature space. A PCA transformation fitted on X_{train} projects all batches onto the 10 leading principal components, consistent with the dimensionality used in the Wasserstein analysis. The two-sample KS statistic is then computed independently for each component $j = 1, \dots, 10$, yielding a p-value p_j for each.

Since 10 simultaneous hypothesis tests are performed, the family wise error rate is controlled through Bonferroni correction:

$$a_{adj} = \frac{0.05}{10} = 0.005 \quad (32).$$

A component is declared drifted if $p_j < a_{adj}$.

The results for each dataset and batch are reported as the number of drifted components out of 10, the fraction drifted, and the minimum observed p-value.

3.8.3. Prediction Drift: Population Stability Index (PSI)

The KS test quantifies shift in the covariate space. PSI complements this by measuring whether the **model's output distribution** has shifted. A more directly actionable signal, as it indicates whether the model is being asked to score compounds in regions of probability space that differ from its training behaviour.

For each dataset and each of the three classifiers, the validation-set predicted probabilities $\hat{p}_{\text{val}}$ serve as the reference distribution, rather than training-set probabilities. This choice is deliberate: using training probabilities as a reference would conflate in-sample fit with distributional shift, producing artificially low PSI values. The validation set, held out during model training, provides an unbiased reference that reflects the model's generalisation behaviour. The batch probabilities $\hat{p}_{\text{test}}$ (Batch A) and $\hat{p}_{\text{ood}}$ (Batch B) are obtained by applying the uncalibrated base models to the respective input matrices.

Both distributions are discretised into $B = 10$ equal-width bins over [0,1]. A small additive constant (10^{-6}) is applied to all bin counts before computing proportions, preventing undefined logarithms in bins that may be empty for one of the two distributions:

$$PSI = \sum_{i=1}^{10}(\hat{p}_{new,i} - \hat{p}_{ref,i}) ln \frac{\hat{p}_{new,i}}{\hat{p}_{ref,i}} \quad (33).$$

PSI is computed independently for each of the three classifiers. The mean PSI across models is reported per dataset as a dataset-level summary statistic, reducing sensitivity to single-model idiosyncrasies. The standard industry thresholds, where the difference is stable if less than 0.10, moderate if between 0.10 and 0.25 and in case of a difference greater than 0.25 the difference is even classified as significant, are applied for interpretation.

3.8.4. Distributional Drift: Wasserstein Distance in PCA Space

The Wasserstein distance is computed in the same PCA space as the KS test, using the 10-component projection fitted on X_{train}. The mean 1-Wasserstein distance across the 10 components is reported as a single scalar per batch:

$$W = \frac{1}{10}\sum_{j=1}^{10} W_1(Z_{ref,j}, Z_{new,j}) \quad (34).$$

where each marginal distance W_1 is computed via scipy.stats.wasserstein_distance. Operating in PCA space, rather than on the raw 2048-dimensional fingerprint vectors,

serves two purposes: it reduces the contribution of uninformative near-zero-variance bit dimensions, and it ensures methodological consistency with the covariate drift analysis. The Batch A value establishes a within-distribution baseline for what constitutes negligible transport cost, while the Batch B value characterises the magnitude of displacement introduced by structurally foreign fingerprint vectors.

Chapter 4. Experimental Results and Discussion

4.1.Baseline Model Performance

The baseline predictive performance of LR, SVM, and XGBoost was evaluated on the held-out test set of each dataset. The purpose of this analysis was to assess how well the three classical ML models could discriminate between positive and negative compounds before applying probability calibration or OOD detection. Four complementary evaluation metrics were used: ROC-AUC, PR-AUC, F1-score, and Balanced Accuracy.

Dataset	Model	ROC-AUC	PR-AUC	F1	BalAcc
Tox21	LogReg	0.7520	0.4401	0.3810	0.6987
Tox21	SVM	0.7870	0.4312	0.4776	0.6715
Tox21	XGBoost	0.7462	0.4476	0.4270	0.6950
BBB	LogReg	0.8559	0.9428	0.9091	0.7815
BBB	SVM	0.8773	0.9539	0.9165	0.7833
BBB	XGBoost	0.8595	0.9471	0.8972	0.7704
DILI	LogReg	0.8171	0.7961	0.7500	0.7222
DILI	SVM	0.8040	0.7972	0.7561	0.7222
DILI	XGBoost	0.7693	0.7637	0.7160	0.6806

***Table 6**. Baseline Discriminative Performance on the Test Sets.*

ROC-AUC measure assesses a model's ability to rank positive over negative samples, calculating its performance over all possible thresholds. A value of 0.5 corresponds to random prediction, while values closer to 1.0 indicate stronger discriminative ability. PR-AUC, also referred to as Average Precision, summarizes the Precision–Recall curve. This metric is of particular note when dealing with imbalanced datasets and measures a model's ability to detect positive instances. Higher values are obtained from models that are able to retrieve positive compounds with higher precision at different levels of recall. The F1-score is the harmonic mean of precision and recall, where the metric returns the F1-score at a threshold of 0.5 by default. This score has the property that both false positives and false negatives are negatively affected. However, unlike ROC-AUC and PR-AUC, the F1 score depends on the selected decision threshold. Balanced Accuracy is the average sensitivity and specificity. This metric is more suitable for imbalanced datasets than accuracy as it treats both classes equally.

Overall, the results show that all three classical ML models achieved meaningful predictive performance across the three ADME-Tox endpoints. However, performance varied substantially across datasets. The BBB_Martins dataset was the easiest prediction task with all models achieving ROC-AUC values above 0.85 and PR-AUC values above 0.94. DILI showed intermediate performance, while Tox21 was the most challenging dataset due to its severe class imbalance, with only approximately 4.3% positive samples.

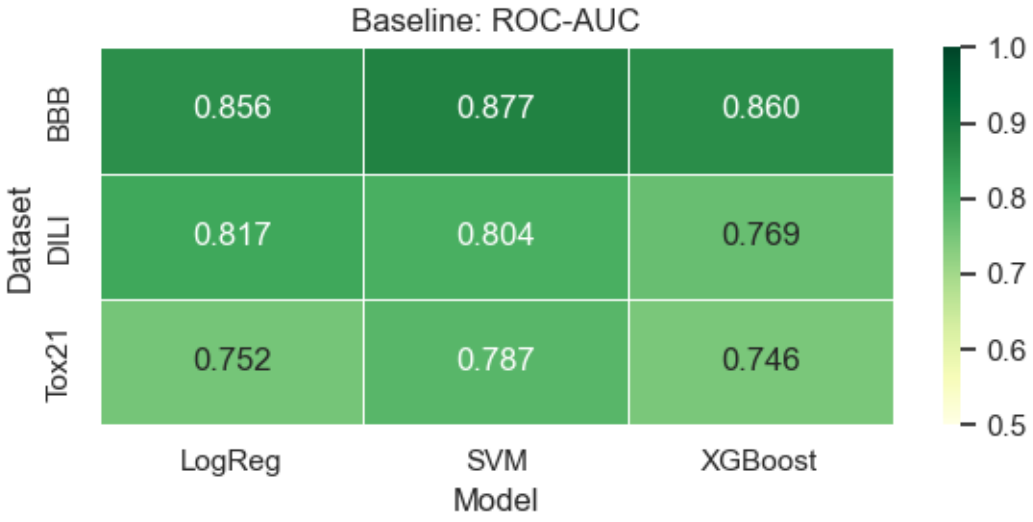


Figure 16. *ROC-AUC Heatmap*

Figure 16 shows the ROC-AUC heatmap for all data-model combinations. For average ROC-AUC values, BBB_Martins clearly outperforms DILI, whereas the Tox21 achieve the lowest values. The individual models, however, show some differences. The SVM achieves the best discrimination, for the Tox21 as well as for the BBB_Martins datasets, as the highest ROC-AUC values are obtained by this model. In contrast to this for the DILI dataset the simple LR outperforms all other models, whereas the tree-based model XGBoost does not perform better than the simpler models on the fixed ECFP4 fingerprint representation.

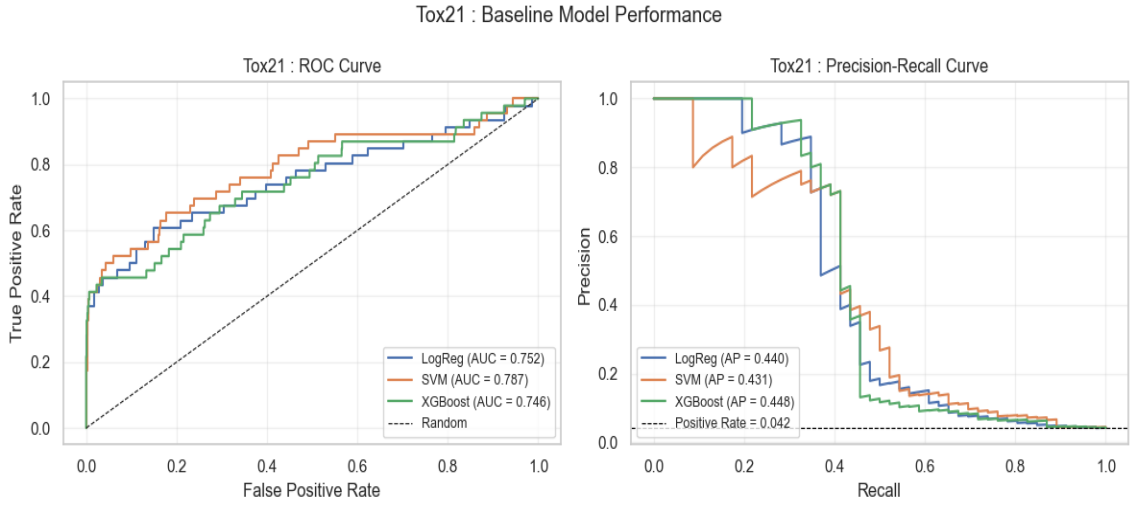


Figure 17. *ROC and Precision-Recall curves for the baseline models on the Tox21/NR-AR dataset*

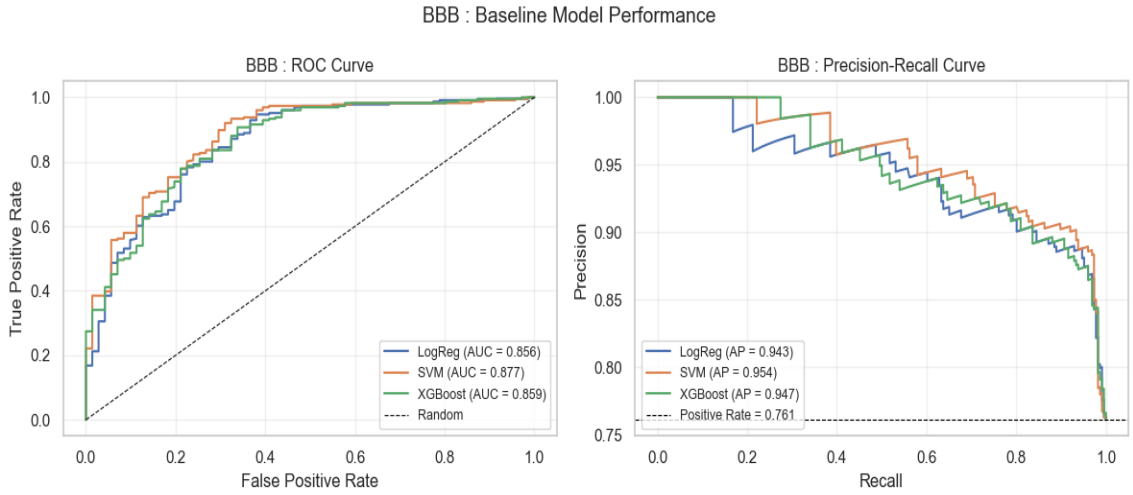


Figure 18. *ROC and Precision-Recall curves for the baseline models on the BBB_Martins dataset*

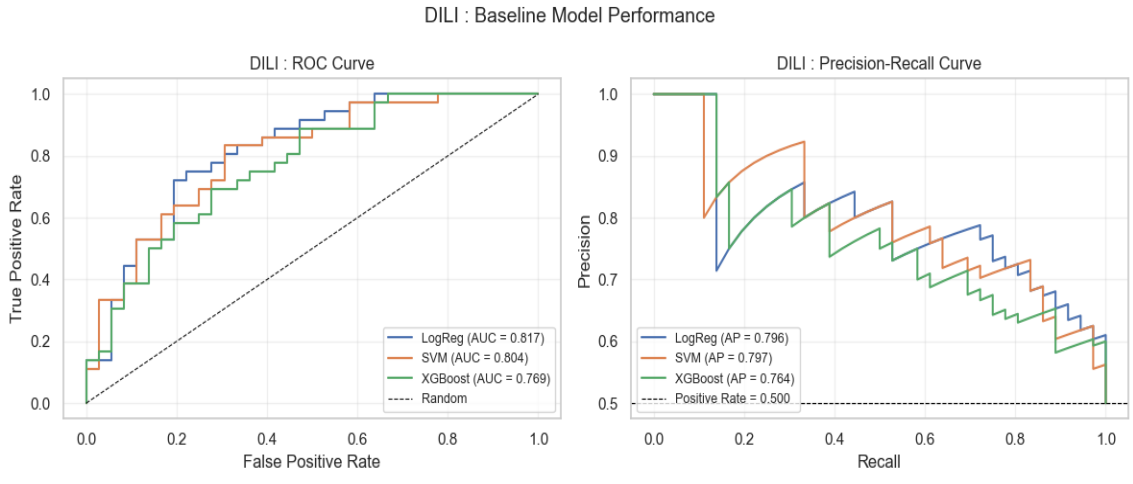


Figure 19. *ROC and Precision-Recall curves for the baseline models on the DILI dataset*

The ROC and Precision–Recall curves further support the numerical results reported in Table 6. In the ROC plots, all models lie clearly above the random baseline, indicating that they provide meaningful discrimination rather than random prediction. In the Precision–Recall plots, the positive-rate line represents the expected precision of a random classifier given the proportion of positive samples in each dataset. Therefore, PR-AUC values should be interpreted relative to this baseline, especially in imbalanced datasets.

For the Tox21/NR-AR endpoint, SVM achieved the highest ROC-AUC value of 0.7870 and the highest F1-score of 0.4776. This suggests that the RBF-kernel SVM was the most effective model for separating active and inactive compounds in this highly imbalanced toxicity endpoint. XGBoost achieved the highest PR-AUC value of 0.4476, which is particularly relevant because PR-AUC is more informative than ROC-AUC under severe class imbalance. LR achieved ROC-AUC = 0.7520 and the highest Balanced Accuracy among the three models, with BalAcc = 0.6987. This indicates that the linear model remained competitive in terms of class-wise recall. However, the relatively low F1-scores across all models reflect the difficulty of identifying the minority positive class at the default decision threshold. Thus, the lower F1 values should not be interpreted as complete model failure, but rather as a consequence of the strong imbalance and the difficulty of detecting rare active compounds.

For the BBB_Martins dataset, all models performed strongly. SVM achieved the best results across all four metrics, with ROC-AUC = 0.8773, PR-AUC = 0.9539, F1-score = 0.9165, and Balanced Accuracy = 0.7833. LR and XGBoost also performed well, with ROC-AUC values of 0.8559 and 0.8595, respectively. The high PR-AUC values observed for BBB_Martins are partly influenced by the majority-positive class distribution, since approximately 76% of the compounds are BBB-positive. For this reason, Balanced Accuracy is particularly important for interpretation. Although all models achieve high F1-scores, the Balanced Accuracy values indicate that the task is still affected by class imbalance and that performance should not be judged by F1-score alone.

For the DILI dataset, LR achieved the highest ROC-AUC value of 0.8171, while SVM achieved the highest PR-AUC value of 0.7972 and the highest F1-score of 0.7561. LR and SVM obtained the same Balanced Accuracy of 0.7222, indicating very similar threshold-based performance. XGBoost performed slightly worse, with ROC-AUC =

0.7693, PR-AUC = 0.7637, F1-score = 0.7160, and Balanced Accuracy = 0.6806. This suggests that, for the small DILI dataset, the more flexible tree-based ensemble model did not generalize as effectively as the simpler linear and kernel-based models. The limited sample size of DILI likely makes complex models more vulnerable to reduced generalization.

Regarding possible underfitting and overfitting effects, the baseline results do not provide direct evidence of severe underfitting, since all models achieve ROC-AUC values clearly above random performance across the three datasets. However, the lower F1-scores observed for Tox21/NR-AR indicate that, although the models can rank compounds reasonably well, the default classification threshold is not optimal for detecting the rare positive class. This behavior is mainly related to class imbalance rather than classical underfitting. Possible overfitting or reduced generalization is more relevant for DILI, where the dataset is small and the endpoint is biologically complex. In this setting, XGBoost performs worse than LR and SVM, despite being the most flexible model. This suggests that simpler and more regularized models can be more reliable than highly flexible models in small ADME-Tox datasets.

In summary, the baseline results demonstrate that classical machine learning models can provide strong predictive performance for ADME-Tox endpoints when combined with ECFP4 molecular fingerprints. However, no single model dominates across all datasets and metrics. SVM appears to be the most robust overall baseline, LR is particularly competitive for small datasets such as DILI, and XGBoost performs well but does not provide a consistent advantage in this experimental setting. These findings justify the subsequent probability calibration analysis, since strong discrimination alone does not guarantee reliable probability estimates.

4.2. Probability Calibration and Reliability Analysis

After evaluating the discriminative performance of the baseline models, the next step was to examine the reliability of their predicted probabilities. A model may achieve high ROC-AUC or PR-AUC values while still producing probability estimates that are not well calibrated. This is particularly important in ADME-Tox prediction, where the numerical probability assigned to a compound may be interpreted as a risk estimate. Therefore, probability calibration was evaluated for all datasets model combinations using the raw model outputs and two post-hoc calibration methods: Platt scaling and isotonic regression.

Dataset	Model	Calibration	ECE	ACE	MCE	Brier	NLL
BBB	LogReg	Isotonic	0.0531	0.0476	0.5398	0.1117	0.3911
BBB	LogReg	Platt	0.0605	0.0559	0.4607	0.1133	0.3865
BBB	LogReg	Raw	0.0867	0.0794	0.4285	0.1191	0.4459
BBB	SVM	Isotonic	0.0672	0.0616	0.1691	0.1104	0.3818
BBB	SVM	Platt	0.0604	0.0602	0.5518	0.1042	0.3601
BBB	SVM	Raw	0.0588	0.0474	0.4084	0.1031	0.3494
BBB	XGBoost	Isotonic	0.0643	0.0702	0.1735	0.1177	0.4453
BBB	XGBoost	Platt	0.0624	0.0641	0.2253	0.1161	0.3887
BBB	XGBoost	Raw	0.0595	0.0630	0.3331	0.1189	0.3937
DILI	LogReg	Isotonic	0.1262	0.1120	0.3431	0.1935	1.9434
DILI	LogReg	Platt	0.1218	0.1530	0.2303	0.1847	0.5542
DILI	LogReg	Raw	0.1277	0.1505	0.2913	0.1896	0.5925
DILI	SVM	Isotonic	0.1183	0.1383	0.3515	0.1959	2.8675
DILI	SVM	Platt	0.0823	0.1303	0.2413	0.1872	0.5561
DILI	SVM	Raw	0.1401	0.1504	0.3847	0.1932	0.5787
DILI	XGBoost	Isotonic	0.1767	0.1460	0.3333	0.2111	1.1471
DILI	XGBoost	Platt	0.1481	0.1617	0.3765	0.2110	0.6170
DILI	XGBoost	Raw	0.1814	0.1716	0.3898	0.2147	0.6419
Tox21	LogReg	Isotonic	0.0055	0.0142	0.5716	0.0297	0.1912
Tox21	LogReg	Platt	0.0134	0.0102	0.4140	0.0327	0.1388
Tox21	LogReg	Raw	0.0518	0.0641	0.7556	0.0523	0.2140
Tox21	SVM	Isotonic	0.0021	0.0113	0.5283	0.0290	0.1284
Tox21	SVM	Platt	0.0065	0.0122	0.7580	0.0295	0.1300
Tox21	SVM	Raw	0.0072	0.0138	0.2721	0.0297	0.1299
Tox21	XGBoost	Isotonic	0.0049	0.0120	0.8398	0.0284	0.1496
Tox21	XGBoost	Platt	0.0110	0.0179	0.3438	0.0299	0.1338
Tox21	XGBoost	Raw	0.0852	0.0864	0.6476	0.0471	0.1958

Table 7*. Calibration and Reliability Metrics before and after Post-hoc Calibration.*

Table 7 summarizes the calibration and reliability metrics for each dataset, model and calibration variant. The metrics reported are Expected Calibration Error (ECE), Adaptive Calibration Error (ACE), Maximum Calibration Error (MCE), Brier Score, and Negative Log-Likelihood (NLL). For all these metrics, lower values indicate better probability reliability. ECE measures the average gap between predicted confidence and observed frequency across probability bins, while ACE performs a similar calculation using adaptive equal frequency bins. MCE captures the worst calibration gap across bins and is therefore more sensitive to sparse or unstable regions. The Brier Score measures the mean squared error between predicted probabilities and true binary levels, while NLL evaluates the likelihood assigned to the true labels and strongly penalizes overconfident wrong predictions.

Overall, the calibration results show the post-hoc calibration improves probability reliability in several cases, but its effect is not uniform across datasets and models. The benefit of calibration depends strongly on the dataset size, class imbalance, and the initial calibration quality of the model. Specifically:

- For the Tox21/ NR-AR dataset, calibration produced the clearest improvement overall. LR improved from ECE = 0.0518 in the raw model to 0.0134 after Platt scaling and 0.0055 after isotonic regression. XGBoost showed an even larger improvement, with ECE decreasing from 0.0852 to 0.0110 after Platt scaling and to 0.0049 after isotonic regression. These reductions indicate that the raw LR and XGBoost probability estimates were initially miscalibrated and that post-hoc calibration substantially improved the agreement between predicted probabilities and observed positive frequencies. SVM behaved differently. It was already relatively well calibrated before post-hoc calibration, with raw ECE = 0.0072 and raw NLL = 0.1299. Isotonic regression further reduced ECE to 0.0021 and slightly improved Brier Score and NLL, whereas Platt scaling produced only a marginal ECE improvement and nearly unchanged NLL. However, the MCE increased after calibration, from 0.2721 in the raw SVM to 0.7580 after Platt scaling and 0.5283 after isotonic regression. This indicates that although the average calibration error improved, some individual probability bins became less reliable after calibration. This behaviour is plausible for Tox21/NR-AR because the dataset is severely imbalanced, with very few positive samples, making worst-case bin-based metrics such as MCE unstable. Overall, calibration was clearly beneficial for LR and XGBoost, while the improvement for SVM was more limited because the raw SVM probabilities were already well calibrated. The results also show that the lowest ECE does not always imply uniformly better probability quality across all metrics. Platt scaling produced lower NLL than isotonic regression for LR and XGBoost, while isotonic regression produced the lowest ECE values. Therefore, calibration performance should be assessed jointly using ECE, Brier Score, NLL, and MCE rather than relying on a single metric.
- For the BBB_Martins dataset, the effect of calibration was more mixed. LR benefited from calibration, with ECE decreasing from 0.0867 in the raw model to 0.0605 after Platt scaling and 0.0531 after isotonic regression. However, SVM and XGBoost were already relatively well calibrated before post-hoc calibration. For SVM, the raw model achieved the lowest ECE, Brier Score, and NLL among its calibration variants. This can be partly explained by the fact that SVC(probability=True) already applies an internal Platt-scaling procedure to produce probability estimates; therefore, the additional post-hoc calibration applied here acts as re-calibration. For XGBoost, the raw model had the lowest ECE, while Platt scaling slightly improved Brier Score and NLL. These results suggest that, for BBB_Martins, post-hoc calibration is useful mainly for LR, while the other models do not consistently benefit from additional calibration.
- For the DILI dataset, calibration was more unstable. This is expected because DILI is the smallest dataset in the study, and the validation and test sets contain a limited number of compounds. Platt scaling provided the most stable improvement, especially for SVM, where ECE decreased from 0.1401 to 0.0823. XGBoost also improved from ECE = 0.1814 to 0.1481 after Platt scaling. LR showed only a small improvement. In contrast, isotonic regression produced

problematic NLL values for all three models, particularly for LR and SVM. This indicates that isotonic regression may overfit the small validation set and assign overly confident probabilities to some test samples. Therefore, although isotonic regression sometimes reduces ECE, it is not the most reliable calibration method for DILI. Platt scaling is preferable for this dataset because it is parametric and more stable under small-sample conditions.

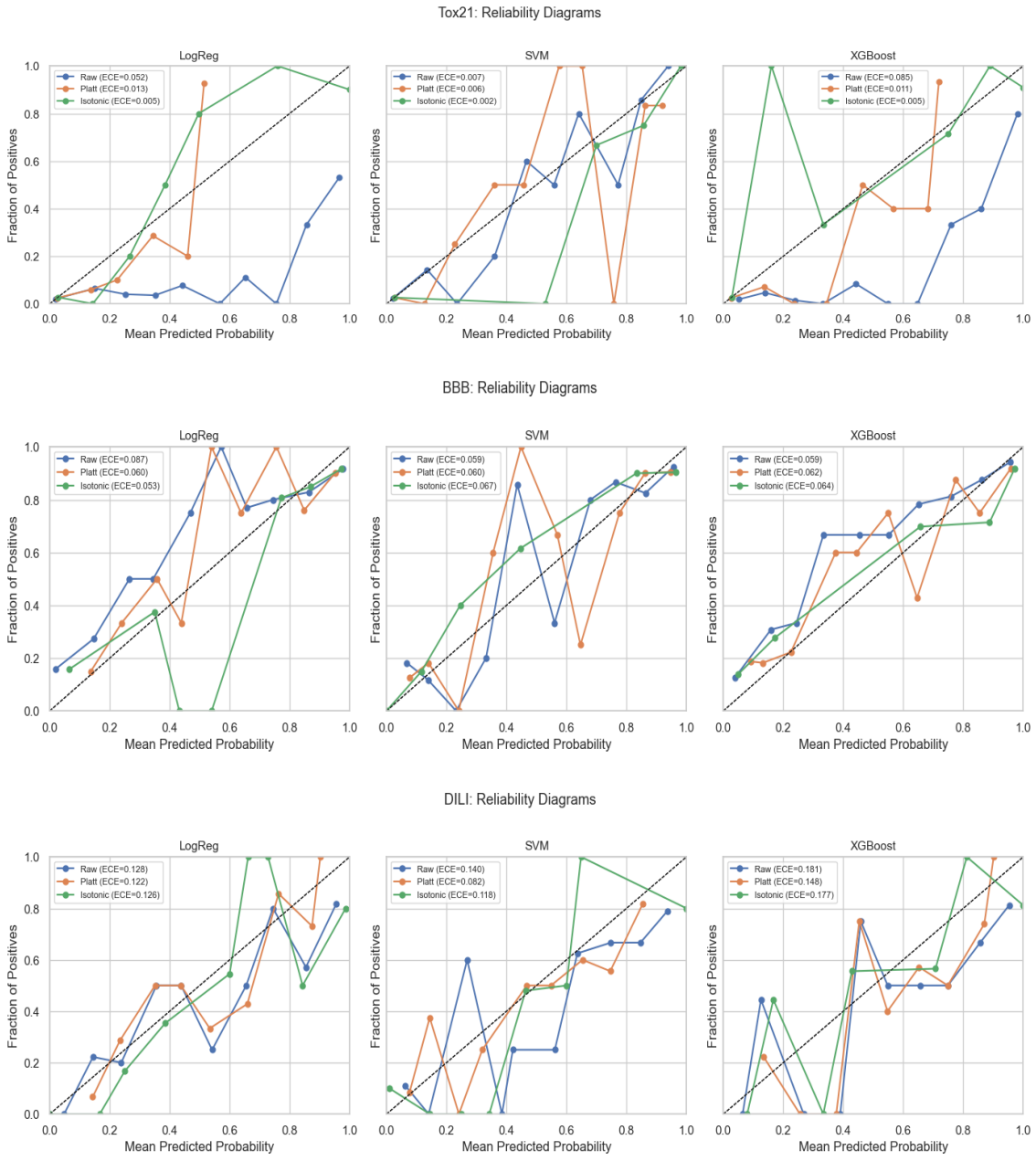


Figure 20. *Reliability diagrams for the dataset.*

Figures 20 present the reliability diagrams for the three datasets. In these plots, the diagonal line represents perfect calibration, where the predicted probability is equal to the observed fraction of positives. Points below the diagonal indicate overconfidence, while

points above the diagonal indicate underconfidence. Therefore, curves closer to the diagonal correspond to more reliable probability estimates.

For the Tox21 dataset, the reliability diagrams are highly irregular, especially for LR and XGBoost. This behavior is expected because Tox21 is severely imbalanced and contains very few positive samples. As a result, several probability bins include only a small number of positive compounds, making the observed fraction of positives unstable. The raw LR and XGBoost curves deviate substantially from the diagonal, mainly indicating overconfident probability estimates in several probability regions. After calibration, especially with isotonic regression, the curves move closer to the diagonal in terms of average calibration error, which is consistent with the strong reduction in ECE. However, the curves remain noisy, and for SVM the MCE increases after calibration. This shows that, although calibration improves the average reliability of the probabilities, some individual bins can still become unstable due to the severe class imbalance.

For the BBB_Martins dataset, the reliability diagrams are more stable than those of Tox21. This is consistent with the larger number of positive compounds and the stronger baseline performance observed for this endpoint. LR benefits from calibration, as its calibrated curves become closer to the diagonal in several regions and its ECE decreases. In contrast, SVM and XGBoost already produce relatively reliable raw probabilities. For SVM, the raw probability curve and the numerical metrics indicate the additional post-hoc calibration does not clearly improve reliability. This is also expected because SVC(probability=True) already applies an internal Platt-scaling procedure. For XGBoost, the raw and calibrated curves remain relatively close, confirming that the effect of calibration on BBB_Martins is more limited and model dependent.

For the DILI dataset, the reliability diagrams also show irregular behavior, but for a different reason. Unlike Tox21/NR-AR, DILI is not severely imbalanced. However, it is the smallest dataset in the study. The limited number of validation and test samples makes the bin-wise estimates of the observed positive fraction less stable. Platt scaling provides the most consistent improvement, especially for SVM, where the ECE is substantially reduced. In contrast, isotonic regression appears less stable and produces large NLL values, suggesting that it may overfit the small calibration set and assign overly confident probabilities to some test samples. Therefore, for DILI, the reliability diagrams should be

interpreted together with the numerical metrics, and Platt scaling appears to be the more reliable calibration method.

Overall, the reliability diagrams confirm that probability calibration must be evaluated using both visual and numerical evidence. Tox21/NR-AR shows strong average improvement after calibration, but its curves remain noisy because of severe class imbalance. BBB_Martins shows relatively stable probability estimates, and calibration is mainly beneficial for LR. DILI shows instability caused by small sample size, with Platt scaling being more robust than isotonic regression.

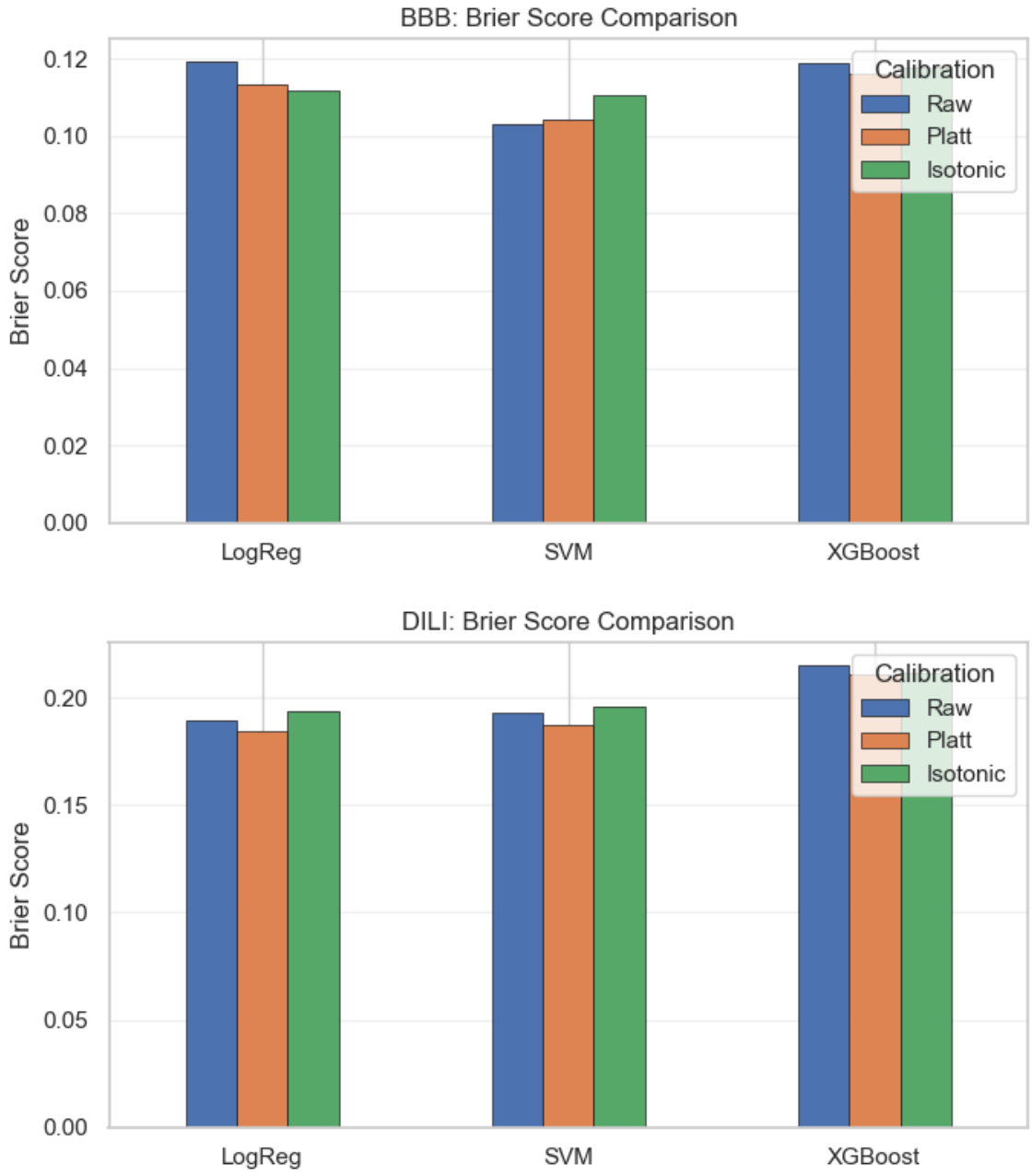


Figure 21. *Brier Score comparison*

The Brier Score measures "the mean squared error between the predicted probabilities and the true binary labels". Lower values indicate better probabilistic predictions. Unlike ECE, which summarizes calibration error across probability bins, the Brier Score evaluates the overall accuracy of the predicted probabilities at the individual-sample level.

For Tox21, calibration substantially reduces the Brier Score for LR and XGBoost, confirming that post-hoc calibration improves not only bin-wise calibration error but also the overall quality of their probability estimates. For SVM, the raw Brier Score is already

low and changes only slightly after calibration, which is consistent with the observation that the raw SVM probabilities were already reasonably calibrated.

For BBB_Martins, the Brier Score results are more mixed. LR improves after calibration, while SVM performs best in its raw form. For XGBoost, Platt scaling gives a small improvement compared with the raw probabilities. This confirms that calibration is model-dependent and does not necessarily improve all models when the raw probabilities are already reliable.

For DILI, the Brier Scores are higher overall, reflecting the greater uncertainty of this endpoint and the instability caused by the small dataset size. Platt scaling provides the most stable improvement for LR and SVM, whereas isotonic regression does not provide a consistent benefit and may worsen the Brier Score. These findings are consistent with the NLL results and support the conclusion that isotonic regression is less reliable for DILI due to possible overfitting.

4.3. OOD Detection and Applicability Domain Analysis

After evaluating discriminative performance and probability calibration, the next step was to examine whether the models can identify samples that fall outside the training distribution. This is important in ADME-Tox prediction because a model may produce confident predictions for compounds that are chemically different from the training data. Such cases are potentially unreliable even when the predicted probability appears high. Therefore, two OOD scoring methods were evaluated: a PCA-based Mahalanobis distance and a k-nearest-neighbour (kNN) mean distance in the original fingerprint space. In both methods, higher scores are intended to indicate a higher likelihood of being out-of-distribution.

For each dataset, OOD thresholds were defined as the 95th percentile of the training OOD scores. This means that approximately 5% of the training samples would be expected to fall above the threshold by construction. In addition to the held-out test set, a synthetic OOD batch of 500 random binary fingerprint vectors was generated for each dataset. These vectors were density-matched to the corresponding training fingerprints, with train and simulated densities of 0.0144 and 0.0145 for Tox21/NR-AR, 0.0209 and 0.0209 for BBB_Martins, and 0.0194 and 0.0195 for DILI. This design prevents trivial OOD

detection based only on fingerprint sparsity and instead tests whether the OOD methods can detect lack of realistic fingerprint structure.

Dataset	Scorer	Threshold (95^{th})	Test flagged	OOD sim flagged	Detection AUROC
Tox21	Mahalanobis	10.318	2.7%	0%	0.0250
Tox21	kNN	6.632	5.4%	1.6%	0.8108
BBB	Mahalanobis	9.992	4.7%	0%	0.0209
BBB	kNN	7.014	6.1%	28.6%	0.8816
DILI	Mahalanobis	10.961	1.4%	0%	0.0260
DILI	kNN	8.348	2.8%	0%	0.6748

***Table 8**. OOD Detection Summary Table*

Table 8 summarizes the OOD detection results. The held-out test sets were only rarely flagged as OOD, with kNN flagging 5.4% of Tox21/NR-AR, 6.1% of BBB_Martins, and 2.8% of DILI test compounds. This indicates that most test compounds remain within the estimated applicability domain of the training data. The Mahalanobis method flagged even fewer test samples, ranging from 1.4% to 4.7%. However, the behavior of the two OOD scorers was very different for the simulated OOD batches. The Mahalanobis-based OOD scorer failed to detect the simulated OOD fingerprints. It assigned lower scores to the simulated OOD vectors than to the in-distribution test samples, producing AUROC values close to zero for all datasets: 0.0250 for Tox21/NR-AR, 0.0209 for BBB_Martins, and 0.0260 for DILI. This does not indicate successful OOD detection; instead, it indicates an inverse ranking. In other words, the Mahalanobis score considered the synthetic OOD fingerprints to be closer to the training distribution than the real held-out test compounds. This result suggests that the PCA-based global Gaussian approximation is not suitable for this type of sparse, density-matched synthetic fingerprint OOD data. In contrast, the kNN distance provided a much more meaningful OOD signal. The kNN OOD AUROC was 0.8108 for Tox21/NR-AR and 0.8816 for BBB_Martins, indicating good separation between in-distribution and simulated OOD samples. For DILI, the kNN AUROC was lower, at 0.6748, suggesting only moderate separation. This weaker performance is expected because DILI is the smallest dataset, and the training chemical space is less densely sampled. The kNN method is therefore more appropriate in this setting because it measures local distance from real training compounds rather than distance from a global distributional center.

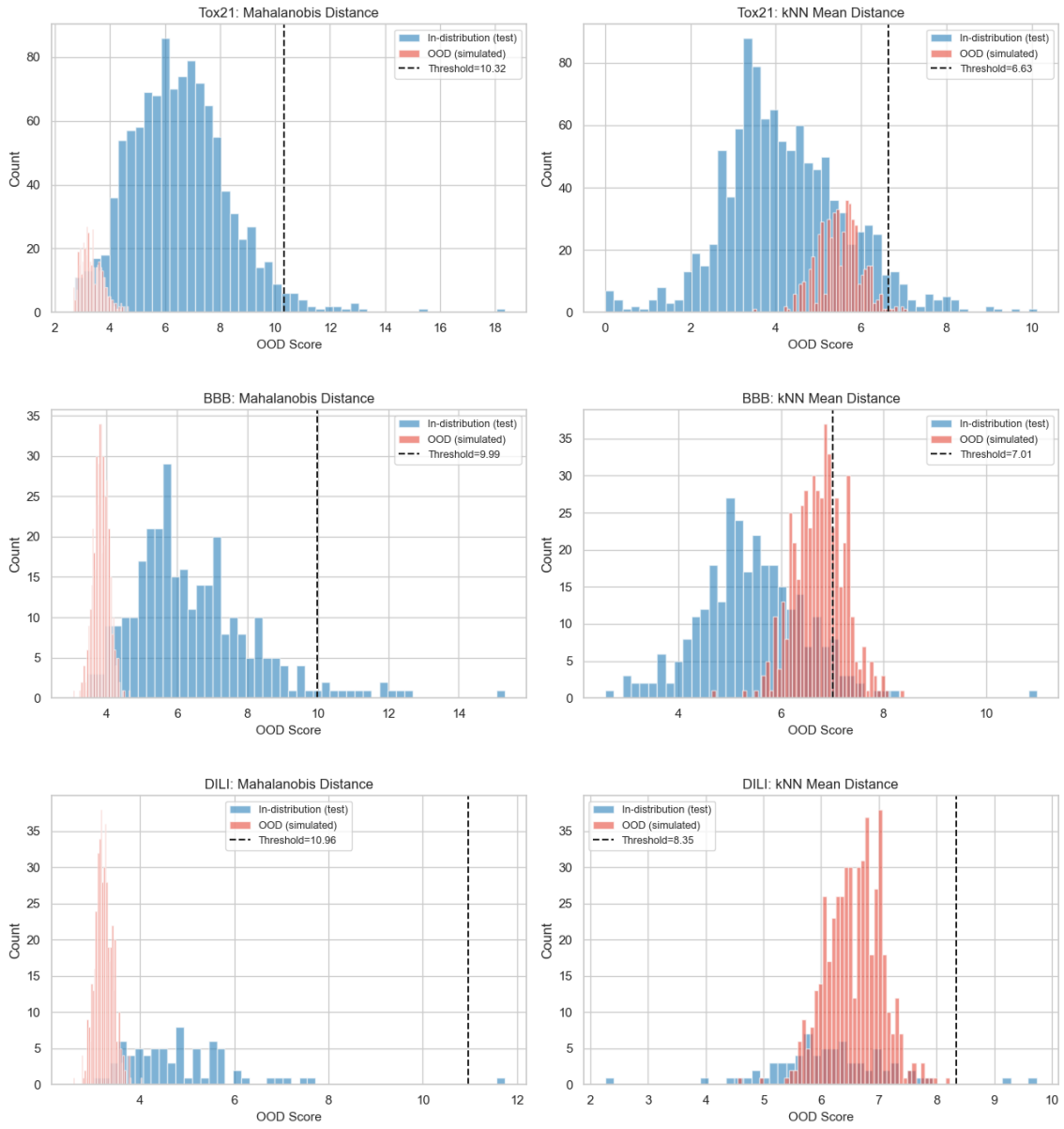
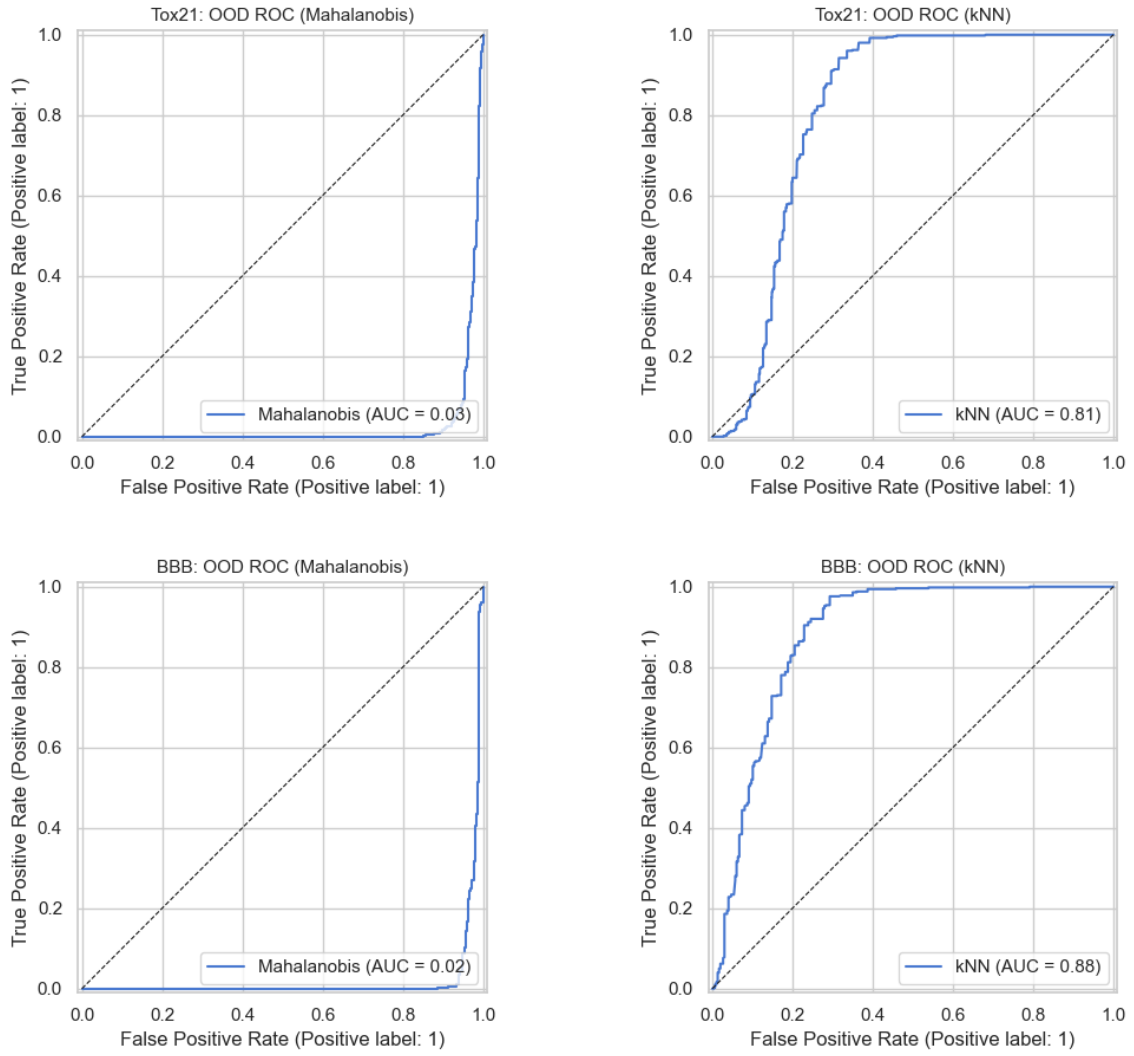


Figure 22. *OOD Score Distributions*

Figure 22 shows the OOD score distributions for Mahalanobis and kNN distance. For the Mahalanobis scorer, the simulated OOD distributions are shifted toward lower scores than the in distribution test samples. This explains why the Mahalanobis AUROC values are close to zero and why no simulated OOD samples are flagged by the 95th percentile threshold. The result is especially clear in BBB_Martins and DILI, where the synthetic OOD scores form a narrow distribution below the real test scores. Therefore, Mahalanobis distance should not be interpreted as a reliable OOD detector in the present setup. The kNN score distributions show a different pattern. For Tox21/NR-AR and BBB_Martins, the simulated OOD scores are shifted toward higher kNN distances compared with the test samples, indicating that the synthetic vectors are farther from real training neighbors.

This confirms that local neighbourhood distance captures OOD structure better than the global Mahalanobis distance. In DILI, the kNN test and OOD score distributions overlap more strongly, which explains the lower AUROC. The conservative 95th-percentile threshold also means that many simulated OOD samples are not flagged, even when their score distribution is shifted relative to the test set. Therefore, the AUROC is more informative than the single threshold count for comparing the two OOD methods.



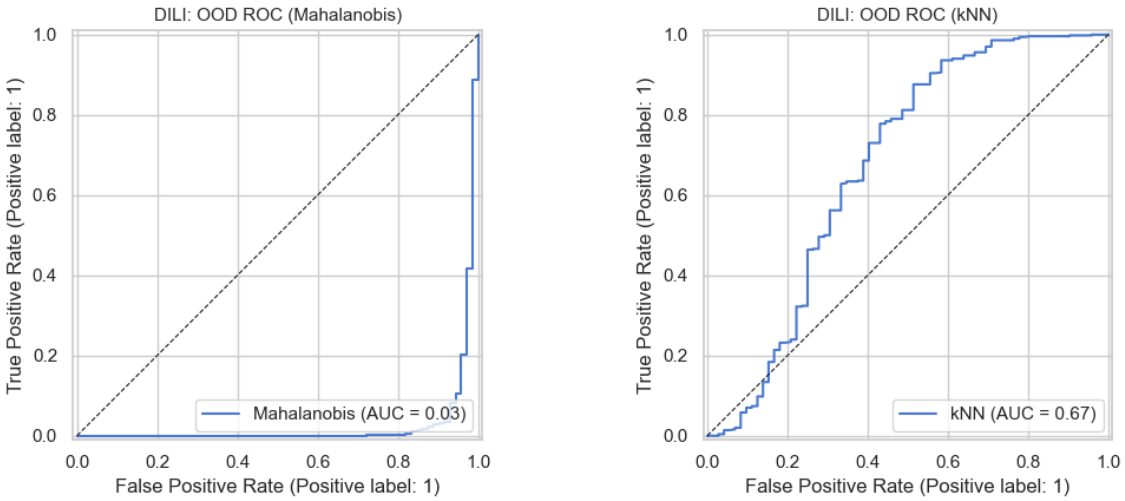


Figure 23. *OOD ROC curves*

Figure 23 present the ROC curves for the OOD detection task, where held-out test samples are treated as in-distribution and simulated fingerprint vectors are treated as OOD. These curves confirm the contrast between the two methods. The Mahalanobis ROC curves lie far below the random diagonal, reflecting the inverse ranking discussed above. In contrast, the kNN ROC curves are clearly above the random line for Tox21/NR-AR and BBB_Martins and moderately above it for DILI. This provides strong evidence that kNN distance is the better OOD scoring method for the sparse ECFP4 fingerprint representation used in this thesis.

Overall, the OOD analysis shows that the two applicability-domain methods behave very differently. The PCA-based Mahalanobis distance is not suitable for detecting density-matched synthetic OOD fingerprints, because it ranks these samples as less OOD than the real test compounds. In contrast, kNN distance provides a more reliable local OOD signal, particularly for Tox21/NR-AR and BBB_Martins. The results support the use of local distance-based applicability-domain methods for sparse molecular fingerprints and show that OOD analysis is necessary alongside classification performance and probability calibration.

4.4. Drift Monitoring Analysis

The final part of the experimental evaluation focuses on drift monitoring. While OOD detection evaluates whether individual samples fall outside the applicability domain, drift monitoring evaluates whether the distribution of an entire incoming batch has shifted

relative to the reference training distribution. This is important in practical ADME-Tox prediction, where models may be deployed on future chemical libraries whose distribution may differ from the data used for model development. Two monitoring batches were evaluated. Batch A corresponds to the held out in distribution test set, while Batch B corresponds to the simulated OOD fingerprint batch generated in the previous section. Batch A is expected to remain close to the training distribution, whereas Batch B is expected to produce a clear distributional shift. Three complementary drift indicators were used. The Kolmogorov-Smirnov test across PCA components, the Population Stability Index (PSI) on predicted probabilities, and the mean Wasserstein distance across PCA components.

Dataset	**KS drifted (Batch A)**	**KS drifted (Batch B)**	**PSI mean (Batch A)**	**PSI mean (Batch B)**	**Wasserstein (Batch A)**	**Wasserstein (Batch B)**
Tox21	0/10	10/10	0.0167	0.3172	0.0263	0.6063
BBB	0/10	10/10	0.0697	4.6574	0.0682	0.7096
DILI	0/10	10/10	0.2829	7.5403	0.1975	0.6964

***Table 9**. Drift Monitoring Summary Table*

Table 9 summarizes the drift monitoring results. The KS test shows a very clear separation between the two batches. For Batch A, none of the first ten PCA components were flagged as drifted in any dataset. In contrast, Batch B was flagged as drifted in all ten PCA components for Tox21/NR-AR, BBB_Martins, and DILI. This indicates that the simulated OOD batches introduce a strong covariate shift relative to the training data, while the held-out test sets remain distributionally close to the reference training sets. The PSI results provide complementary evidence at the model-output level. For Tox21/NR-AR and BBB_Martins, Batch A has low mean PSI values of 0.0167 and 0.0697, respectively, indicating stable predicted-probability distributions. For DILI, Batch A has a higher mean PSI value of 0.2829, which exceeds the common significant-drift threshold of 0.25. This constitutes a false alarm of the monitoring framework and should be interpreted cautiously because DILI is the smallest dataset and predicted-probability distributions can vary substantially between validation and test splits due to limited sample size. In contrast, Batch B produces clearly elevated PSI values in all datasets, with mean PSI values of 0.3172 for Tox21/NR-AR, 4.6574 for BBB_Martins, and 7.5403 for DILI. These values indicate strong output-distribution drift for the simulated OOD batches. The Wasserstein distance results further support the same conclusion. Batch A

remains close to the training distribution, with distances of 0.0263 for Tox21/NR-AR, 0.0682 for BBB_Martins, and 0.1975 for DILI. Batch B shows much larger distances in all datasets: 0.6063 for Tox21/NR-AR, 0.7096 for BBB_Martins, and 0.6964 for DILI. Therefore, the simulated OOD batches are consistently farther from the training distribution than the held-out test batches.

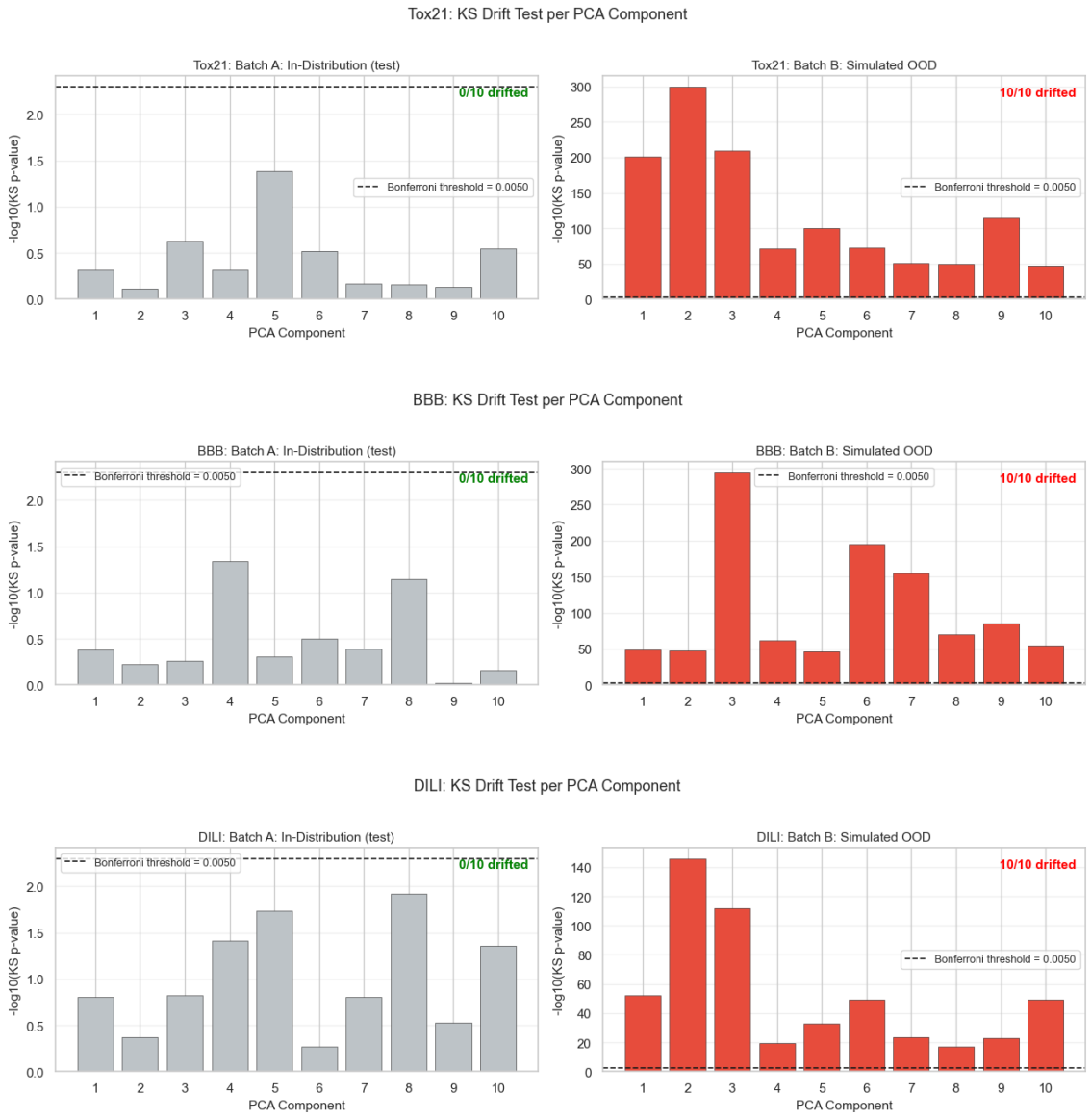
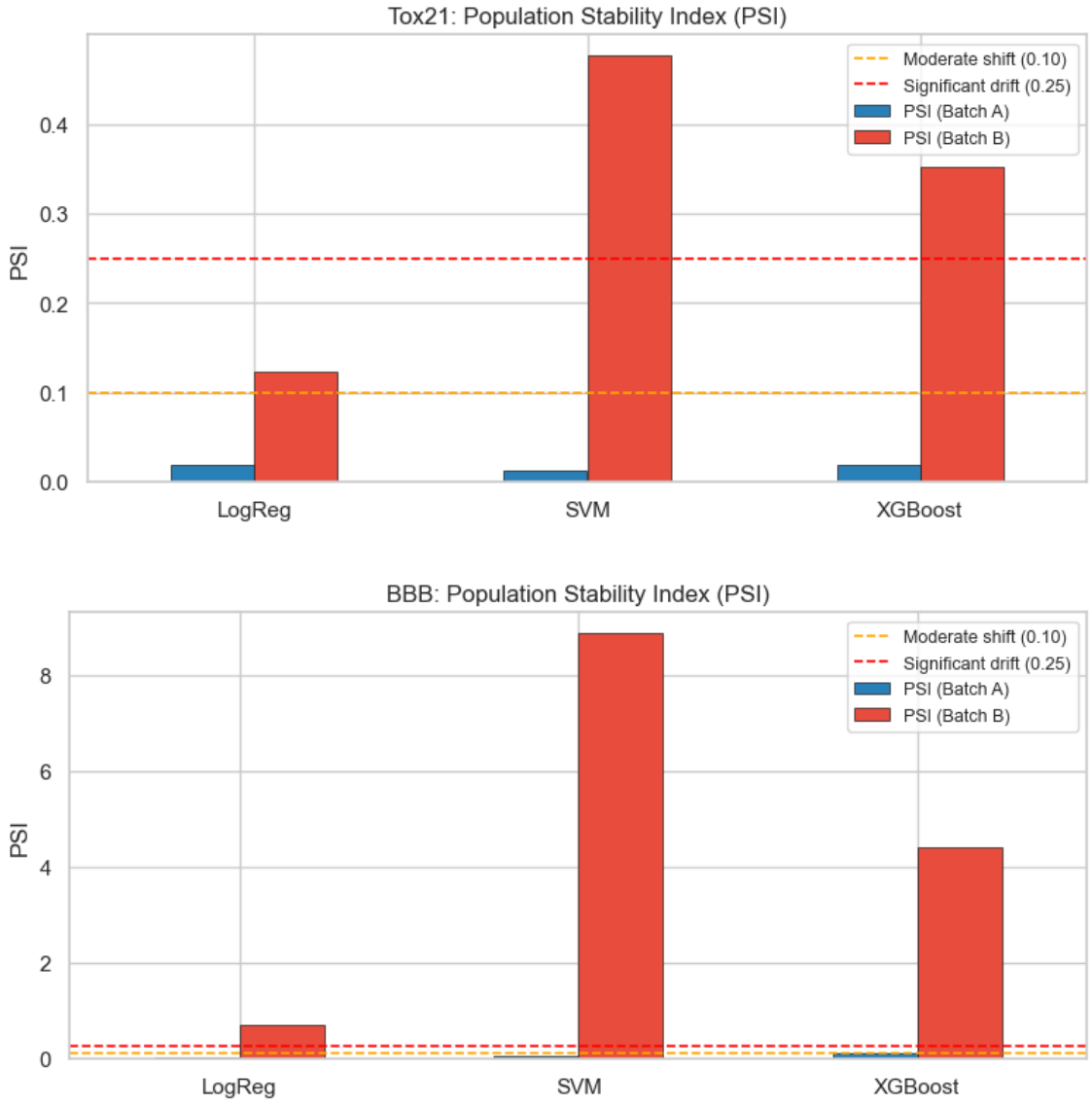


Figure 24. *KS drift test per PCA component*

The y-axis reports -log10(KS p-value), so higher bars indicate stronger statistical evidence for drift. The dashed horizontal line corresponds to the Bonferroni-corrected significance threshold. For Batch A, all bars remain below the threshold across the three datasets, resulting in 0/10 drifted PCA components. For Batch B, all ten PCA components exceed the threshold in every dataset, resulting in 10/10 drifted components. This

confirms that the KS test successfully identifies the simulated OOD batches as distributionally shifted while not falsely flagging the held-out in-distribution test sets.



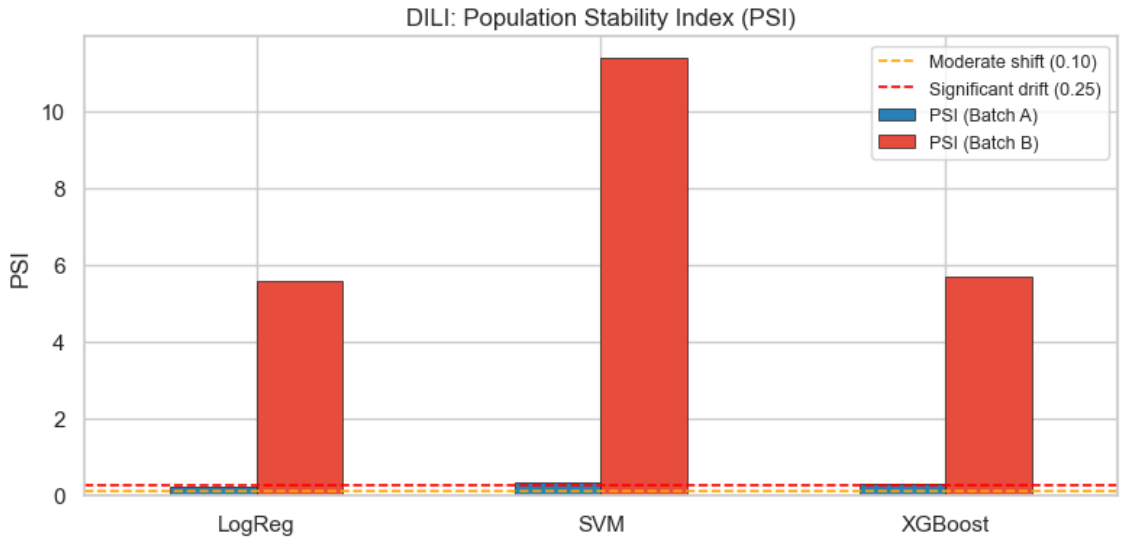


Figure 25. *Population Stability Index*

Figure 25 present the PSI results. The blue bars correspond to Batch A and the red bars correspond to Batch B. The dashed horizontal lines indicate commonly used interpretation thresholds: 0.10 for moderate shift and 0.25 for significant drift. For Tox21/NR-AR, Batch A remains stable for all models, while Batch B shows moderate to significant drift depending on the model. For BBB_Martins, Batch A is mostly stable, although XGBoost shows a small moderate shift. Batch B produces very large PSI values, especially for SVM and XGBoost, indicating a major change in the distribution of model outputs. For DILI, Batch A already shows elevated PSI values, which likely reflects the small size of the dataset and split-to-split instability. However, Batch B is much larger for all three models, confirming severe output-distribution drift.

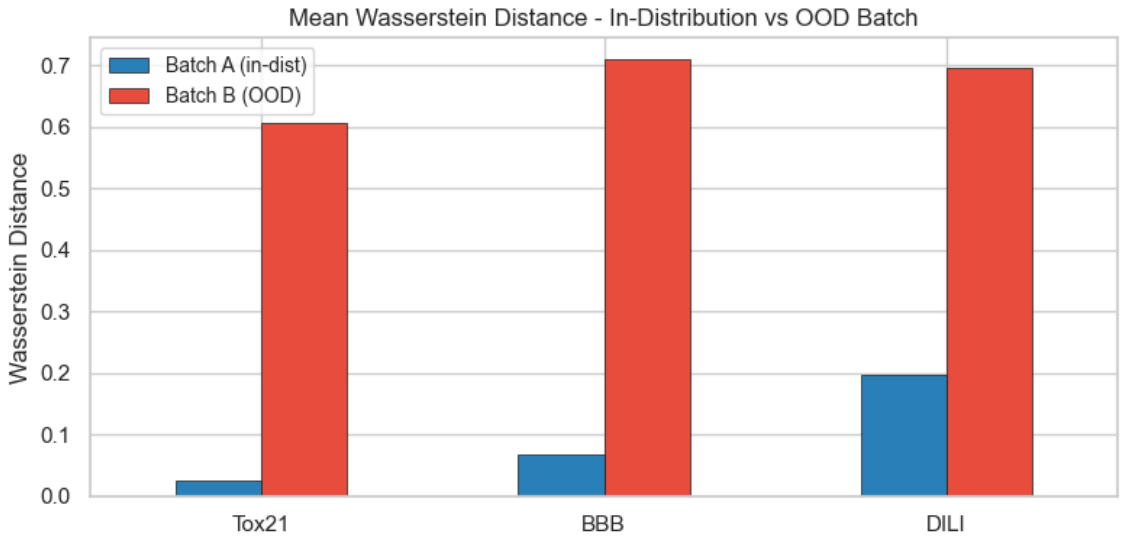


Figure 26. *Mean Wasserstein Distance for Batch A and Batch B across datasets*

Across all datasets, Batch B has substantially larger Wasserstein distance than Batch A. This confirms that the simulated OOD fingerprints are distributionally shifted in the PCA-transformed fingerprint space. DILI also shows the largest Batch A Wasserstein distance, which is consistent with its small sample size and greater variability between train and test splits. Nevertheless, the Batch B distance remains much higher than Batch A for all datasets.

Overall, the drift monitoring analysis confirms that the proposed monitoring framework can distinguish stable in-distribution batches from strongly shifted OOD-like batches. The KS test provides clear evidence of covariate drift in the synthetic OOD batches, PSI detects changes in the model-output distributions, and Wasserstein distance quantifies the magnitude of the distributional shift in PCA space. The only caution concerns DILI, where Batch A already shows elevated PSI and Wasserstein values due to small-sample instability. Taken together, these results show that drift monitoring provides complementary information to OOD detection and probability calibration and is necessary for assessing the reliability of ADME-Tox models under changing input distributions.

4.5. Conclusions and Future Work

Overall, the results demonstrate that reliable ADME-Tox prediction requires more than good classification performance. Although the baseline models achieved meaningful ROC-AUC, PR-AUC, and F1-score values, the additional analyses demonstrated that probability calibration, out-of-distribution detection, and drift monitoring provide essential complementary information. In particular, calibration improved probability reliability mainly for Tox21/NR-AR and for Logistic Regression on BBB_Martins, while the DILI dataset showed greater instability due to its limited sample size. The OOD analysis showed that kNN distance was more suitable than Mahalanobis distance for sparse ECFP4 fingerprints, and the drift monitoring experiments successfully distinguished stable in-distribution batches from strongly shifted simulated OOD batches.

These findings suggest that classical machine learning models with ECFP4 fingerprints can serve as strong and interpretable baselines for ADME-Tox modelling, but their predictions should not be evaluated only through discrimination metrics. Reliability

checks are necessary before such models can be used in practical chemical decision-making.

Future work could extend this study in several directions. First, ECFP4 fingerprints could be combined with physicochemical descriptors to provide richer molecular representations. Second, graph-based models such as Graph Neural Networks could be evaluated, since they operate directly on molecular graphs and may capture structural information that fixed fingerprints cannot fully represent. Third, more realistic OOD and drift scenarios should be tested using external chemical libraries, time-split datasets, or compounds from different experimental sources. Finally, uncertainty-aware methods such as conformal prediction, deep ensembles, or Bayesian approaches could be explored to provide more robust estimates of prediction reliability.

References

[1] T. T. Van Tran, A. Surya Wibowo, H. Tayara, and K. T. Chong, “Artificial Intelligence in Drug Toxicity Prediction: Recent Advances, Challenges, and Future Perspectives,” *J. Chem. Inf. Model.*, vol. 63, no. 9, pp. 2628–2643, May 2023, doi: 10.1021/acs.jcim.3c00200.

[2] J. Zhang *et al.*, “Computational toxicology in drug discovery: applications of artificial intelligence in ADMET and toxicity prediction,” Sep. 01, 2025, *Oxford University Press*. doi: 10.1093/bib/bbaf533.

[3] A. Vashishat, P. Patel, G. Das Gupta, and B. Das Kurmi, “Alternatives of Animal Models for Biomedical Research: a Comprehensive Review of Modern Approaches,” *Stem Cell Rev. Rep.*, vol. 20, no. 4, pp. 881–899, May 2024, doi: 10.1007/s12015-024-10701-x.

[4] T. Rudroff, “Artificial Intelligence as a Replacement for Animal Experiments in Neurology: Potential, Progress, and Challenges,” *Neurol. Int.*, vol. 16, no. 4, pp. 805–820, Aug. 2024, doi: 10.3390/neurolint16040060.

[5] A. S. Bhatia, M. K. Saggi, and S. Kais, “Quantum Machine Learning Predicting ADME-Tox Properties in Drug Discovery,” *J. Chem. Inf. Model.*, vol. 63, no. 21, pp. 6476–6486, Nov. 2023, doi: 10.1021/acs.jcim.3c01079.

[6] T. Hastie, R. Tibshirani, and J. Friedman, “Springer Series in Statistics The Elements of Statistical Learning Data Mining, Inference, and Prediction.”

[7] J. Platt, “Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods,” *Adv. Large Margin Classif.*, vol. 10, Jun. 2000.

[8] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[9] Z. Wu *et al.*, “MoleculeNet: A benchmark for molecular machine learning,” *Chem. Sci.*, vol. 9, no. 2, pp. 513–530, 2018, doi: 10.1039/c7sc02664a.

[10] K. Huang *et al.*, “Therapeutics Data Commons: Machine Learning Datasets and Tasks for Drug Discovery and Development,” Aug. 2021, [Online]. Available: http://arxiv.org/abs/2102.09548

[11] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On Calibration of Modern Neural Networks,” 2017.

[12] L. L. G. Ferreira and A. D. Andricopulo, “ADMET modeling approaches in drug discovery,” May 01, 2019, *Elsevier Ltd*. doi: 10.1016/j.drudis.2019.03.015.

[13] J. G. Topliss, “Utilization of operational schemes for analog synthesis in drug design,” *J. Med. Chem.*, vol. 15, no. 10, pp. 1006–1011, Oct. 1972, doi: 10.1021/jm00280a002.

[14] C. A. Lipinski, B. W. Dominy, and P. J. Feeney, “drug delivery reviews Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings,” 1997.

[15] A. Lavecchia, “Machine-learning approaches in drug discovery: Methods and applications,” 2015, *Elsevier Ltd*. doi: 10.1016/j.drudis.2014.10.012.

[16] Sakai JB., “Pharmacokinetics: The Absorption, Distribution, and Excretion of Drugs. In: Practical Pharmacology for the Pharmacy Technician.,” pp. 27–40, 2009.

[17] M. P. Doogue and T. M. Polasek, “The ABCD of clinical pharmacokinetics,” 2013. doi: 10.1177/2042098612469335.

[18] “What Is ADME? Written by Nicole Gleichmann.”

[19] O. M. Ajisafe, Y. A. Adekunle, E. Egbon, C. E. Ogbonna, and D. B. Olawade, “The role of machine learning in predictive toxicology: A review of current trends and future perspectives,” Oct. 01, 2025, *Elsevier Inc.* doi: 10.1016/j.lfs.2025.123821.

[20] M. M. Hossain and K. Roy, “The development of classification-based machine-learning models for the toxicity assessment of chemicals associated with plastic packaging,” *J. Hazard. Mater.*, vol. 484, Feb. 2025, doi: 10.1016/j.jhazmat.2024.136702.

[21] P. Majumdar, *In-depth Logistic Regression*. 2026. doi: 10.5281/zenodo.19446181.

[22] “Sample Data for Gender and Recommendation for Remedial Reading Instruction.”

[23] G. James, D. Witten, T. Hastie, and R. Tibshirani, “An Introduction to Statistical Learning with Applications in R Second Edition,” 2021.

[24] C. Cortes, V. Vapnik, and L. Saitta, “Support-Vector Networks Editor,” Kluwer Academic Publishers, 1995.

[25] J. H. Friedman, “Greedy Function Approximation: A Gradient Boosting Machine,” 2001. [Online]. Available: http://www.jstor.orgURL:http://www.jstor.org/stable/2699986

[26] B. Van Calster *et al.*, “Calibration: The Achilles heel of predictive analytics,” *BMC Med.*, vol. 17, no. 1, Dec. 2019, doi: 10.1186/s12916-019-1466-7.

[27] J. Nixon, G. Brain, M. W. Dusenberry, L. Zhang Google, G. Jerfel, and D. T. Google Brain, “Measuring Calibration in Deep Learning.”

[28] D. J. . Hand, Daniel. Keim, Raymond. Ng, Osmar. Zaïane, and Randy. Goebel, *KDD-2002 : proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining : July 23-36, 2002, Edmonton, Alberta, Canada*. ACM, 2002.

[29] F. Pedregosa *et al.*, “Scikit-learn: Machine Learning in Python,” Jun. 2018, [Online]. Available: http://arxiv.org/abs/1201.0490

[30] C. F. Dormann, “Calibration of probability predictions from machine-learning and statistical models,” *Global Ecology and Biogeography*, vol. 29, no. 4, pp. 760–765, Apr. 2020, doi: 10.1111/geb.13070.

[31] P. Naeini, G. F. Cooper, and M. Hauskrecht, “Obtaining Well Calibrated Probabilities Using Bayesian Binning.” [Online]. Available: www.aaai.org

[32] J. Hernández-Orallo and P. Flach PETERFLACH, “A Unified View of Performance Metrics: Translating Threshold Choice into Expected Classification Loss C` esar Ferri,” 2012.

[33] T. Silva Filho, H. Song, M. Perello-Nieto, R. Santos-Rodriguez, M. Kull, and P. Flach, “Classifier calibration: a survey on how to assess and improve predicted class probabilities,” *Mach. Learn.*, vol. 112, no. 9, pp. 3211–3260, Sep. 2023, doi: 10.1007/s10994-023-06336-7.

[34] K. Lee, K. Lee, H. Lee, and J. Shin, “A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks.”

[35] D. Hendrycks and K. Gimpel, “A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks,” Oct. 2018, [Online]. Available: http://arxiv.org/abs/1610.02136

[36] W. Liu, X. Wang, J. D. Owens, and Y. Li, “Energy-based Out-of-distribution Detection.” [Online]. Available: https://github.com/wetliu/

[37] J. G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N. V. Chawla, and F. Herrera, “A unifying view on dataset shift in classification,” *Pattern Recognit.*, vol. 45, no. 1, pp. 521–530, 2012, doi: 10.1016/j.patcog.2011.06.019.

[38] J. Quionero-Candela, A. Schwaighofer, and N. Lawrence, “Dataset Shift in Machine Learning,” Jan. 2009.

[39] R. Parrondo-Pizarro, J. Lanini, and R. Rodríguez-Pérez, “Uncertainty Quantification in Molecular Machine Learning for Property Predictions under Data Shifts,” *J. Chem. Inf. Model.*, Jan. 2026, doi: 10.1021/acs.jcim.5c02381.

[40] R. P. Sheridan, “Time-split cross-validation as a method for estimating the goodness of prospective prediction.,” *J. Chem. Inf. Model.*, vol. 53, no. 4, pp. 783–790, Apr. 2013, doi: 10.1021/ci400084k.

[41] “Current Status of Methods for Defining the Applicability Domain of (Quantitative) Structure–Activity Relationships”.

[42] S. Weaver and M. P. Gleeson, “The importance of the domain of applicability in QSAR modeling,” *J. Mol. Graph. Model.*, vol. 26, no. 8, pp. 1315–1326, Jun. 2008, doi: 10.1016/j.jmgm.2008.01.002.

[43] J. Lu, A. Liu, F. Dong, F. Gu, J. Gama, and G. Zhang, “Learning under Concept Drift: A Review,” Apr. 2020, doi: 10.1109/TKDE.2018.2876857.

[44] K. Huang *et al.*, “Therapeutics Data Commons: Machine Learning Datasets and Tasks for Drug Discovery and Development,” Aug. 2021, [Online]. Available: http://arxiv.org/abs/2102.09548

[45] R. R. Tice, C. P. Austin, R. J. Kavlock, and J. R. Bucher, “Improving the human hazard characterization of chemicals: A Tox21 update,” Jul. 2013. doi: 10.1289/ehp.1205784.

[46] E. Diamanti-Kandarakis *et al.*, “Endocrine-Disrupting Chemicals: An Endocrine Society Scientific Statement,” *Endocr. Rev.*, vol. 30, no. 4, pp. 293–342, Jun. 2009, doi: 10.1210/er.2009-0002.

[47] R. Daneman and A. Prat, “The blood–brain barrier,” *Cold Spring Harb. Perspect. Biol.*, vol. 7, no. 1, Jan. 2015, doi: 10.1101/cshperspect.a020412.

[48] I. F. Martins, A. L. Teixeira, L. Pinheiro, and A. O. Falcao, “A Bayesian Approach to in Silico Blood-Brain Barrier Penetration Modeling,” *J. Chem. Inf. Model.*, vol. 52, no. 6, pp. 1686–1697, Jun. 2012, doi: 10.1021/ci300124c.

[49] M. Chen, V. Vijay, Q. Shi, Z. Liu, H. Fang, and W. Tong, “FDA-approved drug labeling for the study of drug-induced liver injury,” Aug. 2011. doi: 10.1016/j.drudis.2011.05.007.

[50] Y. Xu, Z. Dai, F. Chen, S. Gao, J. Pei, and L. Lai, “Deep Learning for Drug-Induced Liver Injury,” *J. Chem. Inf. Model.*, vol. 55, no. 10, pp. 2085–2093, Oct. 2015, doi: 10.1021/acs.jcim.5b00238.