MASTERARBEIT | MASTER'S THESIS

Titel | Title

Systematic analysis of the AOP-Wiki and QSAR modeling of developmental and reproductive toxicity targets

verfasst von | submitted by

Eda Inal BSc

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Magistra pharmaciae (Mag. pharm.)

Wien | Vienna, 2025

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 605
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Pharmazie
Betreut von | Supervisor:	Univ.-Prof. Mag. Dr. Gerhard Ecker

Acknowledgments

First of all, I would like to express my deepest gratitude to my supervisor Univ.-Prof. Mag. Dr. Gerhard Ecker for giving me the opportunity to work on my thesis in his group and for his comprehensive and helpful guidance over the past months. I would also like to thank the Pharmacoinformatics Research Group for their help and advice. Thanks to this opportunity, I discovered my interest in computational science, which made me highly motivated and happy to invest my time and energy in this work.

Finally, I am deeply grateful to my family and friends, especially my parents, siblings and grandfather. I am eternally thankful to all of them and will never forget their endless support and encouragement over the past five years. Lots of thanks also to my friends that I met during my pharmacy studies. I have collected many wonderful and joyful memories with them that I will cherish forever.

I want to note, that ChatGPT was used for assistance in refining the wording of this work and for improving clarity and academic style.

Table of Contents

Abstract

The risk assessment of developmental and reproductive toxicity (DART) is particularly complex, as it involves various mechanisms, target organs and life stages. Of all toxicological endpoints, DART requires the highest number of animals and incurs the highest cost for toxicological testing, which is why there is a growing need for new approach methodologies (NAMs), such as quantitative structure-activity relationship (QSAR) models. However, there are currently only few well-established and reliable QSAR models for DART. The aim of this work was therefore to develop QSAR models for biological targets derived from the AOP-Wiki, a publicly accessible database for adverse outcome pathways (AOPs). AOPs describe how a molecular initiating event (MIE) leads to an adverse outcome (AO) via key events (KEs). In this work the data in AOP-Wiki is systematically analyzed, and subsequently twelve targets from MIEs were identified, whose associated AOs are assigned to DART.

For each target, IC50 data was retrieved from ChEMBL35, and regression models were created. Model performance was evaluated using cross-validation, a test set, prospective validation with the updated ChEMBL36 and two additional external datasets (the developmental neurotoxicity list from Risk-Hunt3r project and the DART-list from ECHA). Internal validation showed robust and moderate results (9 targets with $R^2 > 0,5$; 3 targets with $R^2 = 0,33\text{-}0,45$). Prospective validation showed low predictive performance, indicating limited generalizability of the models. In contrast, classification-based external validation showed high sensitivity of 96% and 97%, respectively, suggesting that most truly toxic substances could be correctly identified as toxic.

Overall, the results show that AOP-based QSAR models are a promising method that still needs further development and requires additional research to further improve mechanistic coverage and prediction reliability.

Zusammenfassung

Die Risikobewertung von Entwicklungs- und Reproduktionstoxizität (DART) ist besonders komplex, da verschiedene Mechanismen, Zielorgane und Lebensphasen betroffen sind. Unter allen toxikologischen Endpunkten braucht DART die höchste Anzahl an Tieren und die größten Kosten für toxikologische Prüfungen, wodurch der Bedarf an New Approach Methodologies (NAMs), wie beispielsweise Quantitative Structure-Activity Relationship (QSAR)-Modellen, stetig zunimmt. Bislang gibt es jedoch nur wenig gut etablierte und verlässliche QSAR-Modelle für DART. Ziel dieser Arbeit war daher die Entwicklung von QSAR-Modellen für biologische Targets, die aus dem AOP-Wiki, einer öffentlich zugänglichen Datenbank für Adverse Outcome Pathways (AOPs), abgeleitet wurden. AOPs beschreiben, wie ein Molecular Initiating Event (MIE) über Key Events (KEs) zu einem Adverse Outcome (AO) führt. In dieser Arbeit werden die Daten im AOP-Wiki systematisch analysiert, und anschließend wurden zwölf Targets aus MIEs identifiziert, deren zugehörige AOs der DART zugeordnet sind.

Für jedes Target wurden IC50-Daten aus ChEMBL35 abgerufen und Regressionsmodelle erstellt. Die Modellleistung wurde mittels Kreuzvalidierung, eines Testsets, einer prospektiven Validierung mit der aktualisierten ChEMBL36 sowie zwei weiteren externen Datensätzen (Developmental Neurotoxicity-Liste des Risk-Hunt3r Projekts und DART-Liste der ECHA) bewertet. Die interne Validierung zeigte robuste und moderate Ergebnisse (9 Targets mit $R^2 > 0,5$; 3 Targets mit $R^2 = 0,33$-$0,45$). Die prospektive Validierung ergab eine geringe Vorhersageleistung, was auf eine eingeschränkte Generalisierbarkeit der Modelle hinweist. Die klassifikationsbasierte externe Validierung zeigte hingegen eine hohe Sensitivität von 96% bzw. 97%, was darauf hinweist, dass die meisten tatsächlich toxischen Substanzen korrekt als toxisch identifiziert werden konnten.

Insgesamt zeigen die Ergebnisse, dass AOP-basierte QSAR-Modelle eine vielversprechende, jedoch noch entwicklungsbedürftige Methode darstellen, die weiterer Forschung bedarf, um die mechanistische Abdeckung und die Vorhersagezuverlässigkeit weiter zu verbessern.

1. Introduction

1.1. Developmental and reproductive toxicity

Reproductive toxicology is a complex field covering multiple life phases and biological levels. DART-related adverse outcomes after exposure to drugs, chemicals or pesticides can affect the functional state of the parental reproductive system, the embryonic and fetal development as well as the birth and postnatal maturation. Accordingly, two major categories are distinguished: developmental toxicity and reproductive toxicity (DART). Reproductive toxicity can impair the entire fertility system by disrupting gametogenesis, hormonal regulation, or reproductive capacity (Toragall et al., 2022). Common outcomes include infertility, whose relevance has increased significantly in recent years due to a strong rise in prevalence (Miller et al., 2024).

Developmental toxicity affects the offspring and results from the transfer of toxicants from parents to the embryo or due to pre- or postnatal exposure, particularly during pregnancy. Such exposure may lead to structural and functional dysplasia, miscarriages, stillbirths, and growth impairments (Toragall et al., 2022). A well-known example is the drug thalidomide, originally used as a sedative, which caused severe congenital malformations (Jiang et al., 2019).

1.2. Potential mechanisms of DART

DART can be triggered by many different mechanisms. The most crucial one is oxidative stress characterized by elevated levels of reactive oxygen species (ROS) and disruption of the antioxidant system through inhibition of enzymes such as superoxide dismutase (SOD) and catalase, or by reduction of glutathione levels. As shown in Figures 1 and 2, which illustrates the mechanism of reproductive toxicity in males and females by the example of commonly encountered poly- and perfluoroalkyl chemicals, oxidative stress can ultimately impair the reproductive system through several ways. High ROS levels and mitochondrial dysfunction can induce apoptosis and autophagy in spermatogonia and

oocytes as well as cause DNA damage in sperm which may be transmitted to the offspring, resulting in developmental toxicity (Shi et al., 2024).

Inflammation is another important pathway, involving activation of signaling pathways such as Nrf2/NK-κB and p38 MAPK, as well as the release of pro-inflammatory mediators including TNF-α, IL-6 and others (Peng and He, 2024). Exposure to harmful substances may also increase the risk of cellular damage, mutations, and malignant transformation in reproductive organs and cells. In addition, toxicants can alter enzyme expression and cause organ damage, impairing detoxification and repair mechanisms which increases toxic effects (Wangikar et al., 2011)

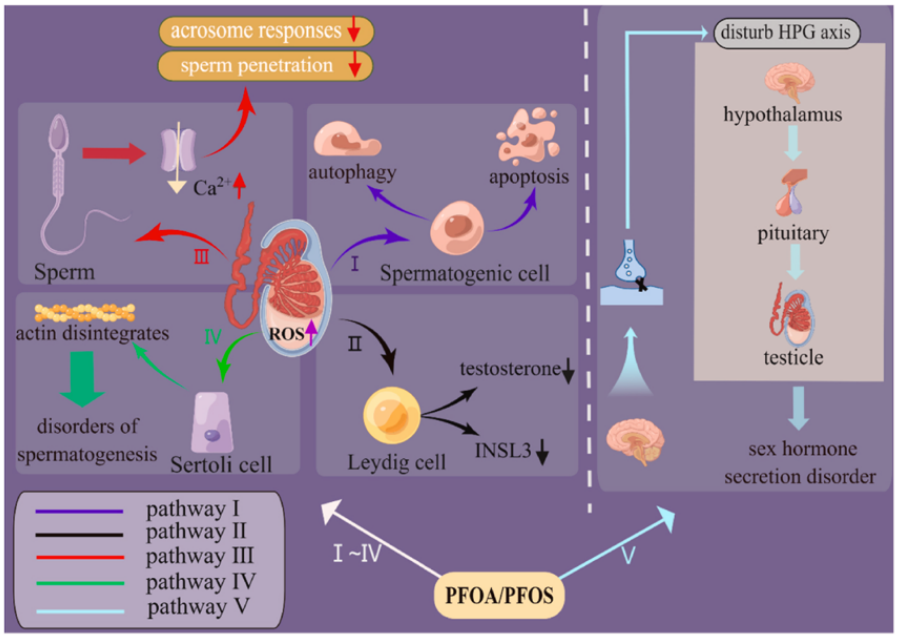

Figure 1: The mechanistic pathways of PFOA/PFOS leading to DART in males (Shi et al., 2024).

In males, Leydig cells are one of the vulnerable targets, leading to reduced testosterone and insulin-like 3 productions. Toxicants can also breakdown the blood-testis barrier by targeting tight and gap junctions and directly damaging Sertoli cells. In both cases, spermatogenesis is subsequently impaired. In females, gap junctions in the cumulus-oocyte complex can be disrupted, and synthesis of steroid hormones in granulosa cells can be reduced by downregulation of the Steroidogenic Acute Regulatory Protein (StAR),

leading to decreased estrogen and progesterone synthesis. This affects follicle and oocyte maturation as well as ovulation (Shi et al., 2024).

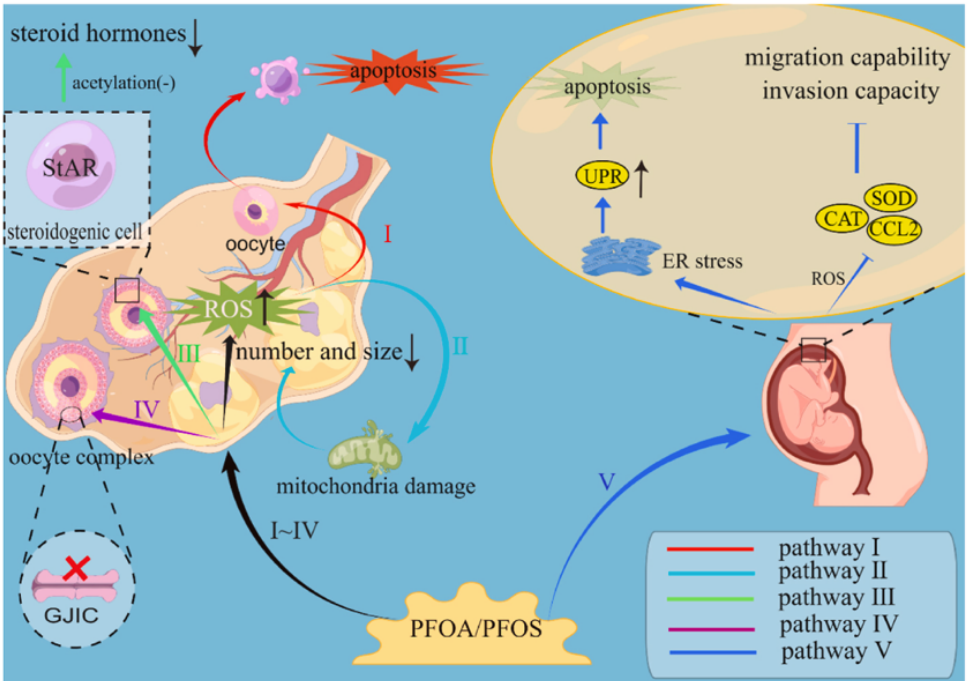

Figure 2: The mechanistic pathways of PFOA/PFOS leading to DART in females (Shi et al., 2024).

Another pathway involves crossing the blood-brain barrier to influence the endocrine system regulated by the hypothalamus and pituitary gland. Normally, gonadotropin-releasing hormone (GnRH) from the hypothalamus stimulates the pituitary to release follicle-stimulating hormone production (FSH) and luteinizing hormone (LH), which in turn induce sex hormone production. Toxicants that disrupt this system reduce serum hormone levels which negatively affects the development, function and maintenance of the reproductive system (Peng and He, 2024; Shi et al., 2024; Wangikar et al., 2011).

A widely known example of developmental toxicity is thalidomide. Although the mechanism is not fully understood, it is assumed that the compound inhibits angiogenic signaling in endothelial cells of the offspring, resulting in growth disturbances, loss of embryonic connective tissue and cell death. Another mechanism is the production of pro-inflammatory mediators and ROS, which damage embryonic DNA directly. Xenobiotics may cross the placenta, causing structural and functional alterations of the placenta itself

that prevents fetal supply and disrupt hormonal balance. As shown in Figure 3, maternal inflammation or oxidative stress induced by exposure to harmful substances by various ways such as inhalation, ingestion, dermal and injection can also trigger the release of signaling molecules that cross the blood-placenta and affect embryonic development (Dugershaw et al., 2020).

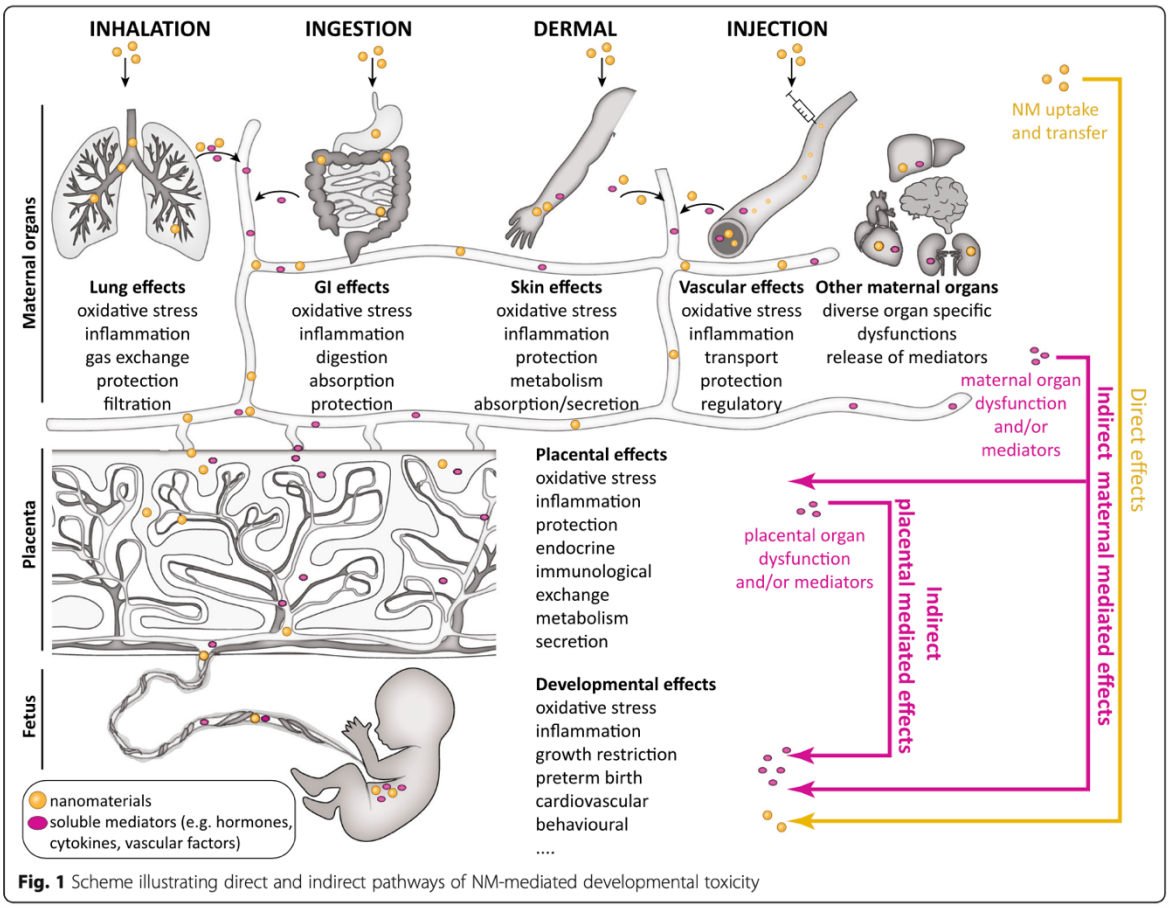

Figure 3: Direct and indirect pathways of nanomaterials leading to Developmental Toxicity (Dugershaw et al., 2020)

1.3. Challenges in the risk assessment of DART

The diversity of mechanisms, life stages, and target systems involved in DART makes its risk assessment highly complex and complicates the attribution of adverse effects to a single causative factor. As a result, extensive studies across different developmental phases and biological systems are required, which is costly, time-consuming, and ethically problematic due to a high number of animals involved (Weyrich et al., 2022). A full

toxicological assessment of a single substance requires approximately 3,200 animals and costs around half a million US dollars, making DART responsible for about half of all toxicological study costs (Zhang et al., 2017). Thus, DART presents a major challenge in drug research and development. Reproductive toxicity is one of the frequent reasons for the discontinuation of approximately 10% of preclinical studies and the withdrawal of about 3% of marketed drugs. Consequently, methods are needed that can identify toxicity-related effects early in development and enable rapid screening of large numbers of compounds (Feng et al., 2021). The complexity of DART has driven the development of alternative methods to replace, reduce, and refine (3R) animal testing. These new approach methodologies (NAMs) include in vitro in silico approaches (Myden et al., 2024).

1.4. QSAR models of DART

One of the in silico methods is quantitative structure-activity relationship (QSAR), which mathematically predicts the toxicity of compounds based on their chemical structure. This approach is particularly suitable for efficient and inexpensive screening of large compound libraries within short time frames (Zhang et al., 2017).

Although DART is among the most critical toxicity categories under the European Union´s REACH regulation, the number and reliability of available computational models for DART are significantly lower than for other toxicological endpoints. But since in silico models provide substantial advantages over animal testing, there is a clear need for the development of well-established QSAR models (Zhang et al., 2017), which is the aim of this thesis.

1.5. Adverse outcome pathways

Adverse Outcome Pathways (AOP) ease the development of NAMs and support the application of the 3Rs principle. AOPs describe a linear sequence of biological events that

represent the initiation and progression of a toxicological mechanism (Knapen et al., 2015) as shown in Figure 4.

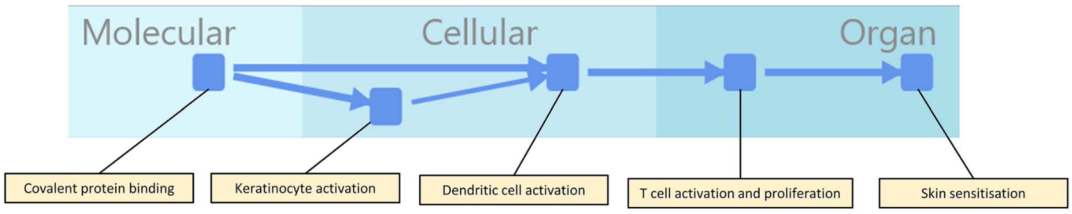

Figure 4: An example of AOP, with MIE at the start, KE in cellular level, and AO at the organ level. KERs are shown as arrows (Ball et al., 2021).

An AOP is described by the following elements, which are linked together in a chain-like sequence:

- MIE (Molecular Initiating Event): The first triggering event of an AOP. It describes the interaction of a stressor with a biological target (e.g., binding to an enzyme or receptor) (Kleinstreuer et al., 2016).
- KE (Key Event): The intermediate steps that occur after the MIE. They describe changes that can take place at the cellular, tissue, or organ level, such as altered gene expression, changes in hormone levels, apoptosis, or impaired cellular or organ function. All KEs must be clearly defined and quantifiable to support the development of assays and NAMs. The causal connections between the KEs are described by KERs (Key Event Relationships) (Knapen et al., 2015).
- AO (Adverse Outcome): The final endpoint of the cascade, representing a disease, an anomaly, or another harmful effect at the individual or population level (Kleinstreuer et al., 2016).

AOPs are collected, documented and stored in the AOP Knowledge Base (AOP-KB), a central platform developed by the OECD that includes the publicly accessible AOP-Wiki (https://aopwiki.org/). AOP-Wiki is a database which provides an online collection of qualitative and descriptive AOPs. Users can easily search up AOPs and the platform provides an accessible interface for retrieving associated information (Knapen et al., 2015).

Since a single AOP represents only a linear sequence of events — from one MIE to one AO — it is often insufficient to capture the full toxicological complexity. Therefore, AOP networks have been developed to represent relationships among several AOPs that share common KEs and AOs (Knapen et al., 2015). Figure 5 shows an example of 2 MIEs (aromatase inhibition and androgen receptor agonism) sharing some identical KEs as well as ending in the same AO – decreased population trajectory.

For this to be possible, identical key events must be consistently defined in the same way to enable their integration into multiple AOPs. In summary, AOP networks allow complex biochemical processes to be represented more realistically, support the discovery of new mechanistic linkages and the development of in vitro and in silico tests for toxicity assessment while reducing animal use (Knapen et al., 2015).

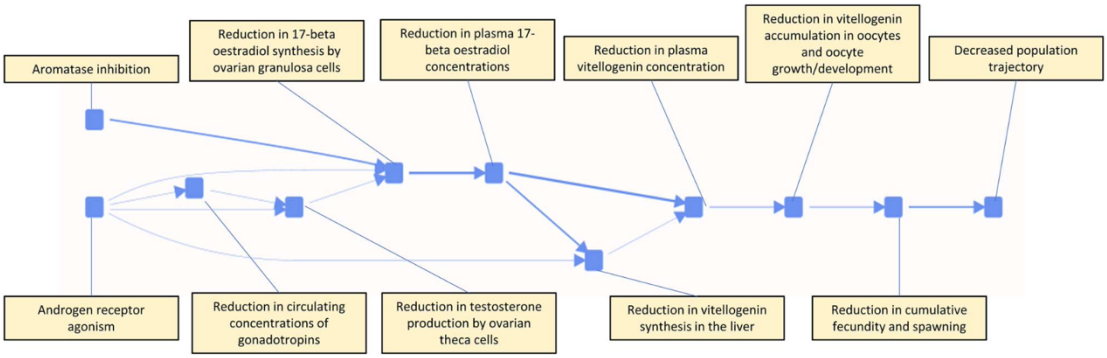

Figure 5: An example of AOP network combined by multiple AOPs for the same endpoint (DART) (Ball et al., 2021).

2. Aim of this thesis

The aim of this work was first to systematically analyze the data available in the AOP-Wiki in order to assess the current state of development of AOPs. Particular attention was given to the molecular targets described within MIEs, as well as to the organs and body systems affected by the associated adverse outcomes.

Based on this analysis, the focus of the thesis is to be placed on biological targets associated with a specific organ/body system. For these targets, the aim is to develop QSAR models to predict the biological activity of chemical substances based on their molecular structure and to validate the models using external datasets to assess their predictive performance and robustness.

3. Methods

3.1. KNIME

KNIME stands for Konstanz Information Miner and is a freely available platform for analyzing, processing and reporting data, providing functions for cheminformatics and machine learning (Fillbrunn et al., 2017). KNIME is very user-friendly and easy to use without coding knowledge, as it works with so called nodes. Each node has its own function, which can be set as required and desired in the configuration page. Connecting all nodes creates a pipeline called “workflow”. One can check their results in between before continuing with the next node and visualize the results. To provide an overview, parts of workflows can also be combined in a metanode, which then looks an can be used like a single node (Sydow et al., 2019).

Version 5.4.4 was used for this work.

3.2. UniprotKB

UniprotKB provides information about protein sequences of various organisms. A distinction is made between UniprotKB/Swiss-Prot and UniprotKB/TrEMBL. The former is a collection of reviewed proteins whose functional information has been manually curated from scientific papers and is therefore of high quality, while the latter consists of unreviewed proteins that are automatically annotated. Furthermore, additional input from the community can add to the database. Uniprot provides a summary of the protein for users as well as structured data for machine analysis. UniprotKB can be downloaded in various formats (such as XML, JSON) or retrieved via APIs or URLs (The UniProt Consortium et al., 2023).

3.3. ChEMBL

The ChEMBL database provides freely accessible bioactivity data consisting of information on binding, function, toxicity-related data and ADMET of millions of compounds. This data is regularly curated manually, extracted especially from scientific papers (Gaulton et al., 2012). This allows potential drug candidates to be identified, the correlation between molecular structures and their biological activity to be investigated, and off-target effects to be detected, thus serving as a fundamental basis for drug discovery (Gaulton et al., 2017).

Each compound, target, assay and document has its own ChEMBL-ID. ChEMBL offers the option of accessing the data either by download, web services (RESTful) or web interface (Gaulton et al., 2012). This has enabled data retrieval in KNIME. Versions 35 and 36 were used in this work.

3.4. RCSB-PDB

The RCSB Protein Data Bank is an open-source database containing 3D structures of molecules such as proteins and nucleic acids, which is available for public use. It plays a crucial role in drug discovery and research, as it enables better research into ligand-target relationships on a molecular basis and disease mechanism, thereby helping to identify potential drugs. These 3D structures of biomolecules range from X-ray structures and NMR structures to cyro electron microscopy and are determined experimentally and have high resolutions (Ahmad et al., 2025). The importance of freely accessible structures is reflected in drug development as a large proportion of FDA-approved drugs have used PDB as an aid (Goodsell et al., 2020).

3.5. Machine learning models

Machine Learning (ML) is an in silico approach used to generate algorithms that identify patterns and trends in a dataset and apply them to predict a novel dataset (Handelman et

al., 2018). ML automatically develops analytical algorithms entirely from input data, for example to construct QSAR. QSAR mathematically describes how biological or toxicological effects or properties are influenced by the structural characteristics of a compound. This approach is used, among other applications, to predict the toxicity of substances (Jiang et al., 2019).

Datasets are typically differentiated between a training set, a validation set, and a holdout test set. The model learns the relationships and patterns between features and outcomes from the training set and is optimized using the internal validation set. The test set, which the model has never encountered before, is then used to realistically assess model performance and generalizability. At last, an external validation set then validates the model performance (Jiang et al., 2019). Two types of models are generally distinguished. Regression models predict continuous numerical values, such as pIC50, for new data. Classification models, on the other hand, assign new substances to predefined categories, such as „active“ or „inactive“ (Handelman et al., 2018). In this thesis, only regression models were applied.

3.5.1. Model optimization

In ML, the focus is less on understanding the internal decision-making process of a model and more on improving predictive accuracy and ensuring applicability to new, unseen data. This is primarily achieved through cross-validation and hyperparameter tuning (Handelman et al., 2018). Cross-validation is an internal validation method in which the dataset is split into so-called folds. In a 5-fold cross-validation, for example, 20% of the compounds are withheld while the model is trained on the remaining 80%. The withheld 20% are then predicted. This process is repeated five times so that each compound is predicted once. However, because each compound is used in training four times during cross-validation, this method does not provide a true measure of how well the model would predict entirely new substances. For this reason, it is considered an internal validation method. External validation uses a test set that the model has never seen or been trained on to refine evaluation and assess generalizability (Zhang et al., 2017).

During hyperparameter tuning, predefined ranges of hyperparameters are systematically tested to identify the hyperparameter combination that results in the best performance, thereby optimizing the model and reducing the risk of overfitting or underfitting (Sandunil et al., 2024).

3.5.2. Performance evaluation

For regression models, the coefficient of determination (R^2) is commonly used as a performance metric. It quantifies how much of the variability in the data can be explained by the model. A value 0 indicates that none of the variability is explained, whereas a value of 1 indicates that all variability is accounted for. In general, models with an R^2 above 0.6 are considered to demonstrate good predictive performance. For classification models, metrics such as precision and sensitivity are evaluated. Precision reflects the variability and repeatability of predicted values, while sensitivity is a representation on how many true positive data can be correctly predicted as positive (Handelman et al., 2018).

3.6. Workflow

This thesis focuses on the analysis of data from AOP-Wiki (aopwiki.org), with particular emphasis on MIEs. To enable the extraction, processing, analysis, and visualization of this data, a comprehensive workflow was developed in KNIME. Furthermore, QSAR models were generated and evaluated in KNIME. The workflows consist of several stages, which are described in detail below.

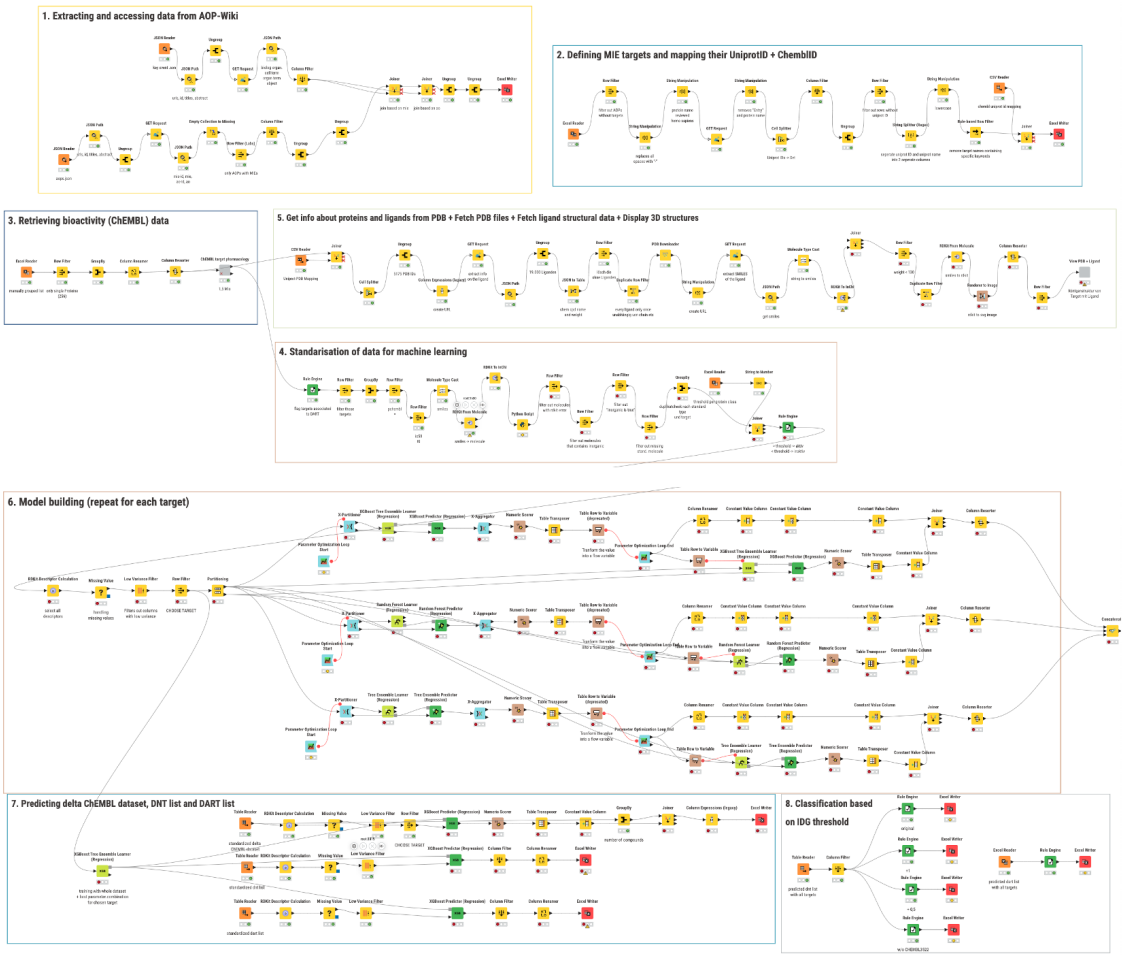

Figure 6: Overview of the whole KNIME workflow

3.6.1. Retrieval of AOP data

The first step involved retrieving two separate datasets from AOP-Wiki via their API in JSON format. The first dataset contains all AOP-related information, including the AOP ID, AOP title, MIE ID, MIE, AO ID, and AO. The second dataset comprises the Key Events, which also include MIEs and AOs but provide more detailed descriptions specifying the precise target involved and the organ or cell type affected by the adverse outcome. Because this information is essential for subsequent querying and analysis, both JSON files were imported into KNIME. The relevant fields were extracted, and the datasets were merged into a single table. This initial integration yielded 414 AOPs and 238 MIEs.

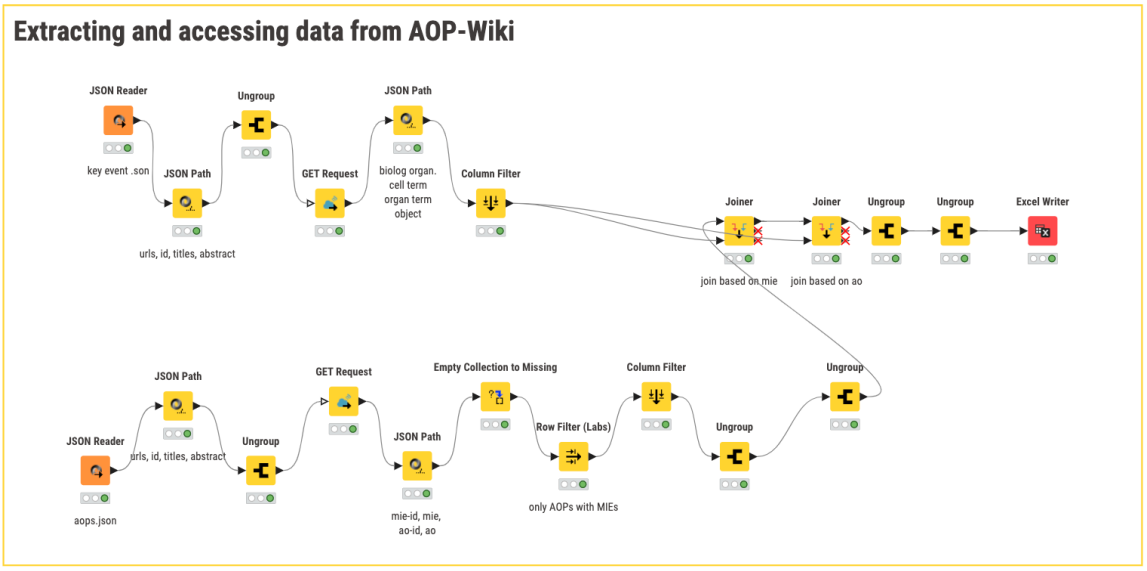

Figure 7: KNIME workflow of extracting and accessing data from AOP-Wiki

The second stage focused on defining the MIE targets and mapping them to their corresponding UniProt and ChEMBL identifiers. Given the importance of target accuracy for the final analysis, MIEs that had missing target names were manually added – which was not possible for each MIE. The dataset was then re-imported into KNIME and filtered to retain only AOPs with defined targets, resulting in 286 AOPs and 157 MIEs. To automatically access UniProt IDs, the UniProt REST API was used. For this, a REST URL was generated for each target using KNIME´s *String Manipulation* node and subsequently queried via *GET* requests node. The URL was constructed to retrieve only proteins that are reviewed in the UniProtKB database and originate from the human organism. The expression for this URL is:

> *„string("https://rest.uniprot.org/uniprotkb/search?query=protein_name:" + urlEncode($target_name$) + "+AND+organism_id:9606 +AND+reviewed:true &format=tsv&fields=accession,protein_name")“*

The results showed that some targets had multiple UniProt IDs assigned, as the query returned all entries containing the target name in any form, including synonyms, domains, and protein families. Therefore, a *Rule-Based Row Filter* was applied to exclude entries containing undesired keywords such as „associated,“ „interacting,“ „binding,“ „cofactor,“

„domain," „precursor," „pseudogene," or „hypothetical". In parallel, the „chembl_uniprot_mapping" file was downloaded from the ChEMBL FTP server. This file lists all UniProt IDs together with their corresponding ChEMBL IDs, protein names, and classifications (single proteins, protein families, or other). The file was imported into KNIME and joined with the original AOP table based on the UniProt ID. Wherever UniProt IDs matched, the corresponding ChEMBL ID was appended to the dataset.

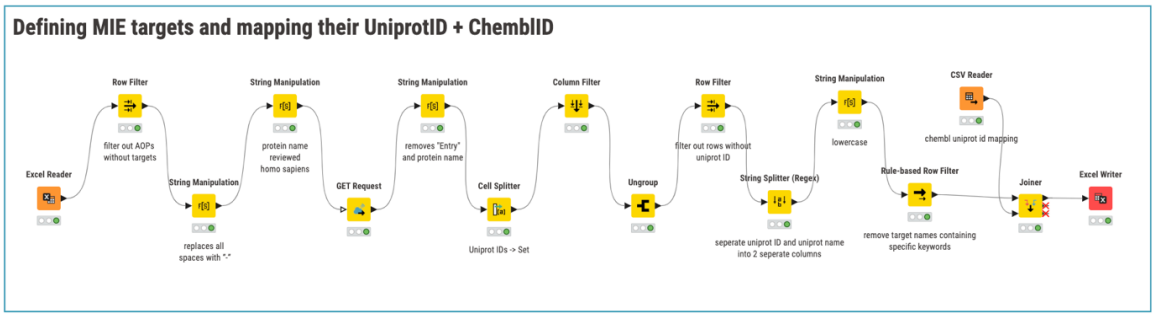

Figure 8: KNIME workflow of defining MIE targets and mapping their Uniprot ID + ChEMBL ID

For further refinement, the table was manually curated in Excel: single proteins were assigned the code „1", while protein families and complexes were assigned „2". The curated dataset was then expanded by grouping each MIE according to the body system or organ affected by the adverse outcome. This dataset was re-imported into KNIME. Since only single proteins were relevant for the subsequent analysis, a Row Filter was applied to remove rows assigned the code „2", resulting in a dataset of 151 AOPs, 70 MIEs, and 49 unique single targets.

The third and one of the most crucial steps involved retrieving bioactivity data for the 49 targets from ChEMBL35. For this purpose, the „*ChEMBL target pharmacology*" metanode developed by Daniela Digles (University of Vienna) was used. This node uses the ChEMBL ID to query bioactivity data, returning information on compounds that interact with the targets in bioactive assays. The output included details such as total count, activity comment, activity ID, molecule identifiers and names, canonical SMILES strings, and measured activity values, including pChEMBL values.

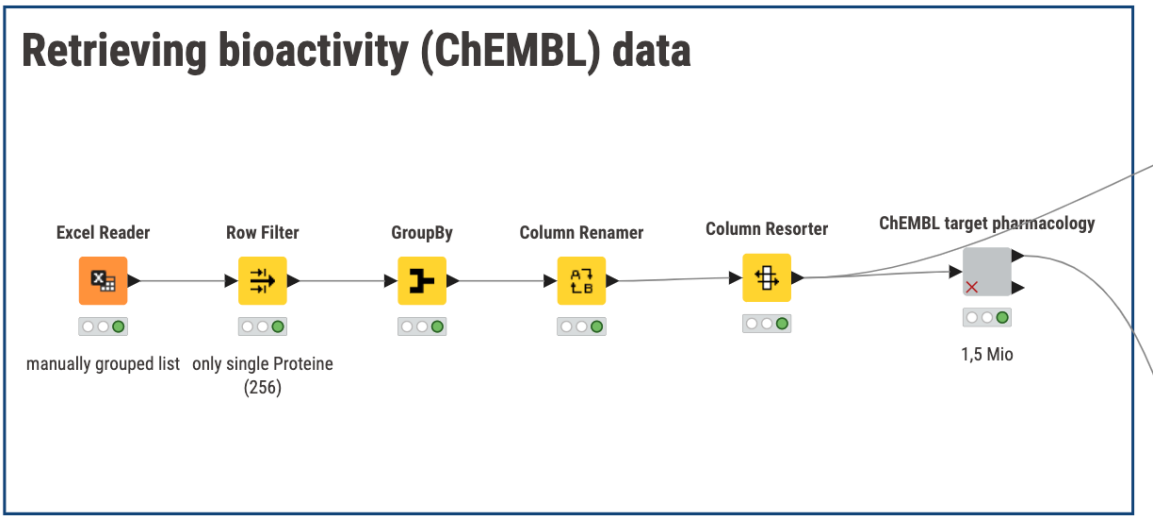

Figure 9: KNIME workflow of retrieving bioactivity data from ChEMBL

3.6.2. Data preparation

After retrieving the bioactivity data from ChEMBL, a series of nodes were applied to prepare the dataset for effective QSAR modelling. Only rows that contained a pChEMBL value, a standard relation of „=“, a standard unit of „nM“, and the standard activity types IC50 and Ki were retained. Additionally, the dataset was standardized using a node developed by Sergey Sosnin (University of Vienna). It cleaned the molecules, disconnected the metals, removed any fragments that were not the sought molecule, uncharged and normalized as well as removed all stereochemistry information to aggregate duplicates and reduce noise. The node also checked whether the molecule contained any inorganic atoms.

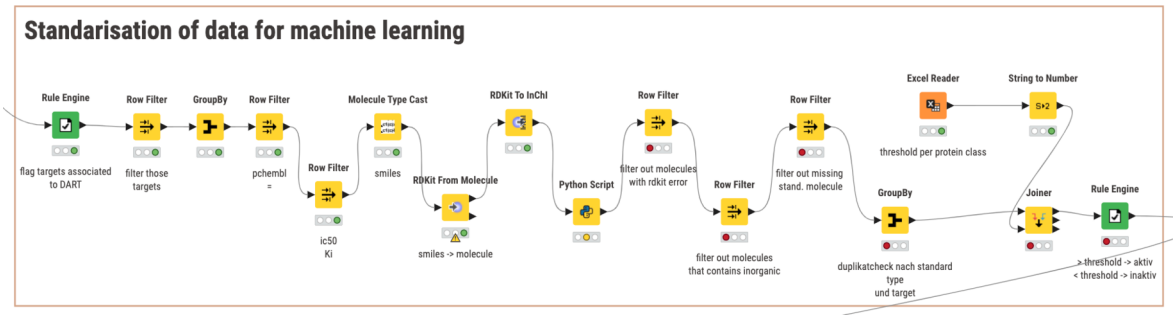

Figure 10: KNIME workflow of standardization of data for machine learning modeling

3.6.3. Decision between classification or regression

To classify molecules as active or inactive, the corresponding protein family for each target was identified in ChEMBL and listed together with their corresponding activity thresholds. For this purpose, the thresholds values defined by Illuminating the Druggable Genome Initiative (IDG) were applied (Table 1). A compound with a pChEMBL value above the threshold was classified as active; otherwise as inactive. Following this, all molecules containing inorganic atoms were removed, and duplicates were eliminated based on the standardized molecular structure, the standard activity type, and the associated targets. Thus, compounds with multiple pChEMBL entries per target were aggregated by calculating the median value, resulting in each molecule appearing only once per target.

protein family	threshold	-log(threshold)
Kinase	30 nM	7,52
GPCR	100 nM	7
Nuclear receptor	100 nM	7
Ion channel	10 µM	5
Non-IDG family targets	1 µM	6

Table 1: IDG Threshold and corresponding -log(threshold) for each protein family (Druggable Genome Initiative, n.d.)

3.6.4. Decision on data to continue with

Each MIE was grouped according to the organ or body system affected by its associated adverse outcome. For example, adverse outcomes that contained terms such as *depression* or *neurotoxicity* were assigned to the nervous system, while terms such as *reduced fertility* were associated with the reproductive system. Given that 49 different targets were identified and their relationships were complex, the analysis was narrowed to a specific group. The three most frequently represented organ/body systems were the nervous system, the reproductive system and the liver (Figure 21). For subsequent analyses, the focus was placed on the fifteen targets associated with the reproductive system.

Furthermore, a decision was made regarding the use of regression or classification machine learning approaches. For each target, the number of compounds, the minimum and maximum activity values, and the proportion of active and inactive compounds were determined. If the number of compounds exceeded 100 and the difference between the minimum and maximum activity values was greater than 2, regression modelling was considered applicable. If the proportions of active and inactive compounds were balanced, classification modelling was considered applicable. Based on the data presented in Figure 22 regression was identified as the most suitable method, with only three out of fifteen targets not fulfilling the required criteria.

3.6.5. Model building for 12 targets

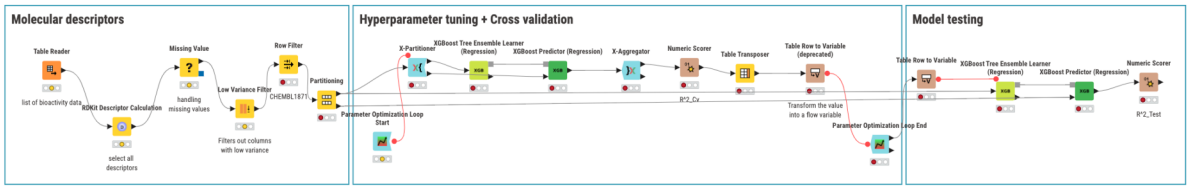

Figure 11: KNIME workflow of model buildung: 1. starting with the calculation of RDKit Descriptors 2. Optimization by hyperparameter tuning and cross-validation 3. Performance evaluation of the model with the best hyperparamter combination using a test set

The first step was to calculate descriptors for all compounds using the *RDKit Descriptors* node. The QSAR-ready dataset was then randomly partitioned into a training set (70%) and a test set (30%) using the *Partitioning* node.

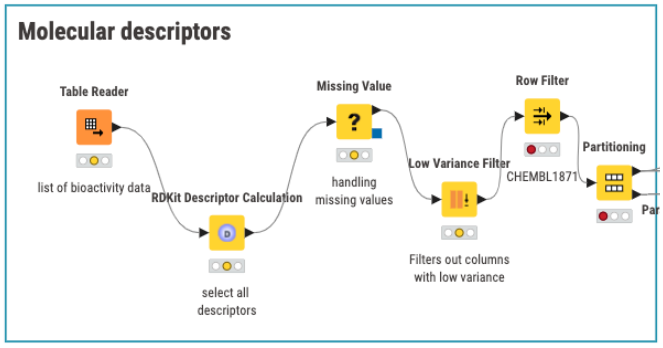

Figure 12: Calculation of RDKit descriptors

Subsequently, the hyperparameter tuning process was initiated to identify the optimal parameters for the models. Three regression learners were selected: XGBoost, Random Forest and Tree Ensemble. A parameter optimization loop was created for each learner, with all parameters controlled as flow variables. Within this loop, an additional loop was implemented in which the training set was randomly divided into 10 folds using *X-Partitioner* node to perform cross-validation. Using the trained model, the predictor generated pIC50 predictions for the test fold. The results of all folds were aggregated within the *X-Aggregator* node. Statistical performance metrics, including R^2, mean absolute error (MAE), root mean squared error (RMSE) and others, were calculated using the *Numeric Scorer* node.

Due to the selected Brute-Force strategy in the *Parameter Optimization Loop Start* node, the loop was repeated until all possible parameter combinations had been evaluated once. The *Parameter Optimization Loop End* node then collected all combinations and their corresponding performance metrics. Based on the chosen objective function R^2, the optimal parameter combination was identified. This R^2 value was referred to as „R^2_cv“ in this work.

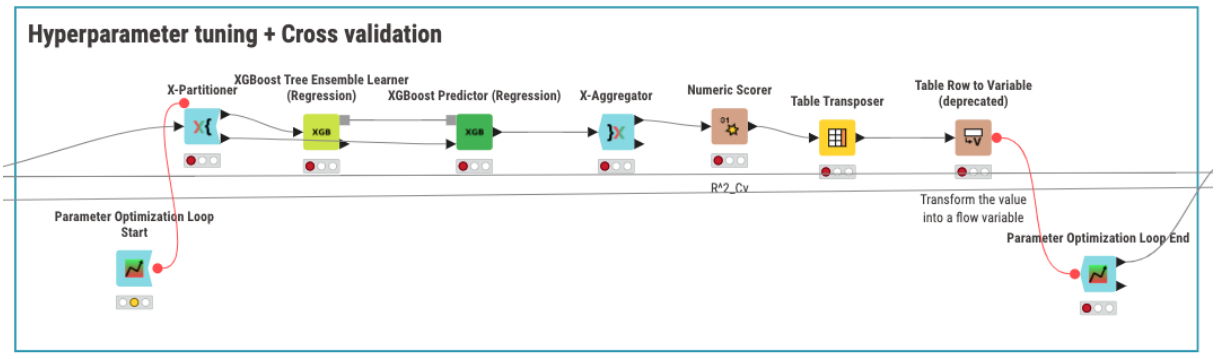

Figure 13: Optimization of the model by hyperparameter tuning and cross-validation

Using these optimal parameters, a final model for each algorithm was trained again – this time on the entire training set without folds. The previously separated 30% test set served as the unseen data on which the model was evaluated. The resulting „R^2_test“ values indicated the model's ability to predict new data and represented the actual performance evaluation metric.

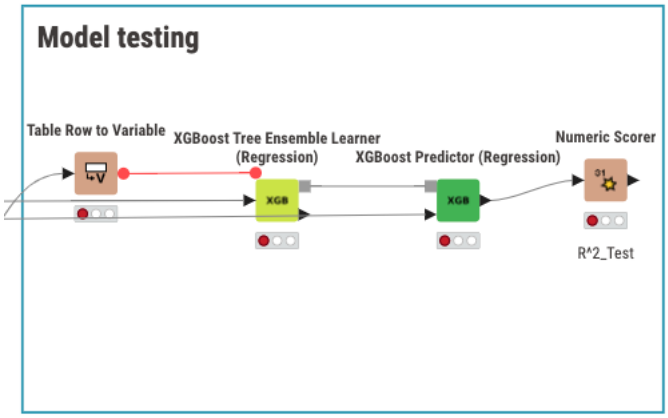

Figure 14: Performance evaluation of the model with the best hyperparameter combination using a test set

3.6.6. Performance evaluation and hazard identification

To obtain a more robust assessment of the model‘s generalizability and predictive power, the models were validated using three external datasets.

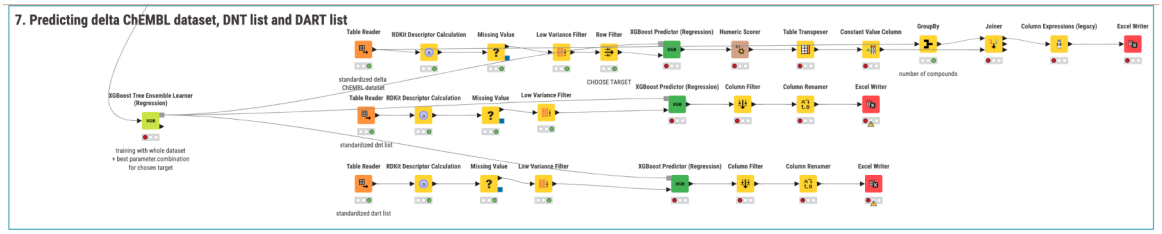

Figure 15: Overview of predicting updated ChEMBL dataset, DNT list and DART list

3.6.6.1. Updated ChEMBL Release

The first external dataset consisted of new entries available in the updated ChEMBL36 that were not present in the earlier version. All steps described in Chapter 3.6.2 were repeated to standardize and clean the dataset. For model building, data from ChEMBL35 served as the training set, while the new entries from ChEMBL36 were used as the external validation set. According to Figure 24, the algorithm with the highest R^2 value was selected for each target, and a corresponding model was built using its optimal parameter configuration. After predicting the plC50 values of the new data, the R^2 value calculated using the *Numeric Scorer* node.

3.6.6.2. DNT-list from Risk Hunt3r Project

The second external test set was the Developmental Neurotoxicity (DNT) list from the Risk-Hunt3r Project. DNT is a subfield of DART, since it affects the developmental of the nervous system of an offspring. The list consists of 1,069 compounds that were classified as either DNT-positive or DNT-negative based on previous research by various authors. The list includes inorganic and organic compounds as well as mixtures, and for most of them the corresponding SMILES strings are already provided. After importing the list into KNIME, only 731 compounds with available SMILES were filtered. These SMILES were converted into RDKit molecules and standardized as described in Chapter 3.6.2. Following the removal of inorganic molecules, 705 compounds remained and were used as the external test set. The previously described model prediction step was repeated with this list; however, R^2 could not be calculated because no observed pChEMBL values were available for comparison. For this reason, the predicted plC50 values were compiled into a table and color-scaled in Excel.

3.6.6.3. DART-list of ECHA

The third external test set was a list of compounds identified by ECHA as causing DART. ECHA, the European Chemicals Agency of the European Union, provides guidance and information on chemical safety and offers a publicly accessible database containing the Classification and Labelling (C&L) Inventory (https://echa.europa.eu/information-on-chemicals/cl-inventory-database). Since the database includes 359,742 substances, a filter was applied to reduce the dataset to compounds classified as „Repr.1A," „Repr. 1B," or „Repr. 2.". These hazard categories correspond to:

- **Repr. 1A**: DART based on human data
- **Repr. 1B**: DART based on animal studies
- **Repr. 2**: Suspected DART based on limited evidence (Scialli, 2008)

After applying the filter, 7,447 substances remained, each associated with ist CAS number, classification, source, and pictograms. This dataset was downloaded as an XLS file and

imported into KNIME. A URL was generated using *String Manipulation* to access the PubChem page of each compound via CAS number and retrieve the canonical SMILES using a *GET Request* node:

> *join(https://pubchem.ncbi.nlm.nih.gov/rest/pug/compound/name/, $CAS no.§, "/property/CanonicalSMILES/TXT")*

The dataset was cleaned, standardized, and prepared for machine learning as described previously. Ultimately, 4,891 organic substances remained and were used as an external test set.

3.6.7. Structural data from PDB

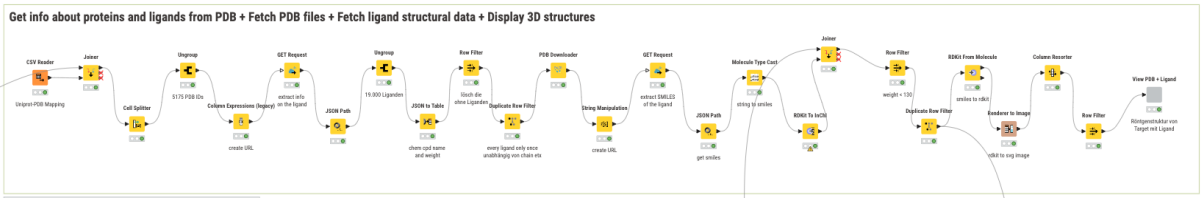

Figure 16: KNIME workflow of retrieving structural data from PDB and displaying 3D structures

An additional stage focused on the retrieval of structural data. A CSV file mapping UniProt IDs to PDB IDs was downloaded from the SIFTS Project FTP server and joined with the existing dataset based on the UniProt ID, resulting in 2,307 distinct PDB IDs. For each PDB entry, a URL was generated and queried via *GET requests* to obtain ligand-monomer information, including ligand identifiers, names, and molecular weights. Structures were downloaded in PDB format from the RCSB PDB using the *PDB Downloader* node. Additional URLs were generated to retrieve SMILES representations of the ligands, which were then converted into InChl codes and InChlKeys, providing unique molecular identifiers for comparison. This step was required to match extracted ligands with compounds obtained from the ChEMBL bioactivity data. An additional filter excluded all molecules with a molecular weight below 130 Da, thereby removing non-drug-like substances such as buffer components. This refinement resulted in 592 distinct ligands. For visualization, SMILES strings were rendered into two-dimensional deceptions using

RDKit, while PDB structures were displayed in a 3D viewer. Both visualization approaches were integrated into a single KNIME component, providing a unified visual representation of ligand and structural data. Through this step, it was possible to identify which targets possess X-ray crystal structures with bound ligands, which may serve as a valuable basis for future structure-based studies such as docking or structure-based pharmacophore modelling. However, this approach was not part of the aim of this thesis.

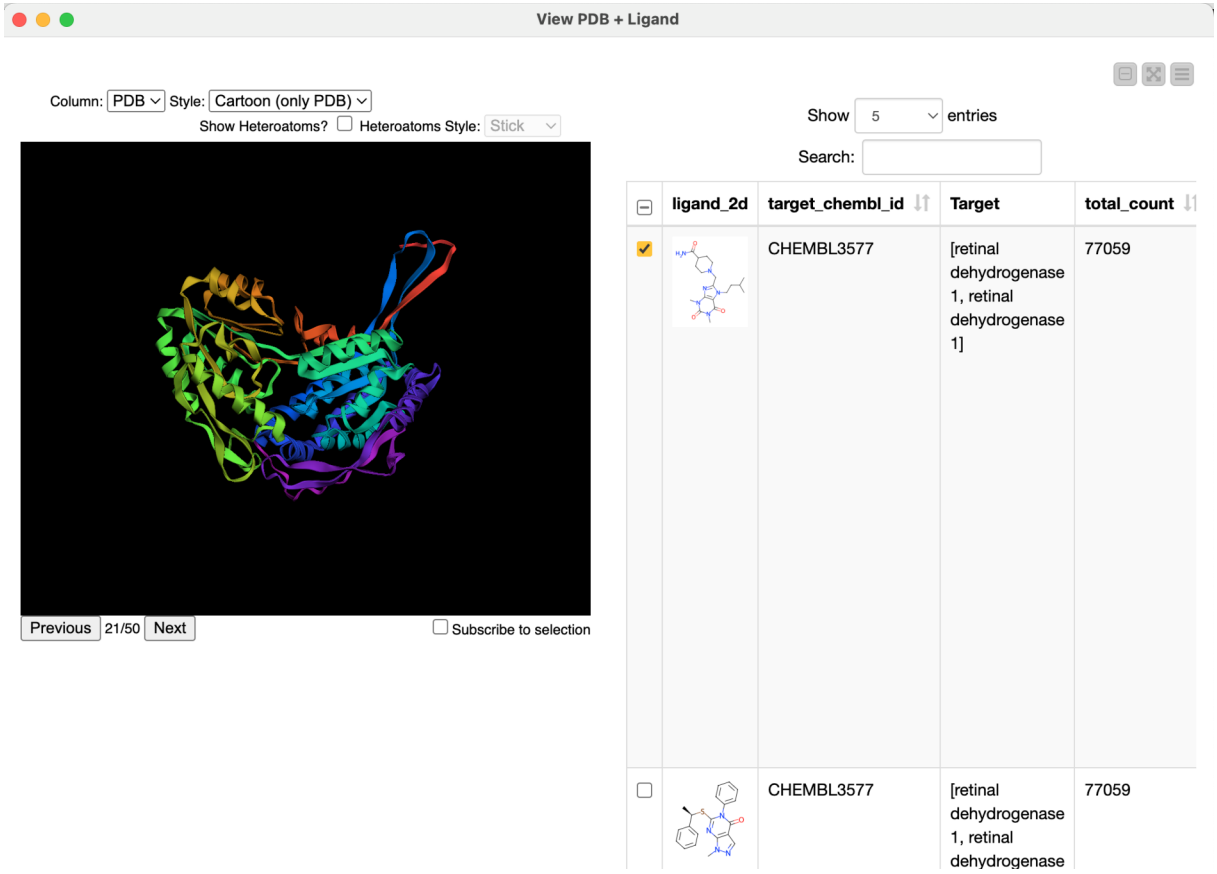

Figure 17: 3D visualisation of proteins (CHEMBL3522 shown in this figure) and their ligands displayed in 2D in the table on right

4. Results and discussion

4.1. Analysis of AOP-Wiki

4.1.1. Dataset Statistics

At the beginning of the analysis, the dataset comprised a total of 519 AOPs, of which 414 contained a defined MIE. Overall, 238 unique MIEs and 190 different AOs were identified, with several MIEs and AOs occurring repeatedly across different AOPs. The most frequently occurring MIEs are summarized in Table 2, where increases in reactive oxygen species, activation of the aryl hydrocarbon receptor (AhR), and energy deposition being the most prevalent events. MIEs appearing in several AOPs highlights the importance of using consistent and standardized terminology when defining MIEs and KEs, since it enables the linking of individual events and the construction of AOP networks for the better understanding of biological mechanistic complexities. In addition, it allows conclusions to be drawn regarding which protein or pathophysiological alterations have a particularly strong impact on toxicological outcomes (i.e. the increase of ROS). Another reason for the frequent occurrence of certain MIEs may be the large number of studies that have been done on these mechanisms. This leads to a certain degree of bias into the dataset and could mean that less explored mechanism may be underrepresented.

MIE-ID	Event	Count
1115	Increase, Reactive oxygen species	21
18	Activation, AhR	19
1686	Deposition of Energy	17
1739	Binding to ACE2	15
279	Thyroperoxidase, Inhibition	9
559	Activation, Nicotinic acetylcholine receptor	9

Table 2: The six most occured MIEs across AOPs.

Of the 238 identified MIEs, only 157 were associated with specific targets. After excluding targets without an UniProt-ID or ChEMBL-ID, 118 MIEs corresponding to 88 targets remained. Among these, 49 targets (55.68%) represented single, uniquely defined proteins, while 38 targets (43.18%) corresponded to protein families, protein complexes, or protein-protein interactions. One target (1,14%) represented a single protein but had no ChEMBL-ID (Figure 18). For QSAR modeling only single proteins are preferred. Since protein families and complexes were not used in further analyses, it led to a loss of information regarding additional mechanisms of action and potential target sites which lead to certain adverse outcomes. Individual members of protein families could be modeled separately, but this was beyond the scope of this work.

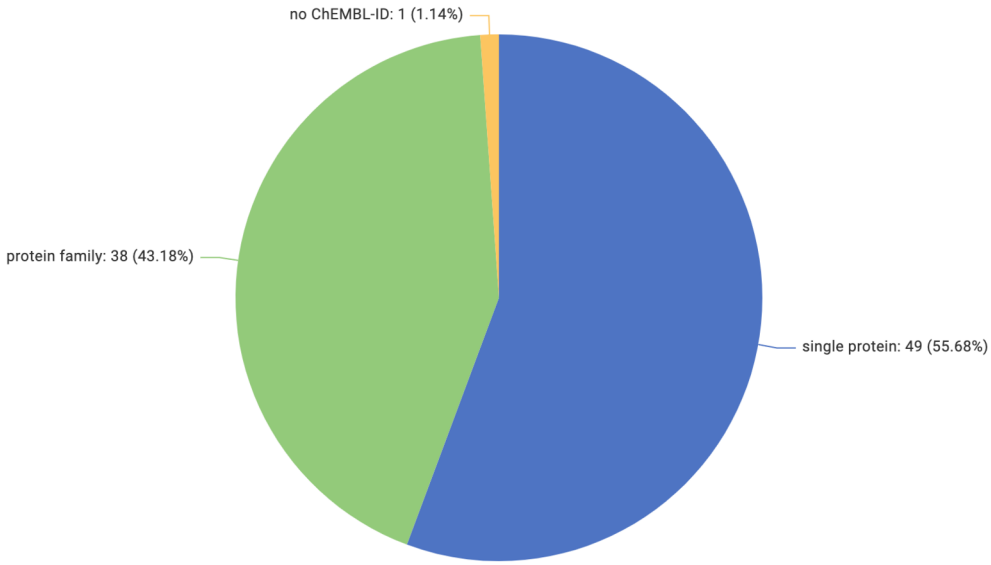

Figure 18: Pie chart showing the frequency of single proteins vs. protein families.

The frequency of protein occurrence across MIEs is illustrated in Figure 19. The aryl hydrocarbon receptor (AhR) was the most frequently represented protein, occurring 46 times, which could be explained by the high prevalence of the MIE „Activation, AhR“ (MIE-ID 18). The second most frequently occurring protein was the androgen receptor with 23 occurrences, followed by the sodium-dependent serotonin transporter with 20 occurrences, which are among the targets of interest in this thesis.

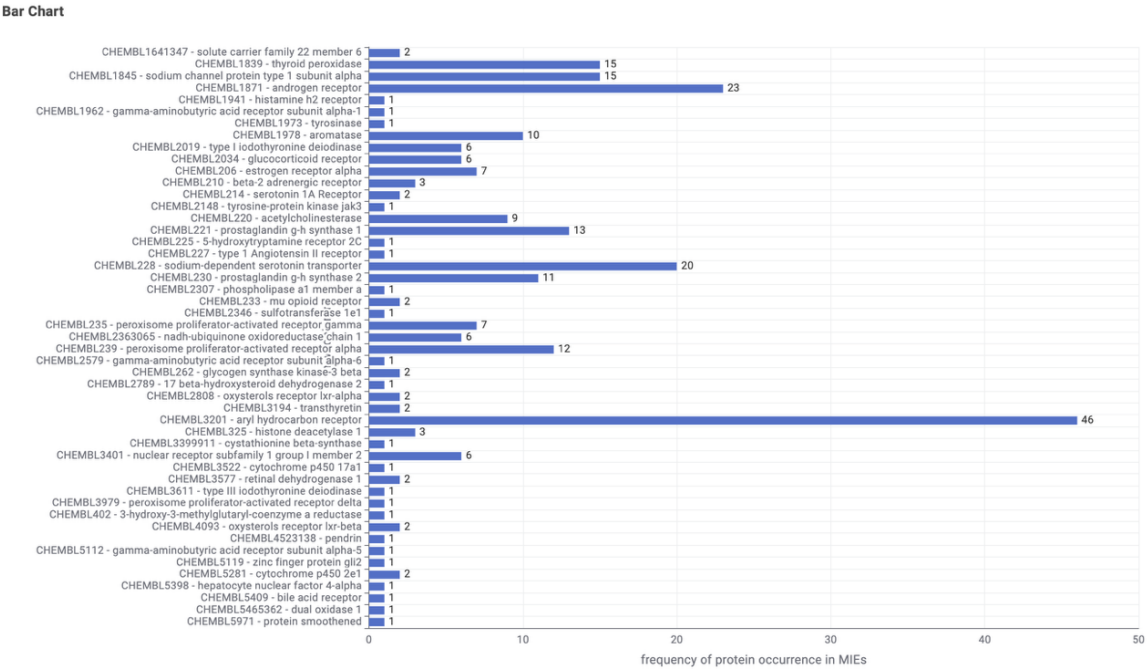

Figure 19: Bar chart showing the frequency of protein occurrences in MIEs.

Figure 20 shows a bar graph of the number of bioactivity data points available in ChEMBL35 for each target. In total, 354.689 bioactivity records were retrieved, with retinal dehydrogenase 1 showing by far the highest number of entries (77.059), followed by the μ-opioid receptor (20.441) and acetylcholinesterase (18.946). The fourth most data-rich target was estrogen receptor alpha (18.287), which was also the most frequently represented target among those of interest in this thesis. On the other hand, dual oxidase (1) and zinc finger protein GLI2 (12) showed the lowest number of bioactivity data.

This imbalance in the distribution of biological activity data among the targets can be explained by the differing research focus on each protein. Retinal dehydrogenase 1 seems to be of much interest for toxicological research, whereas targets like dual oxidase and zinc protein GLI2 may only be studied in specific pathological cases. Due to the low amount of data, it is difficult to build robust and reliable QSAR models for such targets, which highlights the need for more research in these areas.

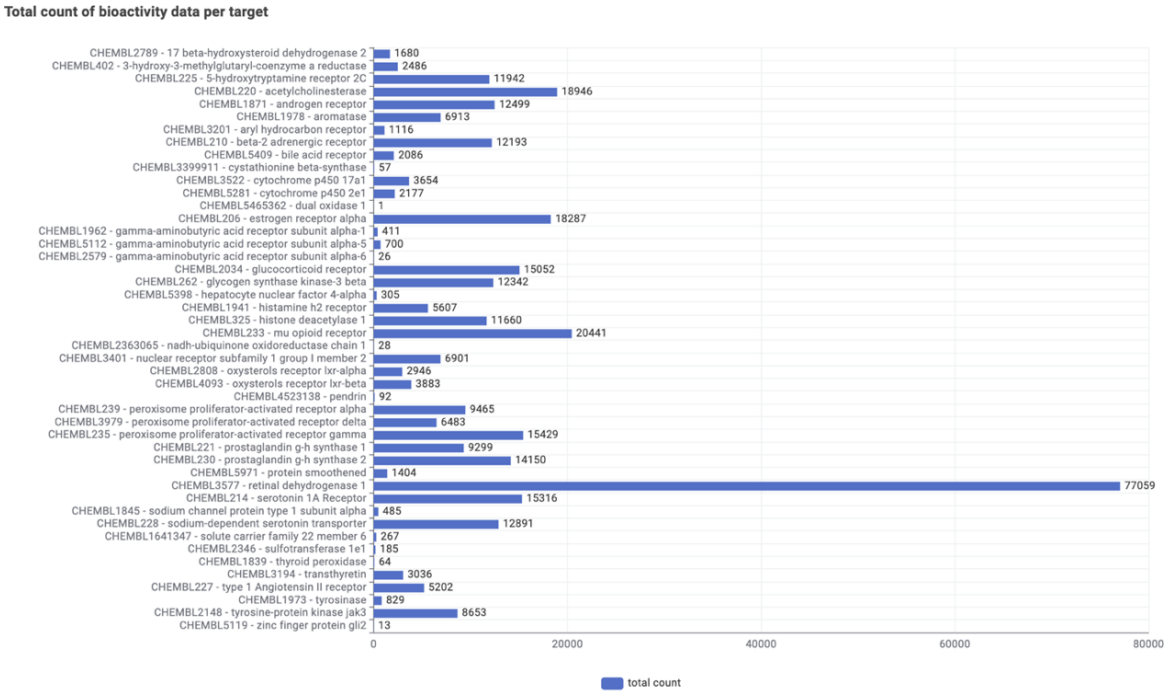

Figure 20: Bar chart showing the total count of bioactivity data per target.

4.1.2. Grouping the Adverse Outcomes

As described previously, all MIEs were grouped according to the adverse outcomes they ultimately led to within their respective AOPs. Out of the 190 identified AOs, the most frequently reported events are listed in Table 3. The most common outcomes included a decrease in population growth rate, reduced growth, increased mortality, hyperinflammation, and increased liver steatosis.

AO-ID	Event	Count
360	Decrease, Population growth rate	65
1521	Decrease, Growth	30
351	Increased Mortality	24
1868	Hyperinflammation	24
459	Increased, Liver Steatosis	20

Table 3: The five most occurred AOs across AOPs.

Most AOs could not be clearly assigned to a specific organ or body system and were therefore categorized as „Unclear/Other.“ When considering only clearly defined organ systems, the nervous system with 16 targets, the reproductive system with 15 targets, and the liver with 14 targets were the most frequently affected. It is important to note that targets may be involved in multiple groups, as they can contribute to several adverse outcomes through different pathways.

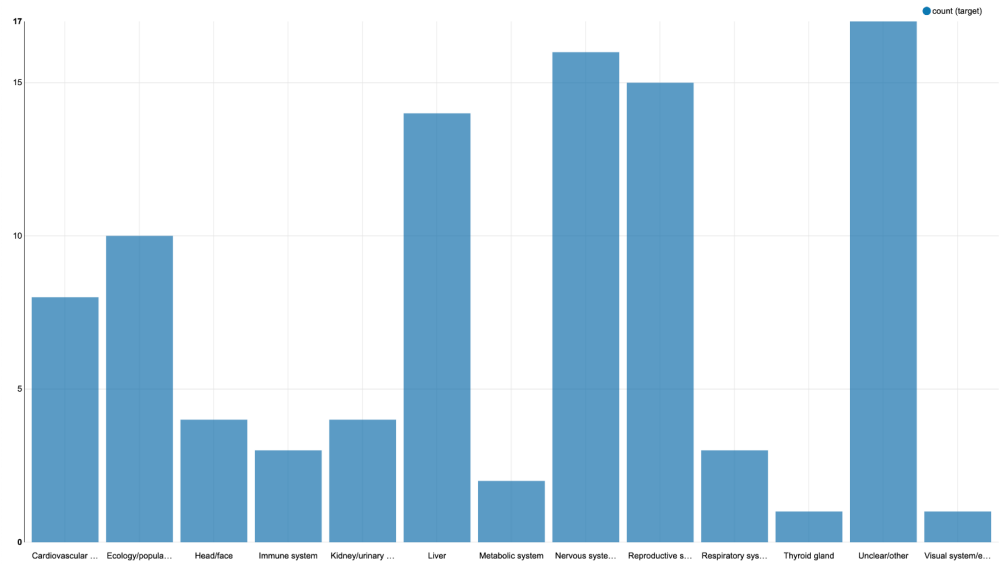

Figure 21: Bar chart showing the number of targets occurring in the body system/organ groups.

In the further stages of this thesis, the focus was placed on the reproductive system, which was identified as the target organ for adverse outcomes in 15 targets, which were:

Target name	ChEMBL ID
thyroid peroxidase	CHEMBL1839
androgen receptor	CHEMBL1871
aromatase	CHEMBL1978
estrogen receptor alpha	CHEMBL206
beta-2 adrenergic receptor	CHEMBL210
prostaglandin g-h synthase 1	CHEMBL221
sodium-dependent serotonin transporter	CHEMBL228
prostaglandin g-h synthase 2	CHEMBL230
sulfotransferase 1e1	CHEMBL2346

peroxisome proliferator-activated receptor alpha	CHEMBL239
aryl hydrocarbon receptor	CHEMBL3201
histone deacetylase 1	CHEMBL325
cytochrome p450 17a1	CHEMBL3522
retinal dehydrogenase 1	CHEMBL3577
3-hydroxy-3-methylglutaryl-coenzyme a reductase	CHEMBL402

Table 4: List of the twelve AOP-derived targets associated to DART and their corresponding ChEMBL ID

Of these, 12 targets were eligible for the development of regression models based on the analysis described in Chapter 3.6.3 and seen in Figure 22. Three targets (aryl hydrocarbon receptor, thyroid peroxidase, sulfotransferase 1e1) were dismissed for the same reason: they did not have at least 100 bioactivity data records. Since models trained on very small datasets would have limited learning capacity, reliable predictions cannot be expected. For this reason, those three molecular targets were excluded. This leads to a loss of information about the pathways leading to DART. Similar to the exclusion of protein families and complexes, this limits the ability to fully capture the biological complexity underlying DART-related adverse outcomes.

chembl_id	mie_ids	lowest activity	highest activity	% active	% inaactive	Regression	Classification	count_IC50	count_Ki
CHEMBL1839	[279]	4,23	6,23	16,7	83,3	Regression ✗	Classification ✗	42	
CHEMBL1871	[25, 26, 1614, 1617]	4,08	9,7	38,1	61,9	Regression OK	Classification OK	1972	888
CHEMBL1978	[36, 408, 2293]	4	10,82	57,6	42,4	Regression OK	Classification OK	2748	497
CHEMBL206	[1065, 1710, 2126]	4	11	47,6	52,4	Regression OK	Classification OK	2624	487
CHEMBL210	[1038, 1039]	3,845	10,92	41,1	58,9	Regression OK	Classification OK	601	558
CHEMBL221	[79, 1103]	4	9	22,1	77,9	Regression OK	Classification ✗	1606	3
CHEMBL228	[619, 1317, 1397]	4	11	54,7	45,3	Regression OK	Classification OK	2684	2354
CHEMBL230	[79]	4	10,7	54,8	45,2	Regression OK	Classification OK	4045	26
CHEMBL2346	[2155]	4,7	6,6	33,3	66,7	Regression ✗	Classification ✗	3	
CHEMBL239	[227, 231, 468, 998, 2219]	4,05	9,55	21,2	78,8	Regression OK	Classification ✗	768	131
CHEMBL3201	[18]	4,6	9,7	55,6	44,4	Regression ✗	Classification ✗	19	8
CHEMBL325	[1502]	4	10,4	71,6	28,4	Regression OK	Classification ✗	6380	124
CHEMBL3522	[1609]	4,03	10,77	86,9	13,1	Regression OK	Classification ✗	718	24
CHEMBL3577	[1880]	4,01	7,92	67,2	32,8	Regression OK	Classification OK	298	13
CHEMBL402	[804]	4,05	10,85	72,8	27,2	Regression OK	Classification ✗	223	9

Figure 22: Overview of the analysis whether the targets where eligible for classification and regression modeling. Mie_ids = mie ids that the target occurred in. Lowest activity = lowest pChEMBL value. Highest activity = highest pChEMBL value. % active = amount of pChEMBL values over the IDG threshold. % inactive = amount of pChEMBL values under the IDG threshold. Count_IC50 = number of IC50 data. Count_Ki = number of Ki data.

Comparing Figure 20 to Figure 22, it is clear that there is a significant loss of bioactivity data points of retinal dehydrogenase 1 (CHEMBL3577). From being the target with the most bioactivity data, there now only remain 298 unique compounds with IC50 values and 13 compounds with Ki values that can be used as a training set. This can be explained by

the fact that there are different types of bioactivity data. For retinal dehydrogenase 1, most available data consisted of potency measurements, while very few entries were left as IC50/Ki values. This shows that the selection of a specific bioactivity type can influence data size and consequently model quality.

4.2. Results of model building

4.2.1. Comparison of IC50 and Ki values

At the beginning of the study, IC50 and Ki values were considered for modeling purposes, independent of the underlying biology of the targets. The possibility of combining both activity measures for modeling was evaluated. For this purpose, compounds with both IC50 and Ki values available were identified for each target. These values were plotted against each other in scatter plots, and R^2 was calculated to assess their correlation.

As shown in Figure 23, most targets exhibited very low or even negative R^2 values, indicating a weak relationship between IC50 and Ki values. Only two targets showed comparatively strong correlations: CHEMBL228 with an R^2 value of 0.696 and CHEMBL239 with an R^2 value of 0.854; however, the latter was based on only 11 data points.

Due to the overall weak correlation between activity measures and – for most of the targets – a small number of Ki values, Ki data were excluded from further analysis. Consequently, all subsequent modeling was performed exclusively using IC50 values to ensure a consistent and homogeneous modeling approach across all targets. Violin plots were additionally generated to support this decision; however, they did not provide strong additional evidence and can be found in the Appendix.

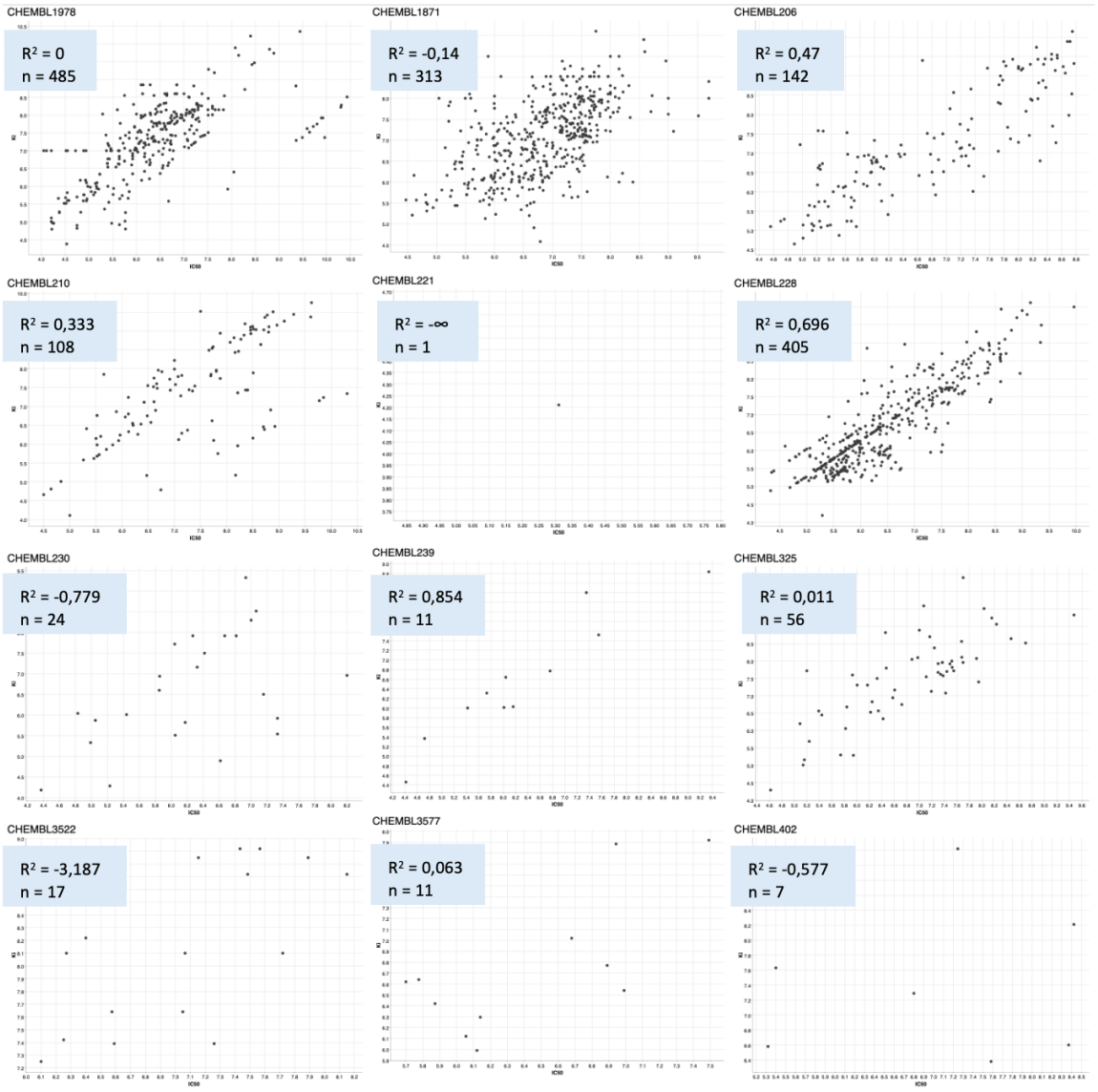

Figure 23: Scatter plots showing the correlation between IC50 and Ki values of each Target. X-axis = IC50. Y-axis = Ki

4.2.2. Hyperparameter tuning

As described in Chapter 3.6.5., hyperparameter tuning was performed for optimizing the model performances. For each target, three different algorithms each with three distinct hyperparameter settings were configurated. As the strategy to find the best combinations of hyperparameters, a GridSearch was applied, which tests every possible combination to obtain the one with the highest R^2.

XGBoost

eta	0.05, 0.15, 0.25
maxLevels	3, 5, 7
nrModels	300, 600, 900

Random Forest

maxLevels	10, 20, 30, 40
nrModels	100, 200, 300, 400, 500
minChildSize	1, 3, 5, 7

Tree Ensemble

maxLevels	5, 15, 25
nrModels	100, 200, 300, 400, 500
minChildSize	1, 3, 5

The hyperparameter eta describes the learning rate of the model, where lower values result in slower but more precise learning (Tiwari et al., 2025). MaxLevels stands for maximum depth of each tree. High values allow the model to train on more complexity within the dataset (Sandunil et al., 2024). NrModels stands for the number of trees, where a larger number generally improves model stability. However, a higher number does not necessarily guarantee better performance and may only increase computational time and memory. MinChildSize determines the minimum of data points required in a leaf node. Lower values allow the model learn more precisely but increase the risk of overfitting (Abubakar et al., 2025).

Figure 24 summarizes all R^2 values obtained from cross validation and from the holdout test set, as well as the best combination of hyperparameters for each algorithm. The rightmost column indicates the number of data points used for model development. Rows colored in gray highlight the algorithms that achieved the best performance based on the R^2_test values. For further analyses, such as outlier detection and model validation, only these gray highlighted settings were used.

target	model	R^2_cv	R^2_test	maxLevels	nrModels	eta	minChildSize	compound_count
CHEMBL1871 - androgen receptor	xgboost	0,537	0,527	5	300	0,05		1891
	random forest	0,519	0,528	30	100		1	
	tree ensemble	0,517	0,542	25	400		3	
CHEMBL1978 - aromatase	xgboost	0,579	0,615	3	900	0,05		2439
	random forest	0,565	0,591	30	500		3	
	tree ensemble	0,565	0,597	15	500		5	
CHEMBL206 - estrogen receptor alpha	xgboost	0,734	0,779	5	300	0,05		2495
	random forest	0,722	0,758	40	400		1	
	tree ensemble	0,721	0,763	25	500		3	
CHEMBL210 - beta-2 adrenergic receptor	xgboost	0,675	0,655	3	300	0,05		568
	random forest	0,676	0,703	20	200		1	
	tree ensemble	0,672	0,695	25	100		1	
CHEMBL221 - prostaglandin g-h synthase 1	xgboost	0,358	0,341	5	300	0,05		1567
	random forest	0,346	0,356	20	400		1	
	tree ensemble	0,341	0,363	25	500		3	
CHEMBL228 - sodium-dependent serotonin transporter	xgboost	0,565	0,550	5	600	0,05		2371
	random forest	0,544	0,543	20	500		1	
	tree ensemble	0,547	0,548	25	300		1	
CHEMBL230 - prostaglandin g-h synthase 2	xgboost	0,483	0,508	7	300	0,05		3948
	random forest	0,479	0,487	40	400		1	
	tree ensemble	0,481	0,492	25	300		3	
CHEMBL239 - ppar alpha	xgboost	0,446	0,463	5	600	0,05		709
	random forest	0,446	0,502	20	400		1	
	tree ensemble	0,458	0,502	15	200		1	
CHEMBL325 - histone deacetylase 1	xgboost	0,629	0,611	7	900	0,05		6250
	random forest	0,578	0,578	40	500		1	
	tree ensemble	0,579	0,581	25	500		1	
CHEMBL3522 - cytochrome p450 17a1	xgboost	0,355	0,452	7	300	0,15		669
	random forest	0,374	0,448	30	500		3	
	tree ensemble	0,377	0,458	15	200		5	
CHEMBL3577 - retinal dehydrogenase 1	xgboost	0,559	0,329	5	300	0,25		295
	random forest	0,566	0,318	20	200		1	
	tree ensemble	0,567	0,336	25	400		1	
CHEMBL402 - HMG-CoA-reductase	xgboost	0,714	0,743	5	300	0,15		211
	random forest	0,729	0,754	30	300		1	
	tree ensemble	0,723	0,737	25	100		3	

Figure 24: An overview of the results of hyperparameter tuning, cross-validation and test set. Yellow = target name. Purple = algorithms. Pink = R^2_cv - R^2 for cross-validation, R^2_test - R^2 for test set. Blue = best hyperparameter combination for each algorithm. Green = amount of compounds used to model. Rows highlighted in grey = the algorithm with the highest R^2_test which are used for subsequent analyses.

As shown in Figure 24, CHEMBL206 (estrogen receptor alpha) and CHEMBL402 (HMG-CoA reductase) had the highest R^2 values for both cross-validation and test set. The dataset of CHEMBL206 comprised 2.495 unique compounds, which can be considered sufficiently large to cover a broad chemical space. This could mean that the model was trained on diverse structural information and that the outcome values are unlikely to be the result of random correlations. In contrast, the same conclusions cannot be drawn for CHEMBL402, as only 206 compounds were available. However, R^2_cv and R^2_test are very similar which indicates that there is likely no overfitting.

All remaining targets, except CHEMBL221 (prostaglandin G/H synthase 1), CHEMBL3522 (CYP17A1) and CHEMBL3577 (Retinal dehydrogenase 1), had R^2 values greater than 0,5 which suggests that the models could be of moderate to good performance. As for the

previously mentioned three targets, R^2 values ranging from 0,33 to 0,45 were obtained indicating a lower predictive ability and model stability.

In most cases, R^2_cv and R^2_test were very similar, suggesting stable model behavior, good generalization and no overfitting.

4.2.3. Outlier detection and residual calculation

Even models with good predictive performance may contain individual data points that exhibit large deviations between observed and predicted values. Such data points are defined as outliers and fall outside a predefined tolerance range. They are quantified by calculating the residual, defined as the difference between the observed and the predicted value. If the absolute residual exceeds a value of 2, this corresponds to an error of approximately two orders of magnitude, and the respective data point is classified as an outlier. Residuals may indicate errors arising during data acquisition or reflect limitations of the model, particularly when a molecule lies outside the chemical space covered by the training set.

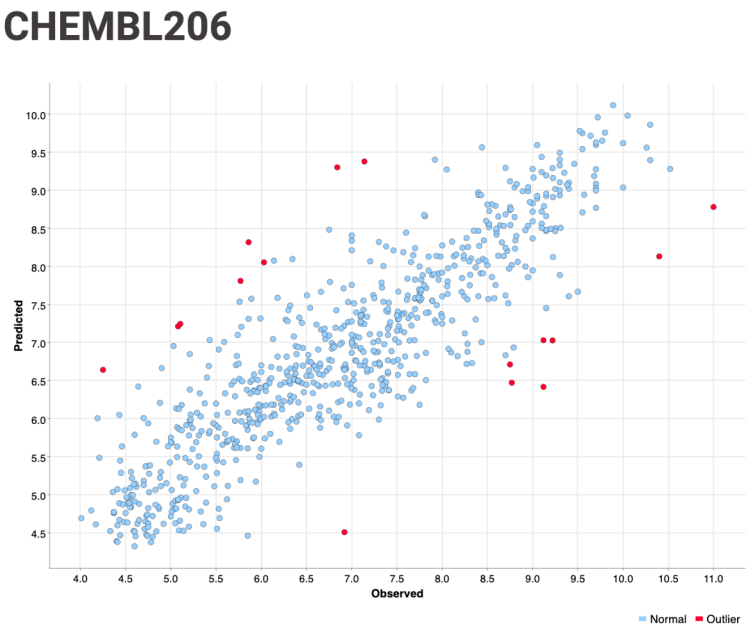

Figure 25: Scatter plot for outlier detection of CHEMBL206 (estrogen receptor alpha). X-axis = observed values. Y-axis = Predicted values. Blue dots = compounds with residual < 2. Red dots = compounds with residual > 2.

For visualization purposes, scatter plots were generated in which observed and predicted values of the test set are plotted against each other. Outliers are highlighted in red. In principle, these outliers could be further investigated with respect to their structural characteristics or potential inaccuracies in the underlying literature. However, since the focus of this thesis is on model evaluation and validation, such in-depth analyses were not performed. As an example, Figure 25 shows the scatter plot for estrogen receptor alpha (CHEMBL206), in which a total of 25 outliers were identified. The corresponding scatter plots for the remaining targets are provided in the Appendix.

4.3. Evaluation of the QSAR models

4.3.1. Results of the updated ChEMBL release prediction

The updated ChEMBL release provided additional compounds for the 12 investigated targets; however, their distribution across targets was not homogeneous. Some targets were associated with a substantially higher number of newly available compounds than others.

These new compounds were subjected to the developed models for prospective validation, resulting in predicted plC50 values for each target. The availability of observed activity values enabled the calculation of R^2 and creating scatter plot analyses for the validation dataset.

In the scatter plots, observed plC50 values are shown on the x-axis, while predicted plC50 values are displayed on the y-axis. The strength of the correlation between observed and predicted values varied considerably among targets. Therefore, the scatter plots of the ten analyzed targets were grouped into three broad categories based on their respective R^2 values and interpreted accordingly.

Group 1: Targets with Moderate Predictive Performance

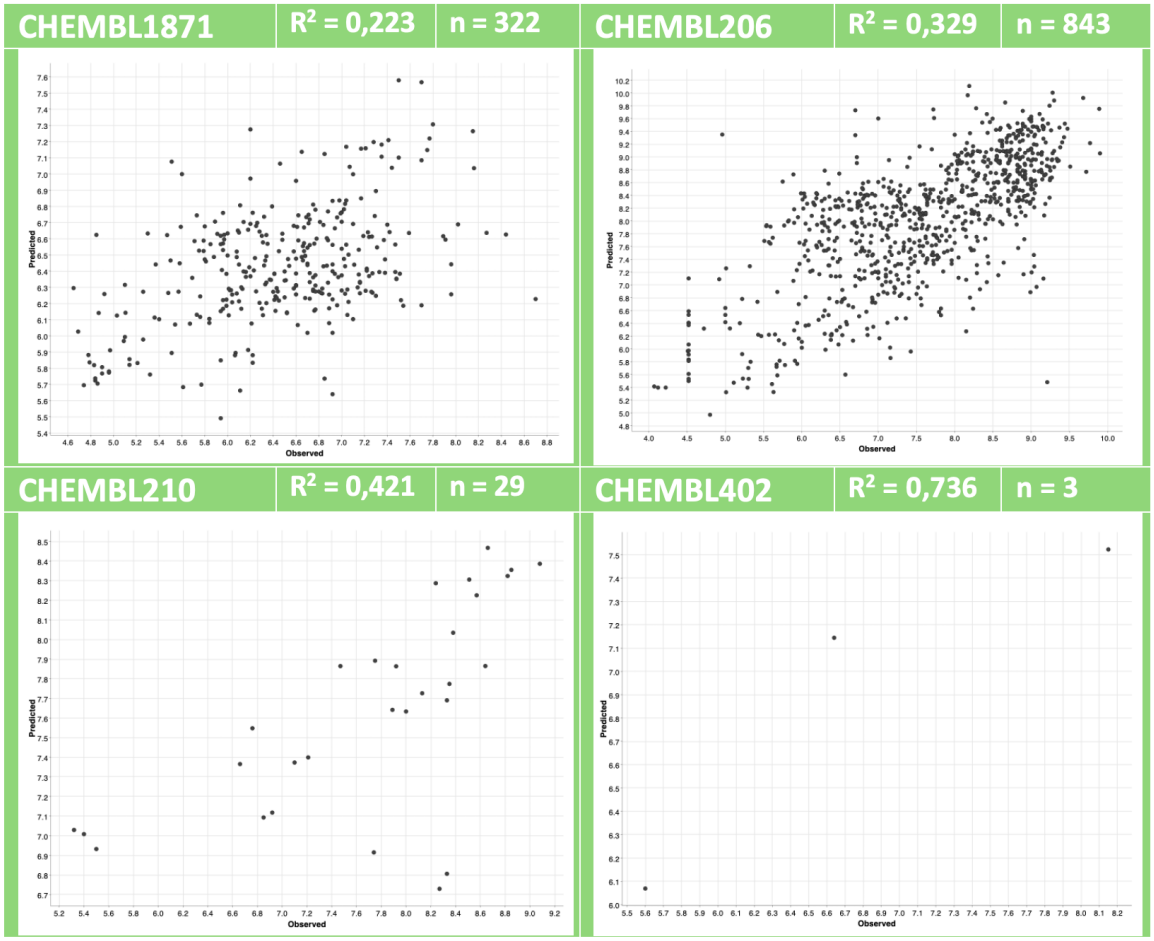

Figure 26: Scatter plots of observed vs. predicted values for prospective validation of new compounds from ChEMBL36 for CHEMBL1871, CHEMBL206, CHEMBL210 and CHEMBL402. X-axis = observed values. Y-axis = predicted values. n = number of new compounds.

The first target in this group is the androgen receptor, which exhibits a rather weak but positive correlation between observed and predicted values. The large number of newly available compounds suggests that this result is not random but reflects a certain degree of statistical stability.

The second target is estrogen receptor alpha, which shows an R^2 value greater than 0,3, indicating moderate predictive performance and a consistent relationship between observed and predicted values. The large sample size supports the statistical reliability of this model. The third target, the $ß_2$-adrenergic receptor, displays the highest R^2 value withing this group, although the sample size is only marginally sufficient. Despite the limited number of new compounds, a positive correlation between observed and

predicted activities is observed, suggesting reasonable generalizability to novel compounds.

The fourth scatter plot corresponds to HMG-CoA reductase. Although a very high R^2 value is observed, the insufficient sample size prevents any meaningful assessment of statistical significance.

Group 2: Targets with Very Low Predictive Performance

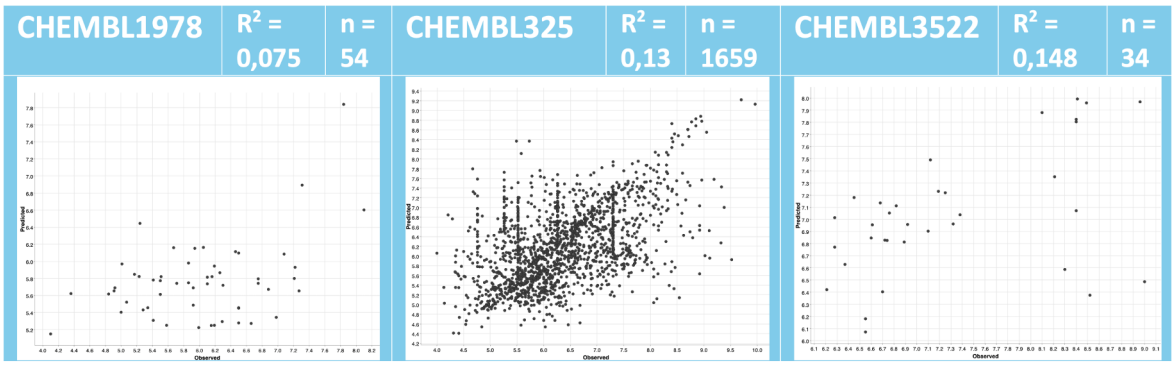

Figure 27: Scatter plots of observed vs. predicted values for prospective validation of new compounds from ChEMBL36 for CHEMBL1978, CHEMBL325 and CHEMBL3522. X-axis = observed values. Y-axis = predicted values. n = number of new compounds.

For the first target in this group, aromatase, the model explains only a negligible fraction of the variability in the observed activity data. Consequently, the model shows poor generalizability to new chemical structures.

The second target, histone deacetylase 1, is associated with the largest number of new compounds. Nevertheless, despite the large sample size, the model exhibits low predictive performance and limited generalization capability.

The third target, CYP17A1, combines a small number of new compounds with a low R^2 value. This indicates a negative correlation between observed and predicted activities and demonstrates that the model fails to capture general trends in the data.

Group 3: Targets with Negative R^2 Values

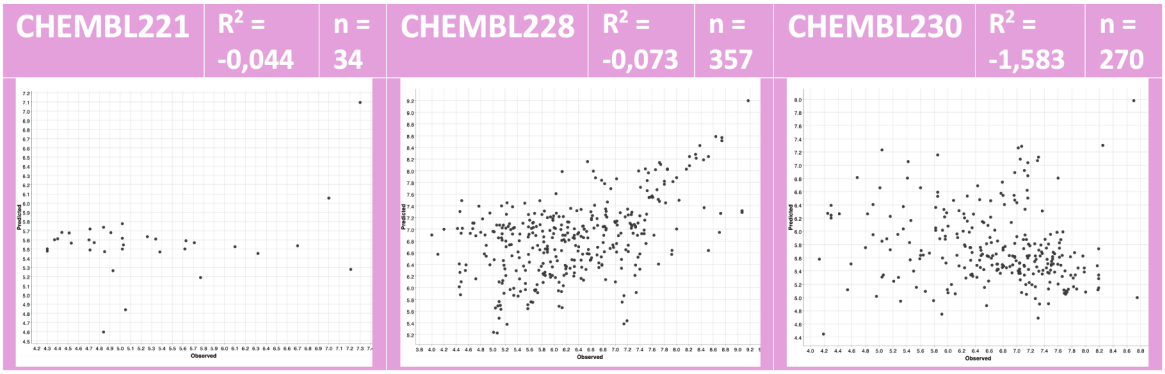

Figure 28: Scatter plots of observed vs. predicted values for prospective validation of new compounds from ChEMBL36 for CHEMBL221, CHEMBL228 and CHEMBL230. X-axis = observed values. Y-axis = predicted values. n = number of new compounds.

For the three targets prostaglandin G/H synthase 1, sodium-dependent serotonin transporter, and prostaglandin G/H synthase 2, similar conclusions can be drawn. The models exhibit a very limited ability to identify trend structures or generate reliable predictions. Negative R^2 values indicate the absence of a linear relationship between observed and predicted values, meaning that the models perform worse than a naive predictor based solely on the mean of the observed activities.

For two targets, namely peroxisome proliferator-activated receptor alpha and retinal dehydrogenase 1, the R^2 values were undefined ($-\infty$) because only a single new compound was available for testing, which is insufficient for the calculation of R^2.

Overall, the R^2 values obtained for the ChEMBL36 validation set are, as expected, lower than those observed for the holdout test set. Low R^2 values are not uncommon in external validation, as newly introduced compounds may differ substantially from the training data and may lie outside the chemical space covered by the models. Additionally, the molecular descriptors employed may have been insufficient, and the inclusion of alternatives such as molecular fingerprints, could improve model stability. Furthermore, for several targets, the limited number of new compounds restricts the statistical interpretability of the results.

Nevertheless, some models demonstrate a certain degree of predictive capability and generalizability toward previously unseen compounds. This validation approach allows for

a realistic assessment of model performance and facilitates the identification of targets and models that require further optimization.

4.3.2. Results of the DNT-list prediction

The DNT list from the RiskHunt3r project is a compilation of chemical compounds that have been evaluated by different authors. Based on the authors´ individual assessments, the compounds were classified as either DNT-positive or DNT-negative in the column called „consensus verdict“. The list included a total of 705 organic compounds which were used as a classification-based validation set for machine learning models. It is imported to note that within the DNT list, 17 compounds were categorized as „clash“ (i.e., conflicting author comments) and 248 compounds as „waiver-only“ (i.e., no experimental data available). These compounds were excluded from further analysis, as a reliable comparison was not possible. Consequently, 440 compounds remained for the evaluation of model performance.

For an analysis of whether the substances pose a potential DNT hazard, a *Rule Engine* node was used in KNIME. The rule was defined such that if the predicted plC50 value of a compound exceeded the IDG-threshold for at least one target, the compound was assigned a „positive“ label in a newly generated column named „comment“. If the predicted plC50 values for all targets were below the threshold, the compound was labeled as „negative“.

As shown in Figures 29-33, predicted plC50 values were tabulated for each compound across all targets. Since no experimentally observed values were available, calculation of an R^2 value was not possible. Therefore, the predicted values were binarized based on whether they were above or below the respective IDG thresholds, as described above. For visualization purposes only, the plC50 values were color-scaled in the table, with the IDG-threshold set as the minimum value – all plC50 values above the threshold were highlighted in red. This visualization helps to rapidly assess the target-based positive/negative distribution of compounds.

It is important to note that the computational approach applied here includes only 12 AOP-derived targets and therefor does not cover all possible biological pathways that may lead to DNT. Consequently, if none of the 12 models predicts a substance classified as DNT-positive by the RiskHunt3r project as positive, this does not necessarily indicate that the substance is non-toxic. Rather, it suggests that the substance induces DNT via different mechanisms.

Preferred Name	Consensus verdict	CHEMBL1871	CHEMBL1978	CHEMBL206	CHEMBL210	CHEMBL221	CHEMBL228	CHEMBL230	CHEMBL239	CHEMBL325	CHEMBL3522	CHEMBL3577	CHEMBL402	comment
1-Methyl-4-phenylpyridinium iodi	Positive	5,94	5,32	4,22	5,98	5,83	4,19	5,08	5,18	4,83	5,67	6,13	6,14	Positive
2-Methoxyethanol	Positive	6,39	5,57	4,64	6,10	5,41	5,89	4,91	5,07	4,65	5,82	6,27	5,93	Positive
3,3'-Iminobispropanenitrile	Positive	7,12	6,04	4,70	6,32	5,46	6,28	5,35	5,22	4,63	5,91	6,13	5,99	Positive
5-Fluorouracil	Positive	6,19	5,98	4,68	6,18	5,56	5,61	5,26	5,29	4,72	5,83	5,89	5,93	Negative
Oxidopamine hydrochloride	Positive	6,15	4,60	4,21	6,86	5,13	5,45	5,18	5,32	4,75	5,70	5,84	6,09	Positive
Acrylamide	Positive	6,45	5,14	4,31	6,02	5,35	4,96	5,30	5,06	5,09	5,66	6,17	5,81	Positive
Aldicarb	Positive	6,40	5,52	4,71	6,30	5,36	5,71	5,41	5,24	4,80	6,12	6,10	5,83	Positive
Captan	Positive	6,52	5,15	5,41	6,61	5,79	5,33	5,90	5,46	4,94	6,55	6,18	5,84	Positive
Carbaryl	Positive	6,30	5,95	4,72	6,74	5,48	5,91	5,55	5,15	5,42	5,87	6,22	6,11	Positive
Chlorpyrifos	Positive	5,95	6,11	5,40	6,42	5,63	5,94	5,50	5,78	5,70	6,83	5,87	5,88	Positive
Colchicine	Positive	6,09	5,74	4,84	6,87	5,29	5,82	5,36	5,98	5,58	6,60	5,49	6,56	Positive
Deltamethrin	Positive	6,37	6,93	6,00	6,74	5,11	5,59	5,50	5,91	5,98	6,50	5,99	6,08	Positive
Di(2-ethylhexyl) phthalate	Positive	5,79	4,89	5,68	6,29	4,94	5,56	5,25	6,34	5,42	6,46	5,81	5,81	Positive
DDT	Positive	5,67	5,84	5,87	6,12	6,24	5,84	6,04	5,51	5,39	6,62	5,89	5,89	Positive
Dieldrin	Positive	6,94	6,08	5,88	6,36	5,67	6,28	5,60	5,70	4,78	6,11	6,17	5,81	Positive
Diethylstilbestrol	Positive	5,65	5,25	7,80	7,37	5,45	4,96	5,48	5,38	5,39	6,50	6,03	6,16	Positive
Heptachlor	Positive	6,90	6,53	5,64	6,24	5,98	5,48	5,49	5,57	4,74	6,20	6,15	5,73	Positive
Hexachlorophene	Positive	5,96	6,01	5,73	6,91	5,96	5,37	6,21	5,75	5,19	6,60	5,90	6,17	Positive
Hydroxyurea	Positive	6,33	5,52	4,53	6,22	5,36	4,77	5,19	5,20	5,33	5,62	5,97	5,76	Negative
Lindane	Positive	6,82	5,77	4,99	6,27	5,90	5,61	5,52	5,57	4,98	6,05	6,39	5,77	Positive
n-Hexane	Positive	6,43	5,55	4,38	6,17	5,78	4,98	4,52	5,02	5,05	5,65	6,35	5,77	Positive
Permethrin	Positive	5,92	5,47	6,84	6,93	5,13	5,17	5,46	5,85	5,65	6,60	5,83	5,89	Positive
Phenobarbital sodium	Positive	5,88	4,26	4,44	5,93	5,86	5,64	6,42	5,58	4,98	6,07	5,82	5,98	Positive
Rotenone	Positive	6,38	5,52	5,02	7,16	5,10	6,24	5,29	5,92	5,09	6,57	5,64	6,48	Positive
Tebuconazole	Positive	5,64	6,67	5,09	6,44	5,40	6,12	5,93	5,81	5,51	6,86	5,82	5,77	Positive
Tetraethylthiuram disulfide	Positive	6,34	5,59	5,12	6,42	5,66	5,86	5,17	5,69	6,02	6,53	6,51	5,89	Positive
Thalidomide	Positive	6,11	4,81	4,58	5,91	5,78	5,97	5,80	5,55	5,77	6,29	5,63	6,00	Positive
Toluene	Positive	6,26	5,53	4,50	5,92	5,87	4,56	4,90	5,07	4,82	5,69	6,27	5,91	Positive
Valinomycin	Positive	6,01	5,24	6,42	7,09	5,36	5,68	4,59	7,01	5,65	6,09	6,29	6,06	Positive
Acetaminophen	Negative	6,39	4,88	4,54	6,39	5,18	4,61	5,31	5,09	5,49	5,75	6,16	5,96	Positive
Ampicillin	Negative	6,15	4,95	5,28	6,31	5,49	5,83	5,86	5,87	5,74	6,83	5,79	5,81	Positive
Aspirin	Negative	6,01	4,87	4,30	6,51	5,19	4,31	5,46	5,19	4,60	5,90	5,89	5,98	Negative
Buspirone	Negative	6,06	5,93	5,63	5,74	5,07	6,15	5,39	6,05	6,50	7,12	5,79	5,87	Positive
Chloramben	Negative	6,39	5,86	4,58	6,08	5,52	4,34	5,67	5,22	5,15	5,73	6,08	5,94	Positive
Chlorpheniramine maleate	Negative	5,90	5,26	5,57	6,21	5,91	7,53	6,20	5,36	5,19	6,76	5,94	5,96	Positive
Cotinine	Negative	6,30	4,60	4,78	5,97	5,58	5,87	5,84	5,15	5,06	5,93	6,26	6,07	Positive
Doxylamine succinate	Negative	5,70	5,43	5,50	6,03	5,90	6,84	5,80	5,43	5,47	6,68	5,96	5,97	Positive
Erythromycin	Negative	6,04	5,03	6,31	6,88	5,18	4,96	5,54	7,23	5,31	6,21	6,30	6,44	Positive
Famotidine	Negative	6,38	6,77	5,04	6,00	5,67	5,73	5,50	5,65	5,95	6,88	5,56	5,61	Positive
Fluconazole	Negative	5,77	4,56	4,75	6,31	5,54	5,30	5,49	5,79	5,41	6,65	5,60	6,02	Positive
Folic acid	Negative	6,05	5,54	5,01	6,01	5,56	5,38	5,45	5,89	5,48	6,67	5,52	6,01	Positive
Phenol	Negative	6,28	4,94	4,44	6,48	5,46	4,62	5,45	5,11	4,66	5,69	6,10	5,93	Positive
Sulfisoxazole	Negative	6,33	5,42	5,02	5,94	5,44	5,79	5,22	5,45	5,62	6,51	5,80	6,03	Positive
Tetracycline	Negative	6,42	5,27	5,09	6,39	5,11	5,57	5,38	6,07	5,54	6,32	5,73	6,03	Positive
Dinotefuran	Negative	6,18	5,95	4,78	5,98	5,53	6,26	5,38	5,49	6,01	6,12	5,70	5,89	Positive
Glycerol	Negative	6,31	5,25	4,43	6,33	5,30	5,58	4,89	5,20	4,46	5,74	6,10	5,91	Positive
Ibuprofen	Negative	6,38	4,95	5,00	5,99	5,66	4,91	5,44	5,17	5,12	5,96	6,30	5,89	Positive
Isoniazid	Negative	6,21	6,03	4,39	5,98	5,56	5,42	5,04	5,45	4,82	5,80	5,93	5,83	Positive
L-Ascorbic acid	Negative	6,24	5,04	4,64	6,96	5,08	4,44	4,70	5,41	4,62	5,85	5,71	6,03	Positive
Metformin	Negative	6,61	5,84	4,37	6,40	5,35	5,95	5,20	5,43	5,47	5,82	6,17	5,72	Positive
Omeprazole	Negative	5,94	6,03	4,70	6,78	5,61	7,09	6,21	5,86	5,58	7,05	5,56	6,51	Positive
Penicillin VK	Negative	6,05	5,25	5,37	6,57	5,36	5,73	5,96	5,84	5,46	6,89	5,73	5,77	Positive
Saccharin	Negative	5,99	5,54	4,70	5,91	5,38	4,07	5,65	5,39	5,02	6,03	6,07	5,84	Positive
Sodium benzoate	Negative	6,20	5,26	4,52	5,98	5,41	4,36	5,53	5,09	4,68	5,80	6,01	5,90	Positive
Anthracene	Negative	5,82	4,91	5,07	6,01	6,08	4,27	5,38	5,17	5,05	5,82	6,19	5,90	Positive
Metoclopramide	Negative	6,11	5,33	4,81	6,59	5,52	6,39	5,39	5,77	5,75	6,49	5,79	5,88	Positive
Amitriptyline	Negative	5,93	5,39	6,25	6,39	6,04	7,51	5,50	5,56	5,49	6,56	5,98	5,79	Positive
Metoprolol	Negative	5,93	4,36	4,52	5,99	5,39	5,62	5,01	5,84	5,55	6,42	5,73	5,87	Positive
Sumatriptan	Negative	6,02	6,23	5,51	6,11	5,37	6,62	5,11	5,74	5,36	7,14	5,60	6,14	Positive
Amoxicillin	Negative	6,24	5,28	5,42	6,89	5,49	5,52	5,98	5,92	6,34	6,67	5,81	5,75	Positive
Diphenhydramine	Negative	5,50	4,87	5,39	6,30	5,84	6,28	5,28	5,44	4,93	6,42	6,04	6,02	Positive
Pomalidomide	Negative	6,21	4,61	5,20	5,91	5,76	5,96	5,78	5,60	5,51	6,26	5,55	5,92	Positive
Warfarin	Negative	5,69	6,05	5,19	6,99	5,17	5,17	5,60	5,45	5,72	6,71	5,75	6,00	Positive
Captopril	Negative	6,81	5,17	4,49	6,23	5,55	5,63	5,09	5,32	5,19	6,13	5,87	5,82	Positive
Dabigatran	Negative	5,84	6,51	5,80	5,92	5,38	6,02	5,84	5,91	6,64	7,30	6,01	6,42	Positive
N,N-Dimethylformamide	Negative	6,59	5,57	4,51	6,12	5,41	5,84	4,97	5,08	4,84	5,80	6,27	5,78	Positive
Dimethyl sulfoxide	Negative	6,50	5,17	4,91	6,09	5,55	5,63	5,09	4,95	4,77	5,85	6,25	5,85	Positive
D-Glucitol	Negative	6,28	5,27	4,42	6,63	5,22	4,62	4,62	5,32	4,52	5,84	5,82	5,75	Negative
Lactose	Negative	6,16	5,02	4,44	6,03	4,96	5,01	5,24	5,81	4,79	6,04	5,59	5,80	Positive
Glucosamine	Negative	6,41	4,66	4,39	6,51	5,07	4,90	4,76	5,46	4,65	5,60	5,84	5,81	Negative
Diethylene glycol	Negative	6,20	5,28	4,55	6,21	5,42	5,90	5,13	5,15	4,64	5,88	6,09	5,90	Positive
(+/-)-Deprenyl	Negative	6,63	6,07	4,59	6,23	5,83	5,82	5,47	5,16	4,78	6,05	6,08	6,08	Positive
Trolox	Negative	6,64	5,82	5,25	7,44	6,01	5,97	5,65	5,56	4,84	6,13	5,93	5,97	Positive
Z-VAD (OMe)-FMK	Negative	5,75	4,57	4,79	6,32	5,43	5,42	5,39	6,47	5,68	6,54	5,75	5,57	Positive
Deferoxamine mesylate	Negative	5,95	5,10	6,40	6,49	5,39	5,29	5,47	6,89	5,39	6,45	5,78	5,55	Positive
Furosemide	Negative	5,86	5,96	4,69	6,00	5,17	4,83	5,34	5,46	6,14	6,90	5,45	6,11	Positive
(+/-)-Verapamil	Negative	6,43	6,08	6,05	6,84	5,40	6,57	5,61	6,47	5,92	6,47	6,07	6,45	Positive
Levetiracetam	Negative	6,46	5,02	4,27	6,09	5,53	5,31	5,28	5,27	5,54	5,67	6,17	5,79	Positive
Quetiapine	Negative	5,75	6,38	5,55	5,82	5,14	5,61	4,98	5,84	4,71	7,01	5,80	5,94	Positive
Atropine	Negative	6,29	5,63	5,36	6,62	5,21	5,87	5,47	5,67	4,77	6,80	5,74	5,86	Positive
Ursodeoxycholic acid	Negative	6,52	5,44	6,11	6,05	4,76	6,05	5,34	5,94	4,83	5,75	6,12	6,04	Positive
Tiotropium	Negative	6,48	5,51	5,31	6,27	5,14	5,50	5,41	6,05	4,89	6,79	5,89	5,79	Positive
Mifepristone	Negative	7,22	6,51	5,56	6,19	4,76	6,69	5,19	5,93	5,71	6,69	6,48	6,25	Positive
Testosterone	Negative	8,04	5,33	6,05	6,29	4,95	5,80	5,42	5,46	4,60	5,75	6,15	5,66	Positive
Chlorpromazine	Positive	6,02	5,75	6,23	6,01	6,28	7,11	6,20	5,46	5,94	6,87	5,85	5,91	Positive
Cocaine	Positive	6,31	5,87	4,61	6,28	5,08	6,40	5,32	5,68	4,63	6,85	5,59	5,94	Positive
Dexamethasone	Positive	6,19	5,01	5,44	6,15	4,84	5,83	5,33	5,97	4,41	6,13	5,96	6,19	Positive
5,5-Diphenylhydantoin	Positive	5,59	5,12	4,97	6,06	5,69	5,21	4,76	5,57	5,75	6,19	6,16	5,96	Positive
L-Domoic acid	Positive	5,97	5,90	4,71	6,09	5,31	5,32	5,11	5,68	4,98	6,44	5,60	5,82	Positive
Ethanol	Positive	6,51	5,36	4,46	6,18	5,49	5,73	4,95	5,01	4,68	5,58	6,31	5,84	Positive
Haloperidol	Positive	5,94	5,88	6,72	5,95	5,17	6,13	5,45	5,79	5,89	7,08	5,88	6,24	Positive
3,6-Diacetylmorphine	Positive	6,41	5,57	4,00	6,39	5,06	5,99	5,41	5,99	4,68	6,93	5,74	6,29	Positive
Ketamine	Positive	6,69	5,79	5,34	6,23	5,76	5,90	5,28	5,41	5,12	5,86	6,50	5,76	Positive
Methylazoxymethanol (MAM)	Positive	6,35	5,19	4,52	6,22	5,21	5,66	5,06	5,21	4,68	5,77	6,08	5,86	Positive

Figure 29: DNT-list of compounds 1-94. Consensus verdict = classification from the authors. Comment = classification based on IDG threshold. Red highlighted fields = predicted pChEMBL > IDG threshold

Maneb	Positive	6,44	4,95	4,66	6,48	5,72	6,14	5,43	5,53	5,20	6,04	6,53	5,80 Positive
3,4-Methylenedioxymethamphetai	Positive	6,42	5,31	4,60	6,91	5,33	6,21	5,41	5,29	4,82	5,76	6,00	6,10 Positive
Methanol	Positive	6,49	5,32	4,48	6,21	5,48	5,58	5,01	5,01	4,66	5,57	6,33	5,86 Positive
1-Methyl-4-phenyl-1,2,3,6-tetrahy	Positive	6,26	5,21	4,42	5,89	5,90	5,69	5,51	5,12	4,94	5,94	6,28	5,96 Positive
(-)-Nicotine	Positive	6,31	5,23	5,07	5,99	5,84	5,98	5,45	5,27	5,23	5,94	6,12	5,88 Positive
Paraquat	Positive	6,21	5,44	4,56	6,07	5,69	4,64	5,72	5,28	5,22	5,81	5,95	6,33 Positive
3,3',4,4',5,5'-Hexachlorobiphenyl	Positive	6,34	5,33	5,68	6,25	6,24	5,15	5,13	5,57	5,31	6,53	5,80	6,15 Positive
Perfluorooctanoate	Positive	6,81	4,65	6,00	6,50	5,52	4,95	6,16	5,55	4,47	6,24	5,90	5,89 Positive
Perfluorooctanesulfonic acid	Positive	6,50	5,10	5,76	6,38	5,44	5,23	5,58	5,79	4,49	6,35	5,91	5,78 Positive
Terbutaline	Positive	6,29	4,88	4,22	7,23	5,37	5,20	5,09	5,40	5,52	5,85	5,79	6,02 Positive
Retinoic acid	Positive	5,85	4,99	6,07	6,46	5,25	5,48	5,95	5,79	5,17	6,48	5,89	5,61 Positive
Valproic acid	Positive	6,42	4,89	4,79	6,15	5,55	4,76	5,07	4,99	4,41	5,80	6,27	5,76 Positive
Halothane	Positive	6,86	5,66	4,71	6,28	5,72	5,18	5,07	5,07	5,31	5,86	6,29	5,89 Positive
Methadone	Positive	5,77	5,46	5,83	6,42	5,52	5,68	5,71	5,74	5,32	6,59	5,86	5,82 Positive
Primidone	Positive	5,87	4,13	4,61	6,08	5,89	6,02	5,61	5,43	4,39	5,91	6,25	5,97 Positive
Bifenthrin	Positive	6,16	5,53	6,90	6,30	5,16	5,45	5,58	5,96	5,38	6,54	6,01	6,50 Positive
Carbofuran	Positive	6,37	5,24	4,55	6,96	5,35	5,98	5,25	5,33	5,53	5,78	5,82	6,09 Positive
Coumaphos	Positive	5,91	6,16	5,09	7,06	5,17	6,41	5,84	5,91	5,59	7,13	5,78	6,02 Positive
Cyclophosphamide	Positive	6,21	6,20	4,69	5,97	5,59	5,87	4,55	5,38	5,17	6,46	5,87	5,88 Positive
Diazinon	Positive	5,91	5,66	4,96	6,34	5,42	6,36	5,31	5,84	5,02	6,98	5,78	5,97 Positive
Dimethoate	Positive	6,04	5,94	4,71	6,01	5,47	6,34	5,25	5,48	4,91	6,32	6,06	5,70 Positive
Endosulfan	Positive	6,83	6,44	5,43	6,31	5,68	5,81	5,93	5,74	4,89	6,37	6,19	5,74 Positive
Hexachlorobenzene	Positive	6,57	6,41	5,17	6,02	6,02	4,76	4,91	5,48	4,75	6,02	6,23	5,96 Positive
Kepone	Positive	6,90	5,71	5,58	6,68	5,61	5,74	5,29	5,73	3,98	6,16	6,26	5,65 Positive
Methimazole	Positive	6,35	5,39	4,53	5,93	5,71	5,62	5,27	5,12	4,78	5,81	6,37	5,82 Positive
Methyl parathion	Positive	6,17	6,14	4,50	6,11	5,28	6,07	5,24	5,42	5,62	6,50	5,51	6,01 Positive
Parathion	Positive	6,10	5,96	4,76	6,15	5,22	6,41	5,05	5,60	5,92	6,75	5,39	5,99 Positive
Procarbazine hydrochloride	Positive	6,20	4,76	4,96	6,34	5,66	5,44	5,66	5,54	5,03	5,83	5,94	5,76 Negative
Retinol acetate	Positive	5,83	4,86	6,15	6,55	5,09	5,35	5,49	5,79	5,42	6,46	5,85	5,67 Positive
Styrene	Positive	6,20	5,10	4,28	5,95	5,86	4,72	5,41	5,09	4,93	5,71	6,24	5,91 Positive
Thiram	Positive	6,49	5,69	5,00	6,29	5,83	6,43	5,38	5,50	4,88	6,40	7,06	5,93 Positive
Trichlorfon	Positive	6,43	6,24	4,70	6,27	5,55	5,79	5,20	5,49	4,61	6,23	6,11	5,76 Positive
Sodium salicylate	Positive	6,14	4,74	4,63	6,63	5,20	4,48	5,36	5,14	4,62	5,79	5,83	5,91 Negative
Terbufos	Positive	6,21	6,21	5,37	6,35	5,60	5,94	5,51	5,54	6,07	6,33	6,11	5,71 Positive
Dextroamphetamine	Positive	6,40	5,00	4,21	6,00	5,66	4,85	5,35	5,05	4,90	5,67	6,24	5,89 Positive
Acephate	Positive	6,40	6,14	4,55	6,05	5,47	5,77	5,17	5,44	4,75	6,18	5,91	5,75 Positive
Cypermethrin	Positive	6,35	6,50	6,07	6,79	5,12	5,61	5,13	5,91	6,00	6,56	5,83	6,03 Positive
Fenpropathrin	Positive	6,58	6,30	6,20	6,90	5,19	5,56	5,93	5,71	5,61	6,58	5,78	6,08 Positive
EPTC (S-Ethyl dipropylthiocarbam	Positive	6,44	6,19	4,67	6,29	5,56	5,25	4,78	5,05	5,27	5,98	6,45	5,85 Positive
Fenamiphos	Positive	6,10	5,85	5,18	6,41	5,26	5,95	5,16	5,72	5,91	6,73	5,77	5,74 Positive
Glufosinate-ammonium	Positive	6,34	5,65	4,45	6,06	5,36	5,92	5,16	5,42	4,49	5,98	5,77	5,71 Negative
Methamidophos	Positive	6,63	5,48	4,47	6,22	5,55	5,72	5,19	5,23	5,16	5,92	6,16	5,81 Positive
Molinate	Positive	6,56	6,14	4,58	6,15	5,70	5,83	4,85	5,07	5,27	5,97	6,48	5,87 Positive
Naled	Positive	6,22	6,36	4,77	6,16	5,62	5,85	5,44	5,64	4,91	6,39	5,93	5,75 Positive
Tri-allate	Positive	6,55	5,73	5,41	6,35	5,61	5,92	5,18	5,52	5,55	6,29	6,06	5,71 Positive
Clodinafop-propargyl	Positive	5,82	6,69	4,75	6,52	5,33	6,63	5,33	5,73	5,22	6,91	5,54	6,43 Positive
Cymoxanil	Positive	6,64	6,15	4,65	6,12	5,60	5,70	5,28	5,50	5,18	6,03	5,75	5,82 Positive
Fenarimol	Positive	5,83	6,30	5,69	6,19	5,47	6,40	5,62	5,53	4,29	6,90	5,84	5,95 Positive
Imidacloprid	Positive	6,03	6,39	4,95	5,85	5,53	6,21	5,89	5,45	5,85	6,37	5,77	6,00 Positive
Phorate	Positive	6,45	6,69	4,93	6,20	5,45	6,12	5,49	5,38	5,72	6,33	6,18	5,79 Positive
Profenofos	Positive	6,08	6,13	5,77	6,48	5,43	5,57	5,59	5,62	5,60	6,66	5,92	5,80 Positive
Chlorfenapyr	Positive	6,73	7,30	6,49	6,55	5,67	6,11	5,99	5,84	5,76	6,87	5,75	6,56 Positive
Fluazinam	Positive	6,39	6,48	5,33	5,86	5,59	5,05	6,20	5,90	5,00	6,97	5,71	6,20 Positive
Flufenacet	Positive	6,09	4,95	4,93	6,50	5,35	6,67	5,47	5,85	5,77	6,94	5,69	6,16 Positive
Etofenprox	Positive	5,43	5,60	6,95	7,08	5,45	6,41	5,99	6,04	5,57	6,43	5,84	6,10 Positive
Ethoprop	Positive	6,25	6,09	5,32	6,28	5,51	5,26	5,42	5,37	5,98	6,29	6,27	5,78 Positive
Pymetrozine	Positive	6,03	6,05	4,60	5,98	5,38	6,18	5,99	5,31	5,93	6,16	5,83	6,04 Positive
Sulfentrazone	Positive	6,20	6,71	5,07	6,04	5,49	6,06	6,28	5,61	5,34	7,04	5,63	6,28 Positive
5-Bromo-2'-deoxyuridine	Positive	6,29	4,64	4,56	6,16	5,60	6,03	5,25	5,62	5,10	6,32	5,56	5,95 Positive
Raloxifene hydrochloride	Positive	5,73	6,03	8,90	6,86	5,25	5,50	5,79	5,78	6,31	7,44	6,38	6,08 Positive
Acetamiprid	Positive	7,00	6,33	4,83	6,05	5,58	6,13	5,40	5,21	5,17	6,12	6,01	5,90 Positive
Amicarbazone	Positive	6,20	4,82	4,81	6,39	5,82	5,17	5,55	5,52	5,33	6,18	5,84	5,95 Positive
Boscalid	Positive	5,96	6,33	6,23	6,03	5,54	5,73	5,82	5,66	6,20	6,76	5,64	6,17 Positive
Clothianidin	Positive	6,17	5,96	4,77	5,99	5,51	6,18	5,80	5,45	5,46	6,31	5,80	5,77 Positive
Topramezone	Positive	5,70	6,36	4,51	5,79	5,32	6,22	5,93	5,82	5,50	7,17	5,55	6,33 Positive
Spirodiclofen	Positive	6,18	5,48	5,39	5,96	5,19	5,57	5,57	6,13	5,22	6,58	5,87	5,63 Positive
Thiacloprid	Positive	6,74	6,33	4,90	5,94	5,56	6,62	6,69	5,26	4,82	6,49	6,03	6,00 Positive
Thiamethoxam	Positive	5,95	5,70	5,01	6,12	5,56	6,04	5,81	5,56	5,34	6,73	5,64	5,86 Positive
Cyfluthrin	Positive	6,39	6,29	6,09	6,69	5,17	5,65	5,25	5,98	5,39	6,57	5,82	6,42 Positive
Prothioconazole-desthio	Positive	6,07	6,18	5,27	6,46	5,49	6,23	6,20	5,85	5,16	6,85	5,93	5,65 Positive
Pyrasulfotole	Positive	5,96	6,11	5,02	5,61	5,17	6,18	6,21	5,78	5,72	7,28	5,55	6,41 Positive
Lidocaine	Positive	6,37	5,25	5,22	6,12	5,53	5,59	5,44	5,29	5,14	6,20	6,22	5,90 Positive
Chlordiazepoxide	Positive	5,79	6,22	5,10	6,18	5,46	5,53	6,11	5,44	4,93	6,99	5,86	5,89 Positive
Tembotrione	Positive	6,10	5,77	5,15	5,73	5,26	6,03	6,04	5,98	5,36	7,01	5,70	6,18 Positive
N-Methylneodecanamide	Positive	6,58	5,48	4,64	6,38	5,69	5,16	5,27	5,04	5,14	5,82	6,29	5,71 Positive
Technical chlordane	Positive	6,84	6,30	5,63	6,27	6,00	6,08	5,89	5,69	4,77	6,17	6,15	5,81 Positive
2-Ethylhexanoic acid	Negative	6,52	5,00	4,79	6,19	5,57	4,73	5,11	5,02	4,46	5,79	6,26	5,76 Positive
4-Nitroaniline	Positive	6,30	5,77	4,24	5,95	5,35	4,80	5,02	5,15	4,66	5,68	5,88	5,93 Negative
Ethylbenzene	Negative	6,17	5,08	4,41	5,94	5,87	4,90	4,85	5,08	4,91	5,65	6,37	5,91 Positive
Benzonitrile	Positive	7,09	5,70	4,49	6,06	5,74	4,63	4,64	5,10	4,94	5,66	6,04	5,93 Positive
Benzyl alcohol	Positive	6,17	4,93	4,46	5,83	5,54	4,84	5,32	5,08	4,76	5,66	6,32	5,93 Positive
Phenylhydrazine	Positive	6,49	5,03	4,33	6,15	5,64	4,57	5,61	5,17	4,73	5,59	6,07	5,97 Positive
Heptachlor epoxide	Positive	6,94	6,17	5,87	6,53	5,91	5,83	5,88	5,73	4,85	6,13	6,26	5,80 Positive
Dialifor	Positive	5,94	5,72	4,83	5,91	5,36	6,51	5,35	5,63	5,68	6,82	5,94	5,86 Positive
Caprolactam (Azepan-2-one)	Positive	6,46	5,56	4,42	6,09	5,54	4,78	5,21	5,04	4,62	5,60	6,34	5,78 Positive
1,4-Benzenediamine (4-Aminoani	Positive	6,61	5,24	4,34	6,30	5,56	4,91	5,39	5,23	4,83	5,57	6,04	6,01 Positive
1,2-Dibromoethane	Positive	6,51	5,63	4,56	6,17	5,81	5,86	4,39	5,01	4,92	5,78	6,50	5,86 Positive
1-Bromopropane	Positive	6,46	5,37	4,14	6,12	5,77	5,32	4,61	4,99	4,76	5,65	6,32	5,81 Positive
1,3-Butadiene	Positive	6,30	5,09	4,22	6,04	5,75	4,55	4,86	5,11	4,77	5,67	6,24	5,89 Positive
Carbendazim	Negative	6,07	4,95	4,82	5,99	5,48	5,14	5,48	5,17	5,43	6,04	5,93	6,20 Positive
Allyl chloride (1-Chloro-2-propen	Positive	6,40	5,26	4,06	6,05	5,75	5,66	4,42	5,06	4,90	5,66	6,31	5,87 Positive
Ethyl chloride	Positive	6,45	5,23	4,52	6,16	5,33	5,71	4,98	5,09	4,63	5,74	6,31	5,88 Positive
Methyl formate	Positive	6,24	5,35	4,67	6,03	5,44	5,52	4,91	5,22	4,81	5,64	6,36	5,85 Positive
Sarin	Positive	6,64	5,37	4,91	6,15	5,41	5,59	5,05	5,04	4,51	5,74	6,17	5,90 Positive
2-Methylpentane	Negative	6,46	5,61	4,33	6,17	5,77	4,91	4,62	4,99	5,05	5,63	6,34	5,77 Positive
Methyl isobutyl ketone	Positive	6,45	4,68	4,38	6,20	5,49	4,84	4,48	5,04	4,64	5,64	6,27	5,75 Positive
Metaldehyde	Positive	6,26	5,09	4,53	6,26	5,33	5,18	5,07	5,29	4,70	5,77	6,12	5,78 Positive
Cyclohexanol	Positive	6,60	4,77	4,70	6,08	5,51	4,65	5,42	5,09	4,73	5,56	6,34	5,74 Positive
Cyclohexanone	Positive	6,71	5,09	4,24	6,14	5,51	4,86	4,72	5,15	4,72	5,54	6,29	5,71 Positive
Ethylene glycol	Positive	6,36	5,52	4,70	6,09	5,41	5,98	4,90	5,08	4,63	5,84	6,25	5,92 Positive
Cyclohexane	Positive	6,58	5,22	4,49	6,09	5,81	4,79	4,92	5,09	4,85	5,56	6,38	5,71 Positive
Pyridine	Positive	6,38	5,14	4,15	5,93	5,79	4,80	4,80	5,12	4,68	5,63	6,05	5,89 Positive
Triethylene Glycol Monomethyl Et	Positive	6,02	5,37	4,60	6,27	5,35	5,56	4,54	5,31	4,38	5,95	5,99	5,88 Negative
Propoxur	Positive	6,14	5,31	4,48	6,70	5,30	5,95	5,13	5,21	5,23	5,85	5,78	6,22 Positive
Dimefox	Positive	6,45	5,24	5,03	6,25	5,54	5,44	4,79	5,17	4,59	5,99	6,06	5,95 Positive
Triphenyl phosphate	Positive	5,61	5,34	5,92	7,03	5,57	5,16	5,62	5,65	5,22	6,47	5,31	6,09 Positive
Fensulfothion	Positive	6,05	5,31	4,92	6,67	5,20	5,88	5,69	5,73	5,31	6,70	5,59	5,82 Positive
TCEP	Positive	6,06	6,56	4,87	6,10	5,56	5,73	4,52	5,79	4,96	6,31	5,86	5,81 Positive
Spiroxamine	Positive	6,47	6,10	4,73	6,36	5,30	6,22	5,71	5,77	4,79	6,25	5,98	5,47 Positive
Transfluthrin	Negative	6,30	5,42	5,57	6,21	5,45	5,43	6,01	5,84	4,99	6,67	5,84	5,93 Positive

Figure 30: DNT-list of compounds 95-204. Consensus verdict = classification from the authors. Comment = classification based on IDG threshold. Red highlighted fields = predicted pChEMBL > IDG threshold

Isolan	Positive	6,36	5,62	4,72	6,56	5,59	5,42	5,34	5,26	5,25	6,12	5,86	6,07 Positive
Cyclonite (RDX)	Positive	6,08	5,73	4,54	6,02	5,55	5,27	5,12	5,37	5,22	6,15	5,57	5,92 Positive
Fenitrothion	Positive	6,16	6,01	4,56	6,14	5,27	6,28	5,39	5,44	5,95	6,57	5,40	6,00 Positive
Hydroquinone	Positive	6,29	4,57	4,79	6,83	5,32	4,58	5,25	5,19	4,86	5,70	6,10	6,04 Positive
Bisindolylmaleimide	Positive	5,80	5,37	6,23	5,94	5,31	6,99	6,35	5,61	6,35	7,00	6,21	6,20 Positive
Tributyl phosphate (TNBP)	Positive	6,04	5,25	5,14	6,21	5,35	5,33	5,00	5,81	5,43	6,29	5,93	5,66 Positive
Chloroprene	Positive	6,28	5,81	4,13	5,99	5,77	4,30	4,90	5,05	4,99	5,75	6,23	5,87 Positive
Tetrachloroethylene (Perchloroeth	Positive	6,45	5,72	4,20	6,13	5,91	4,36	5,15	5,29	4,72	5,79	6,46	5,86 Positive
Phosphamidon (Dimecron)	Positive	6,08	5,59	4,87	6,15	5,46	6,44	5,08	5,61	5,28	6,51	5,83	5,85 Positive
GenX Free Acid	Negative	6,64	5,51	4,99	6,37	5,49	4,65	5,07	5,36	4,87	6,18	5,95	5,85 Positive
Bis1	Positive	5,85	5,11	6,33	5,74	5,41	6,97	6,49	5,60	6,66	7,21	6,15	6,21 Positive
TDCIPP	Positive	5,95	6,07	5,10	6,24	5,48	5,54	4,92	6,02	5,30	6,53	5,87	5,73 Positive
Ethyl acetate	Positive	6,39	4,99	4,58	6,08	5,42	5,65	5,02	5,03	4,82	5,73	6,14	5,82 Positive
Thiouracil	Positive	6,28	5,71	4,29	6,00	5,51	5,45	5,16	5,12	4,19	5,83	6,15	5,80 Positive
Isoxaflutole	Positive	6,01	6,44	4,84	5,91	5,01	5,69	5,35	5,82	6,07	7,08	5,74	6,08 Positive
Fluoroacetate	Positive	6,30	5,47	4,66	6,10	5,37	5,41	5,19	5,01	4,90	5,77	6,03	5,94 Positive
Y-27632	Positive	6,65	6,27	5,38	6,12	5,46	6,17	5,35	5,64	5,78	6,45	5,92	5,85 Positive
Cytosine arabinoside	Positive	6,33	4,72	4,03	6,16	5,55	5,35	5,06	5,58	4,70	6,05	5,48	5,95 Positive
SAHA	Positive	5,97	5,07	4,77	6,24	5,39	5,68	5,47	5,46	7,30	6,42	5,95	5,94 Positive
Merphos	Positive	6,13	5,52	5,21	6,46	5,65	5,66	5,22	5,83	5,70	6,28	6,04	5,71 Positive
Schradan	Positive	6,05	6,46	5,11	6,31	5,55	5,52	4,67	5,57	4,76	6,52	5,83	5,72 Positive
Imatinib (Glivec)	Positive	5,75	5,79	6,78	5,60	5,43	5,99	6,30	5,64	7,02	7,60	6,11	6,47 Positive
α-Chloralose	Positive	6,37	4,26	5,00	6,63	5,36	5,41	5,14	5,52	4,38	6,08	5,85	5,80 Positive
Nelfinavir mesylate	Negative	6,17	5,48	7,64	6,45	5,15	6,24	5,85	7,49	7,22	6,89	6,30	6,19 Positive
Methomyl	Positive	6,10	5,69	4,72	6,27	5,44	5,81	5,48	5,29	5,72	6,02	6,13	5,90 Positive
PD98059	Positive	6,13	5,24	5,31	6,79	5,12	5,44	5,11	6,12	4,90	6,53	5,48	6,45 Positive
Edifenphos	Positive	5,46	5,87	5,83	6,02	5,59	5,93	5,23	5,31	5,81	6,61	5,98	5,81 Positive
3-(Dimethylamino)-propanenitrile	Positive	7,38	6,00	4,59	6,26	5,61	5,63	4,52	5,16	4,56	5,81	6,10	5,81 Positive
2,3,7,8-Tetrachlorodibenzo-p-dio	Positive	6,55	5,88	5,92	6,71	5,97	5,25	5,31	5,45	4,13	6,51	5,76	6,10 Positive
Gefitinib (Iressa)	Positive	5,96	6,10	4,94	6,02	5,15	6,69	5,37	6,07	7,00	7,27	5,58	6,56 Positive
Isoxathion	Positive	5,79	5,84	4,83	6,47	5,40	6,73	6,08	5,68	5,63	7,04	5,62	6,20 Positive
Wortmannin	Positive	6,39	4,97	4,25	6,05	5,04	5,39	5,58	6,15	4,27	6,71	5,96	5,99 Positive
DAPT	Positive	5,70	4,80	5,09	6,23	5,36	6,10	5,67	6,03	6,20	6,44	5,76	6,30 Positive
Bromophos (Brofene)	Positive	6,27	5,86	5,22	6,43	5,75	5,80	5,50	5,67	5,72	6,68	5,86	5,95 Positive
Leptophos	Positive	5,74	6,37	5,60	6,73	5,62	6,16	5,57	5,79	5,61	6,84	5,77	6,22 Positive
U0126	Positive	6,38	6,53	4,90	6,34	5,40	5,96	5,55	5,87	6,23	6,81	5,85	6,35 Positive
Tetrachlorvinphos	Positive	6,00	6,32	4,97	5,99	5,60	6,16	5,70	5,72	4,93	6,84	5,77	5,97 Positive
Heptenophos	Positive	6,65	6,81	4,82	6,06	5,36	5,82	5,30	5,49	4,67	6,29	5,91	5,91 Positive
Mirex	Positive	6,77	4,70	5,74	6,64	5,84	5,31	5,69	5,81	4,15	6,22	6,29	5,71 Positive
Chlormephos	Positive	6,25	6,18	4,93	6,18	5,63	5,85	5,86	5,44	5,83	6,24	6,34	5,83 Positive
Formothion	Positive	6,07	6,22	4,67	5,92	5,47	5,66	5,17	5,52	4,72	6,45	5,82	5,74 Positive
Hexabromocyclododecane (HBCD	Positive	6,46	4,88	5,25	6,47	5,80	5,06	5,20	5,83	5,42	6,21	6,27	5,69 Positive
Methylchloroisothiazolinone	Positive	6,46	6,01	4,39	5,92	5,61	5,16	5,34	5,09	5,42	5,78	6,39	5,87 Positive
PFPeS	Positive	6,59	5,11	4,69	6,29	5,46	4,71	5,21	5,09	4,49	5,67	6,30	5,68 Positive
Endothion	Positive	5,97	5,83	4,63	6,64	5,26	5,95	5,25	5,75	5,76	6,58	5,56	5,97 Positive
SB-216763	Positive	5,96	5,63	4,91	5,88	5,84	5,79	6,22	5,75	5,90	7,25	5,93	6,00 Positive
PFESA1	Negative	6,48	5,00	4,75	6,41	5,46	4,97	5,07	5,80	5,14	6,29	5,84	5,89 Positive
MCPA.EHE (2-ethyl hexyl ester)	Negative	5,96	5,50	5,64	6,73	5,23	5,51	5,47	6,13	5,85	6,58	5,84	5,55 Positive
Isobenzan	Positive	6,84	6,20	5,75	6,58	5,92	5,89	5,94	5,65	4,49	6,17	6,22	5,71 Positive
Carbamazepine	Positive	5,92	5,22	4,69	5,94	5,64	4,29	5,84	5,34	5,49	6,09	6,35	6,13 Positive
Ethiofencarb (Croneton)	Positive	6,62	5,85	4,88	6,50	5,56	6,17	5,67	5,16	5,84	6,07	6,14	5,99 Positive
desvenlafaxine hydrochloride	Positive	6,64	5,54	5,76	7,08	5,41	7,04	5,30	5,55	5,17	6,42	6,00	5,73 Positive
Oxydemeton-methyl	Positive	6,22	5,89	4,75	6,40	5,37	5,85	5,05	5,66	5,08	6,39	5,82	5,79 Positive
PFHxA	Positive	6,90	4,89	5,10	6,42	5,46	4,73	6,15	5,30	4,42	6,11	5,98	5,89 Positive
Aldrin	Positive	6,84	5,86	5,70	6,29	5,83	6,05	5,40	5,56	4,73	6,17	6,20	5,74 Positive
Mexacarbate (Zectran)	Positive	6,44	5,62	4,70	6,48	5,39	6,85	5,25	5,40	5,32	6,21	5,83	6,12 Positive
5-Azacytidine	Positive	6,30	4,93	4,14	6,08	5,61	4,98	5,19	5,61	5,13	6,12	5,52	5,94 Positive
Trichloronate	Positive	6,22	6,16	5,34	6,57	5,79	5,95	5,49	5,57	5,43	6,44	6,09	5,87 Positive
Aminonicotinamide	Positive	6,38	5,06	4,49	6,16	5,57	4,84	5,60	5,29	5,44	5,66	6,00	5,90 Negative
Amitraz	Positive	5,94	5,74	6,25	6,11	5,48	6,22	5,90	5,60	5,81	6,67	5,91	5,93 Positive
Sulprofos	Positive	6,12	6,40	5,75	6,55	5,52	6,02	5,80	5,54	5,82	6,61	5,96	5,72 Positive
rivaroxaban	Negative	5,94	5,80	4,64	5,70	5,01	6,13	5,69	5,74	6,19	7,02	5,45	6,10 Positive
Mipafox	Positive	6,55	5,56	4,91	6,42	5,52	5,71	5,12	5,15	5,31	5,85	6,14	5,82 Positive
Perfluorobutanesulfonic acid (PFE	Positive	6,69	5,07	4,94	6,35	5,34	4,90	5,17	5,32	4,55	6,12	5,99	5,85 Positive
PFHpS	Positive	6,62	4,76	5,15	6,43	5,45	5,08	5,41	5,71	4,44	6,32	5,94	5,79 Positive
Perfluorohexanesulfonic acid pot	Positive	6,66	4,61	5,01	6,33	5,40	5,06	5,30	5,60	4,34	6,27	5,95	5,79 Positive
Diethylene glycol diacrylate	Positive	5,93	5,12	4,60	6,10	5,21	4,90	4,81	5,55	4,54	6,25	5,76	6,03 Positive
Triadimefon	Positive	5,87	6,28	4,73	6,88	5,44	5,88	6,19	5,45	5,49	6,76	5,88	5,96 Positive
Diazepam	Positive	5,84	5,55	5,27	5,80	5,86	5,45	5,57	5,37	5,00	6,71	6,38	5,91 Positive
Cyclopamine	Positive	5,75	5,83	5,90	6,24	5,92	7,04	5,37	5,54	5,63	6,51	5,89	5,89 Positive
Pyroxasulfone (KIH-485)	Positive	6,23	6,27	4,72	6,50	5,26	4,83	6,40	6,12	5,45	6,82	5,80	6,00 Positive
Chlorfenvinphos	Positive	5,89	6,12	5,30	5,96	5,40	5,94	5,81	5,78	5,46	6,71	5,79	5,78 Positive
Chlorothion	Positive	6,12	5,99	4,57	6,23	5,39	5,81	5,41	5,50	5,36	6,70	5,49	5,97 Positive
Vinclozolin	Negative	6,21	5,87	4,70	5,97	5,64	4,97	6,16	5,53	5,03	6,54	5,94	5,85 Positive
Trimethyl phosphate	Positive	6,35	5,70	4,57	6,16	5,42	5,38	4,83	5,34	4,47	6,02	5,83	5,83 Positive
Fenvalerate	Positive	6,08	6,68	6,62	6,90	5,28	6,06	5,29	5,72	5,92	6,39	5,85	6,30 Positive
Rapamycin	Positive	6,10	5,43	7,10	6,49	5,22	5,62	5,27	7,14	5,38	6,34	6,31	6,27 Positive
4,6-Dinitro-o-cresol	Positive	6,37	5,36	4,38	6,20	5,22	4,69	5,06	5,33	4,98	5,85	5,50	6,01 Positive
Pyriminil (Pyrinuron, Vacor)	Positive	5,73	5,76	4,64	5,69	5,50	5,95	5,61	5,48	5,49	6,59	5,52	6,17 Positive
Ethylbis(2-chloroethyl)amine	Positive	6,46	5,68	4,82	6,19	5,78	6,31	4,70	5,11	5,20	6,03	6,30	5,92 Positive
1,3-Dichloropropene	Positive	6,44	5,66	4,36	6,02	5,81	5,80	4,50	5,05	5,20	5,70	6,45	5,86 Positive
PBDE-47	Positive	5,82	5,75	6,10	6,76	5,98	5,72	5,44	5,62	5,02	6,48	5,77	5,99 Positive
Fenthion	Positive	6,22	6,08	4,98	6,49	5,54	6,14	5,22	5,53	5,97	6,64	5,66	5,93 Positive
Tris(2-chloroethyl)amine (Trichlor	Positive	6,46	6,11	4,66	6,13	5,74	5,67	4,62	5,25	5,04	6,12	6,32	5,95 Positive
Chlorpyrifos oxon	Positive	5,93	5,84	5,02	6,32	5,63	6,12	5,10	5,72	5,30	6,79	5,74	5,94 Positive
Ethion	Positive	5,95	5,70	5,17	6,30	5,29	6,24	5,55	6,15	5,92	6,52	6,02	5,79 Positive
1-2-Propyleneglycol	Positive	6,56	5,12	4,71	6,17	5,36	5,96	5,17	5,00	4,85	5,77	6,25	5,83 Positive
norfluoxetine hydrochloride	Positive	6,08	4,86	6,56	6,87	5,69	7,55	6,17	5,42	5,61	6,47	5,79	6,13 Positive
Caffeine	Positive	6,23	6,15	4,64	5,94	5,33	5,09	5,72	5,52	4,96	6,22	5,84	5,99 Positive
Methotrexate	Positive	6,05	4,98	4,65	6,09	5,49	5,48	5,22	5,91	6,08	6,92	5,70	6,19 Positive
Methyl-n-butyl ketone (2-Hexanor	Positive	6,44	4,88	4,36	6,28	5,47	4,97	4,45	5,05	4,67	5,63	6,30	5,75 Positive
Cyclosporine	Positive	5,90	4,54	6,53	7,09	5,39	5,89	5,57	7,21	6,38	6,31	6,30	5,98 Positive
Aniline	Positive	6,39	5,27	4,21	6,01	5,70	4,61	5,56	5,07	4,68	5,63	6,11	5,97 Positive
Dichlorvos (DDVP, Vapona)	Positive	6,33	6,23	4,52	6,19	5,43	5,49	4,96	5,42	4,92	6,12	6,07	5,86 Positive
Fluoroacetamide	Positive	6,48	5,13	4,56	6,11	5,27	5,81	5,27	5,06	5,02	5,66	6,30	5,90 Positive
carbamazepine 10,11-epoxide	Negative	6,27	5,89	4,90	6,23	5,62	5,18	6,18	5,39	5,22	6,30	6,19	6,04 Positive
Avermectin	Positive	6,14	4,67	6,44	6,44	5,12	5,21	5,31	7,21	5,00	6,17	6,35	6,56 Positive
Forskolin	Positive	6,28	4,93	4,77	5,96	4,93	5,44	5,13	5,96	4,67	6,04	6,00	6,12 Positive
Acetone	Positive	6,44	4,59	4,40	6,17	5,46	5,24	4,65	5,02	4,58	5,65	6,32	5,75 Positive
Chloroform	Positive	6,48	5,54	4,65	6,13	5,82	5,00	4,74	5,08	5,00	5,76	6,57	5,76 Positive
Cyhalothrin	Positive	6,39	6,77	6,55	6,48	5,14	5,59	5,49	6,14	5,43	6,52	5,92	6,47 Positive
Monocrotophos	Positive	6,15	5,72	4,75	6,00	5,39	6,04	5,41	5,44	5,06	6,22	5,75	5,92 Positive
Flumethrin	Negative	6,13	6,07	6,79	6,44	5,41	5,65	5,39	5,88	5,32	6,43	6,36	6,36 Positive
1-Pentanol	Positive	6,43	5,39	4,61	6,29	5,47	5,11	5,02	5,00	4,76	5,65	6,31	5,84 Positive
Benzene	Positive	6,25	5,27	4,32	5,96	5,87	4,43	4,94	5,10	4,86	5,67	6,16	5,92 Positive
dichlorodiphenyldichloroethylene	Positive	5,79	5,88	5,46	6,10	6,17	5,15	5,69	5,37	5,06	6,50	5,76	5,90 Positive
Propaphos	Positive	5,84	5,59	5,39	6,44	5,27	5,60	4,86	5,87	5,95	6,76	5,75	5,77 Positive
Mevastatin	Positive	6,05	4,98	5,55	6,18	4,79	6,06	5,70	6,31	5,29	6,24	5,90	7,37 Positive
Methyl bromide	Positive	6,50	5,00	4,46	6,14	5,79	5,37	4,67	5,01	4,65	5,59	6,35	5,81 Positive

Figure 31: DNT-list of compounds 205-313. Consensus verdict = classification from the authors. Comment = classification based on IDG threshold. Red highlighted fields = predicted pChEMBL > IDG threshold

Ethylene	Positive	6,39	5,18	4,22	6,09	5,79	4,73	4,78	5,09	4,91	5,56	6,19	5,87 Positive
Methyl chloride (Chloromethane)	Positive	6,51	5,26	4,23	6,19	5,79	5,87	4,63	5,05	4,78	5,58	6,30	5,82 Positive
Methyl iodide	Positive	6,49	5,08	4,46	6,13	5,81	4,98	4,44	5,01	4,69	5,60	6,44	5,81 Positive
Bensulide	Positive	5,82	5,26	5,14	6,03	5,33	6,21	5,21	5,87	5,63	6,85	5,77	5,73 Positive
Vinyl chloride (Chloroethene)	Positive	6,42	5,18	4,08	6,03	5,76	4,36	4,74	5,05	4,81	5,65	6,21	5,85 Positive
Dichloromethane	Positive	6,53	5,14	4,49	6,04	5,83	5,29	4,39	5,07	4,86	5,65	6,42	5,81 Positive
Ethylene oxide	Positive	6,43	5,20	4,73	6,08	5,68	5,51	4,52	5,11	4,79	5,70	6,29	5,89 Positive
1,2-Propylene oxide	Positive	6,50	5,11	4,76	6,08	5,67	5,58	4,67	5,07	4,88	5,61	6,26	5,85 Positive
Acetone cyanohydrin (2-Hydroxy-2	Positive	7,33	4,87	4,66	6,25	5,45	5,03	5,10	5,05	4,95	5,67	6,19	5,80 Positive
Dimethyl sulfate	Positive	6,43	5,23	4,59	6,16	5,38	4,89	4,82	5,32	4,62	6,00	5,93	5,87 Positive
Lactofen	Positive	6,19	5,95	5,27	6,34	5,57	6,33	5,72	6,08	5,49	6,67	5,68	6,77 Positive
Mevinphos	Positive	6,13	6,15	4,49	5,92	5,28	5,35	4,88	5,49	4,96	6,18	5,81	5,86 Positive
Tri-o-cresylphosphate	Positive	5,65	5,16	5,78	6,99	5,29	5,30	5,56	5,66	5,56	6,51	5,67	6,08 Positive
Dioxathion	Positive	5,93	5,14	4,56	6,40	5,25	6,10	5,35	6,34	5,56	6,53	6,05	5,88 Positive
Tribufos	Positive	6,19	5,34	5,20	6,57	5,58	5,35	5,52	5,79	5,51	6,32	5,97	5,73 Positive
Isophorone	Positive	6,49	4,78	4,17	6,16	5,52	5,17	4,94	5,13	5,01	5,59	6,29	5,80 Positive
2-Methylpropanenitrile	Positive	7,27	5,22	4,50	6,20	5,68	4,86	4,25	5,03	5,09	5,61	6,20	5,78 Positive
Methyl ethyl ketone	Positive	6,42	4,93	4,26	6,19	5,49	5,14	4,44	5,03	4,86	5,65	6,29	5,80 Positive
Carbophenothion (Trithion)	Positive	6,01	6,64	5,43	6,13	5,51	6,07	5,52	5,61	6,09	6,65	5,99	5,79 Positive
1,1,2,2-Tetrachloroethane	Positive	6,46	5,32	4,49	6,25	5,82	5,27	5,05	5,29	4,98	5,74	6,48	5,85 Positive
Dichloroacetic acid	Positive	6,47	5,02	4,58	6,09	5,46	4,92	5,47	5,04	4,82	5,81	6,24	5,76 Positive
2-Nitropropane	Positive	6,68	5,17	4,36	6,09	5,38	5,08	5,08	5,04	4,86	5,64	6,18	5,77 Positive
TBBPA	Positive	5,83	4,91	6,29	6,83	5,73	5,63	6,01	5,74	5,31	6,69	6,01	6,16 Positive
Tefluthrin	Positive	6,44	5,59	5,51	6,24	5,28	5,48	5,95	6,08	4,92	6,57	5,89	5,95 Positive
Simvastatin	Positive	6,14	4,93	6,44	6,07	4,82	6,50	5,51	6,32	5,32	6,07	6,07	7,71 Positive
Hexaconazole	Positive	5,85	6,15	5,15	6,52	5,59	5,87	5,55	5,70	4,99	6,83	5,89	5,78 Positive
BPS	Positive	5,76	5,64	5,80	6,45	5,09	5,25	5,24	5,37	5,47	6,31	5,70	6,23 Positive
Tetramethylenedisulfotetramine (T	Positive	6,46	4,92	4,89	6,23	5,69	5,49	5,07	5,63	5,07	6,23	5,97	5,87 Positive
Methyl methacrylate	Positive	6,31	5,24	4,50	5,98	5,40	5,18	4,88	5,06	4,84	5,74	6,13	5,90 Positive
Toxaphene	Positive	6,74	5,42	5,44	6,53	5,90	6,01	5,75	5,64	4,57	6,31	6,20	5,65 Positive
Systox (Demeton)	Positive	6,47	6,50	4,99	6,19	5,39	5,60	5,44	5,58	5,78	6,34	6,08	5,78 Positive
Di-N-butyl phthalate	Positive	5,74	5,00	4,63	6,04	5,26	4,98	4,97	5,79	5,46	6,62	5,91	5,94 Positive
Benzyl butyl phthalate	Positive	5,37	5,16	5,10	6,07	5,34	5,09	5,08	5,72	5,40	6,69	5,77	5,99 Positive
Pentachlorophenol	Positive	6,59	6,17	5,29	6,51	5,74	4,58	5,58	5,43	4,43	5,94	6,33	5,94 Positive
Dinoseb	Positive	6,53	5,56	4,47	6,35	5,39	4,61	5,10	5,37	5,17	5,97	5,58	5,97 Negative
1,2-Benzenedicarbonitrile (1,2-Dic	Positive	7,19	5,10	4,52	6,26	5,56	4,56	4,52	5,24	4,97	5,63	6,02	6,08 Positive
Demeton-S-methyl	Positive	6,52	6,33	4,84	6,07	5,50	6,16	5,54	5,36	5,45	6,41	6,15	5,87 Positive
2,4,5-Trichlorophenoxyacetic acid	Positive	6,30	5,78	4,56	6,26	5,57	4,81	5,79	5,20	4,34	6,11	6,18	5,94 Positive
venlafaxine hydrochloride	Positive	6,42	5,53	5,83	7,05	5,45	7,09	5,06	5,57	5,28	6,46	5,91	5,72 Positive
Fonofos	Positive	6,29	5,73	4,93	6,26	5,93	6,08	4,99	5,30	5,53	6,42	6,30	5,81 Positive
Cyolane (Phospholan)	Positive	6,19	6,81	4,92	6,09	5,45	5,52	5,60	5,37	5,04	6,40	5,97	5,74 Positive
Methidathion (Suprathion)	Positive	5,89	7,01	4,86	6,37	5,57	6,01	5,33	5,66	5,64	6,70	5,86	5,77 Positive
ADONA	Negative	6,47	6,00	4,94	6,43	5,44	5,00	5,81	5,74	4,62	6,22	5,87	5,85 Positive
1,2-Dibromo-3-chloropropane (D	Positive	6,63	6,33	4,69	6,26	5,86	5,87	4,83	5,25	4,99	5,87	6,47	5,84 Positive
3-Methylpentane	Negative	6,49	5,59	4,29	6,16	5,75	4,79	4,69	4,99	4,90	5,63	6,34	5,77 Positive
Methylcyclopentane	Positive	6,56	5,65	4,60	6,06	5,79	4,82	4,87	5,11	4,85	5,53	6,35	5,74 Positive
Soman	Positive	6,76	5,30	5,01	6,18	5,56	5,07	5,03	5,07	4,88	5,79	6,19	5,89 Positive
Tebupirimfos	Positive	6,03	5,04	4,89	6,47	5,43	5,86	5,13	5,84	4,95	6,77	5,79	5,69 Positive
Pyridaben	Positive	5,79	6,00	6,12	6,05	5,58	6,37	6,46	5,93	6,08	6,74	5,95	5,86 Positive
Dichlofenthion	Positive	6,13	6,37	5,45	6,40	5,45	6,13	5,13	5,63	5,59	6,53	5,73	5,89 Positive
Cumene	Positive	6,24	5,39	4,30	5,99	5,85	4,71	4,96	5,04	5,12	5,69	6,36	5,90 Positive
Nitrobenzene	Positive	6,40	6,10	4,20	5,97	5,45	4,70	4,98	5,08	4,74	5,73	6,01	5,91 Positive
N-Acetyl-L-aspartic acid	Negative	6,14	5,36	4,40	6,21	5,35	5,14	5,44	5,38	5,09	5,91	5,84	5,77 Negative
1,1'-(1,1,2,2-tetramethylethylene	Negative	5,88	5,36	5,56	6,34	6,28	5,10	5,41	5,45	5,40	6,22	6,13	5,88 Positive
1,1',1",1'"-ethylenedinitrilotetrapr	Positive	5,94	5,59	4,57	6,44	5,30	5,52	4,99	5,77	5,30	6,25	5,72	5,71 Positive
1,1,3,3-tetramethylbutyl peroxyn	Positive	6,21	5,45	5,33	6,52	5,23	5,85	5,50	5,99	5,77	6,27	5,80	5,44 Positive
1,1'-Ethane-1,2-diylbis[pentabron	Positive	5,99	6,04	5,75	6,07	5,86	5,47	6,13	5,68	5,39	6,68	6,28	6,03 Positive
1,3-diethyldiphenylurea (N,N'-Die	Negative	5,73	6,01	5,68	6,00	5,75	6,81	5,34	5,42	5,90	6,64	5,89	6,15 Positive
1,6-hexamethylene diisocyanate	Negative	6,06	6,16	4,89	6,18	5,28	4,66	5,12	5,34	5,29	6,01	6,14	5,74 Positive
2-(4-tert-butylbenzyl) propionalde	Negative	6,40	4,88	4,27	6,11	5,76	5,35	4,93	5,31	5,23	5,91	6,21	5,84 Positive
2,2'-dimethyl-4,4'-methylenebis(c	Positive	7,12	6,26	5,70	6,58	5,47	6,45	5,35	5,63	5,53	5,71	6,04	5,70 Positive
2,2'-Iminodiethanol	Positive	6,27	5,96	4,52	6,24	5,46	6,21	5,01	5,15	4,60	5,78	6,17	5,81 Positive
2,4,6-trinitrotoluene	Positive	6,36	5,64	4,77	6,01	5,39	4,98	4,47	5,49	5,09	5,96	5,48	6,01 Positive
2-dimethylaminoethanol	Negative	6,51	5,22	4,75	6,10	5,44	5,84	4,95	5,06	4,71	5,84	6,22	5,88 Positive
2-Pyrrolidone	Positive	6,45	5,58	4,59	6,00	5,55	5,83	5,23	5,03	4,69	5,62	6,29	5,78 Positive
3,3',4,4'-Tetrachloroazobenzene	Positive	6,20	5,65	5,38	6,11	6,03	5,09	6,04	5,45	5,30	6,48	5,89	5,95 Positive
3,5-Dimethylpyrazole	Negative	6,31	5,34	4,64	5,99	5,83	5,55	4,91	5,10	4,96	5,73	6,22	5,88 Positive
3-aminopropyldiethylamine	Negative	6,43	5,44	4,86	6,29	5,60	6,05	5,12	5,10	4,54	5,86	6,21	5,83 Positive
Acetophenone	Positive	6,12	5,69	4,22	6,01	5,54	4,65	5,05	5,05	4,66	5,68	6,32	5,92 Positive
Acibenzolar-S-methyl	Positive	6,24	5,42	4,70	5,94	5,62	5,33	6,09	5,22	5,10	6,02	6,31	5,89 Positive
Alitame (sweetner)	Negative	6,05	5,42	4,73	6,35	5,36	5,79	5,48	5,89	6,00	6,38	5,85	5,49 Positive
Aluminum citrate	Positive	6,08	5,44	4,53	6,24	5,30	5,17	4,79	5,33	4,36	5,92	5,79	5,82 Negative
Atorvastatin	Positive	5,63	5,48	6,35	6,09	5,56	5,60	5,68	5,86	5,05	6,50	6,33	8,22 Positive
Azinphos-methyl	Negative	5,86	5,81	5,08	6,29	5,42	6,46	5,80	5,61	4,54	6,99	5,78	5,87 Positive
Biphenyl	Negative	5,77	4,94	4,35	6,15	6,02	4,20	5,56	5,13	5,11	5,72	6,09	6,11 Positive
Biphenyl-4,4'-diol	Negative	5,90	4,78	5,74	6,96	5,39	4,62	5,22	5,17	5,08	5,88	5,93	6,30 Positive
Bis(2,2,6,6-tetramethyl-4-piperidy	Negative	5,99	5,03	6,27	6,26	5,26	6,19	5,59	7,47	6,28	6,23	6,15	5,94 Positive
Calcium cyclamate	Negative	6,53	4,73	4,28	6,18	5,58	5,02	5,05	5,24	4,89	5,85	6,04	5,68 Positive
Carbon disulfide	Positive	6,39	5,11	4,22	6,04	5,76	4,43	4,61	5,09	5,04	5,75	6,32	5,78 Positive
Chlorpyrifos-methyl	Negative	6,23	6,12	4,89	6,36	5,70	5,59	5,65	5,55	5,11	6,71	5,86	5,98 Positive
CI-943	Positive	6,59	5,50	4,76	6,35	5,52	5,71	5,94	5,57	5,28	6,21	6,10	5,95 Positive
Decabromodiphenylether	Positive	6,04	5,08	5,63	6,44	5,81	5,44	5,75	5,70	5,20	6,37	6,27	6,10 Positive
Demiditraz	Positive	6,33	5,83	4,72	6,11	5,81	6,18	5,99	5,28	5,07	5,86	6,20	5,98 Positive
Dicrotophos	Positive	6,18	6,04	4,89	5,88	5,41	6,02	4,87	5,43	4,47	6,27	5,69	5,97 Positive
Ethylene dibenzoate	Positive	5,41	5,33	5,18	6,21	5,43	4,36	5,20	5,43	4,67	6,47	5,94	6,11 Positive
Ethylene dichloride	Negative	6,53	5,97	4,56	6,19	5,81	5,89	4,33	5,06	4,92	5,80	6,41	5,84 Positive
Ethylenethiourea	Positive	6,40	5,20	4,44	6,24	5,81	6,13	5,48	5,25	4,32	5,85	6,36	5,82 Positive
FD&C Red 3	Negative	6,20	5,70	6,79	6,75	5,16	5,25	5,27	5,69	4,39	6,40	6,32	6,11 Positive
FD&C Red 40	Positive	5,77	5,35	5,77	6,54	5,43	5,69	5,52	5,68	6,43	6,70	5,63	6,27 Positive
Fenamidone	Negative	5,65	6,75	5,29	5,94	5,24	5,56	5,67	5,45	5,81	7,00	5,93	5,87 Positive
Fenpropidin	Positive	6,26	5,48	5,96	6,37	5,92	5,79	6,31	5,70	5,31	6,58	5,97	5,62 Positive
Flubendiamide	Positive	6,28	5,33	6,05	5,92	5,54	6,56	5,23	5,71	5,47	7,09	6,14	6,71 Positive
Flupyradifurone	Positive	5,99	5,93	4,89	6,43	5,41	6,31	5,35	5,47	5,54	6,74	5,85	6,04 Positive
Geraniol	Positive	6,32	5,60	4,97	6,18	5,32	5,22	5,15	5,00	5,02	5,78	6,31	5,89 Positive
Hexahydro-1,3,5-trimethyl-1,3,5-t	Positive	6,32	5,33	4,71	6,02	5,84	5,40	5,24	5,20	4,29	5,94	6,07	5,96 Positive
Hydroxycyclohexyl phenyl ketone	Negative	6,36	4,77	4,67	6,21	5,75	5,02	5,74	5,31	4,68	5,87	6,43	5,80 Positive
Indaziflam	Positive	6,42	6,04	5,50	6,21	5,51	5,96	6,01	5,77	5,75	6,86	5,78	6,04 Positive
Indoxacarb	Negative	6,15	5,77	5,26	6,27	5,28	5,94	6,21	6,11	5,76	6,94	6,08	6,24 Positive
Lasofoxifene	Negative	5,78	5,72	8,64	6,75	4,97	6,37	5,05	5,64	5,08	6,52	6,25	5,79 Positive
L-Valyl-L-prolyl-L-proline and L-Is	Negative	5,89	4,73	4,96	6,71	5,31	6,28	5,89	5,97	6,03	6,61	5,66	5,73 Positive
Malathion	Positive	5,92	5,51	4,83	6,12	5,27	5,15	5,20	5,71	5,33	6,50	5,85	5,76 Positive
Mancozeb	Negative	6,44	4,95	4,66	6,48	5,72	6,14	5,43	5,53	5,20	6,04	6,53	5,80 Positive
Melamine	Negative	6,33	5,56	4,35	6,26	5,78	5,18	5,38	5,47	4,82	5,97	5,79	5,84 Negative
Mequinol (4-Methoxyphenol)	Positive	6,18	4,47	4,78	6,72	5,39	5,28	4,94	5,16	4,93	5,68	6,25	6,13 Positive
Methacrylamide	Positive	6,53	5,04	4,16	6,03	5,38	5,02	5,39	5,04	5,12	5,62	6,29	5,83 Positive
Methyl 2-naphthyl ether	Positive	6,24	4,88	4,93	6,66	5,64	5,36	4,93	5,09	5,04	5,68	6,24	6,00 Positive
Methyl 4-hydroxybenzoate [methy	Positive	6,25	4,75	4,63	6,37	5,23	4,87	5,15	5,06	5,03	5,82	5,94	5,97 Negative
N,N-Dicyclohexyl-2-benzothiazole	Positive	6,29	5,38	5,76	6,16	5,79	5,33	6,30	5,59	6,35	6,69	5,98	5,47 Positive
N-1,3-dimethylbutyl-N'-phenyl-p-	Positive	6,01	5,45	6,59	6,43	5,75	6,23	5,30	5,52	6,28	6,50	5,85	5,96 Positive
Nitisinone	Positive	6,39	5,74	4,72	5,87	5,20	6,00	5,81	5,62	5,71	6,60	5,39	5,89 Positive

Figure 32: DNT-list of compounds 313-423. Consensus verdict = classification from the authors. Comment = classification based on IDG threshold. Red highlighted fields = predicted pChEMBL > IDG threshold

N-Methyl-2-pyrrolidone	Positive	6,59	5,59	4,60	5,99	5,51	5,77	5,29	5,11	4,90	5,81	6,30	5,79 Positive
Octocrilene	Negative	6,21	5,43	6,20	6,74	5,28	5,40	5,89	5,95	5,27	6,52	5,88	5,82 Positive
Orthosulfamuron	Positive	6,10	5,73	5,00	5,97	5,52	5,73	5,89	5,87	5,01	7,00	5,39	6,17 Positive
Penthiopyrad (MTF-753)	Positive	6,22	5,88	5,39	6,16	5,24	6,15	5,98	5,96	5,48	6,80	5,80	5,94 Positive
p-Menthane-3,8-diol	Negative	6,94	5,11	4,72	6,32	5,47	5,55	5,58	5,25	4,79	5,75	6,20	5,67 Positive
Propineb	Negative	6,56	4,78	4,71	6,53	5,81	6,28	5,47	5,42	5,36	5,93	6,53	5,74 Positive
Propylparaben (Propyl 4-hydroxyl	Negative	6,15	5,27	4,87	6,44	5,32	5,07	5,30	5,09	5,20	5,89	5,94	5,94 Negative
Resorcinol	Positive	6,31	4,71	4,59	6,84	5,32	4,65	5,36	5,19	4,74	5,70	6,10	6,04 Positive
Silver acetate	Positive	6,46	5,58	4,51	6,06	5,41	5,17	5,11	5,15	4,81	5,68	6,33	5,82 Positive
Sodium citrate dihydrate	Positive	6,08	5,44	4,53	6,24	5,30	5,17	4,79	5,33	4,36	5,92	5,79	5,82 Negative
Sulfloxaflor	Positive	6,91	6,11	4,99	6,09	5,50	5,51	5,25	5,48	5,81	6,46	5,74	5,98 Positive
Tartrazine	Positive	5,59	6,34	5,24	5,65	5,37	5,59	4,85	5,74	6,19	6,89	5,66	6,05 Positive
Trifludimoxazin	Negative	6,23	7,53	4,60	6,16	5,57	5,70	5,61	5,84	6,22	7,17	5,40	6,22 Positive
Tritosulfuron	Negative	6,32	5,16	5,58	6,09	5,62	5,73	6,01	5,80	4,96	7,01	5,52	6,13 Positive
Vinyl neononanoate	Positive	6,25	4,72	4,39	6,12	5,56	4,75	5,10	4,98	5,05	5,90	6,21	5,73 Positive
Zinc bis(dibutyldithiocarbamate)	Positive	6,41	5,66	4,70	6,29	5,75	5,39	4,83	5,13	5,36	6,03	6,42	5,71 Positive
Ziram	Positive	6,65	5,32	4,53	6,20	5,73	5,63	5,06	5,11	5,02	5,93	6,47	5,76 Positive

Figure 33: DNT-list of compounds 424-440. Consensus verdict = classification from the authors. Comment = classification based on IDG threshold. Red highlighted fields = predicted pChEMBL > IDG threshold

Precision describes the proportion of substances predicted as toxic that are indeed toxic, while sensitivity describes the proportion of truly toxic substances that are correctly identified as toxic (Rainio et al., 2024). Initially, the numbers of true positives (TP), false positives (FP), and false negatives (FN) were determined in Excel in order to calculate precision and sensitivity using the standard formulas (Figure 34).

$$\text{Pre.} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad \text{Sen.} = \text{Rec.} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Figure 34: Standard formulas for precision and sensitivity. TP = true positives. FP = false positives. FN = false negatives (Rainio et al., 2024).

Compounds classified as DNT-positive by both RiskHunt3r and the *comment* column were assigned to the TP group. Compounds predicted as positive by the model but classified as DNT-negative by RiskHunt3r were categorized as FP, while the opposite case constituted FN. The results are summarized in Table 5. The target-based screening approach identified a very high proportion of DNT-positive substances as toxic (96%), Moreover, when the overall model predicted a substance as toxic, it was indeed toxic in 78% of cases. Overall, the results indicate a highly sensitive and reliable model.

	original	w/o CHEMBL3522	threshold + 0.5	threshold + 1
TP	328	275	124	23
FP	94	75	39	12
FN	12	65	216	317
Precision	78%	79%	76%	66%
Sensitivity	96%	81%	36%	7%

Table 5: Results of validation with DNT-list. TP = true positive. FP = false positive. FN = false negative. Original = all targets & IDG threshold. w/o CHEMBL3522 = leave CHEMBL3522 out & IDG threshold. Threshold + 0,5 = all targets & IDG threshold+0,5. Threshold + 1 = all targets & IDG threshold+1.

The distribution of pChEMBL values above the IDG threshold was very heterogeneous across targets. As shown in Figures 29-33, the targets CHEMBL3522 (cytochrome P450 17A1) exhibited a particularly high number of highlighted values. This target is an enzymatic protein that plays a crucial role in steroid biosynthesis.

To assess the influence of CHEMBL3522 on the overall classification and to evaluate how the performance metrics would change, the same analysis was repeated while excluding this target. The results showed a slight increase in precision but a substantial reduction in sensitivity. This indicates that while some false positives were eliminated, many DNT-positive substances were no longer correctly identified. Thus, CHEMBL3522 is critical for identification of numerous harmful substances, and its biological role in sex hormone production may explain the high number of red pChEMBL values. Given that high sensitivity is particularly important in toxicity assessments, this enzyme should not be excluded from the overall classification.

Another approach was to investigate whether the results were dependent on the IDG threshold since the originally developed models were regression models that predicted numeric values which were transformed into binary activity classes. For this, the IDG threshold was increased by 0,5 and 1. For example, targets with an original IDG-threshold of 6 were re-evaluated using thresholds of 6,5 and 7 to determine whether the original threshold was too loose or not. Increasing the threshold by 0,5 resulted in a slight increase in precision but a significant decrease in sensitivity. While the numbers of TP and FP decreased markedly, the number of FN increased eighteen-fold. Increasing the threshold by 1 produced the same effect but to a much greater extent. Thus, increasing the threshold proved overly restrictive and did not lead to improved performance metrics.

In summary, the original screening approach yielded the most favorable results and appears to be practically useful. For this thesis, high sensitivity is prioritized, as it is preferable to incorrectly classify a substance as toxic rather than to incorrectly classify it as safe.

4.3.3. Results of the DART-list prediction

The DART list provided by the ECHA was processed and evaluated using the same method as applied to the RiskHunt3r DNT list. The key difference is that all substances in this list are classified as DART-positive, with no DART-negative compounds included. This is because, only substances known to cause DART effects were filtered on ECHAs website and then downloaded. Otherwise, the dataset would have comprised more than 300.000 substances, which would have exceeded the available computational capacity. As described previously, all pChEMBL values exceeding the IDG-threshold were highlighted in a blue color (Figure 35). Again, CHEMBL3522 stood out with a high number of highlighted values, which could be explained by its strong involvement in mechanisms relevant to reproductive toxicity again.

CAS no.	Name	CHEMBL1871	CHEMBL1978	CHEMBL206	CHEMBL210	CHEMBL221	CHEMBL228	CHEMBL230	CHEMBL239	CHEMBL325	CHEMBL3522	CHEMBL3577	CHEMBL402	comment
76420-72-9	Enalaprilat	5,656	5,014	4,985	5,984	5,381	6,138	6,047	5,903	5,372	6,784	5,679	5,792	Positive
53179-09-2	Disisomicin pentakis(sulphate)	6,316	5,488	5,482	5,920	5,211	5,349	5,436	6,400	5,645	6,025	5,788	5,797	Positive
100-41-4	ethylbenzene	6,169	5,076	4,409	5,940	5,869	4,896	4,847	5,082	4,910	5,654	6,368	5,906	Positive
745-65-3	Alprostadil	5,893	5,132	4,832	6,173	4,964	6,254	5,126	6,176	5,497	6,332	5,854	5,788	Positive
108-94-1	cyclohexanone	6,707	5,094	4,243	6,139	5,514	4,864	4,723	5,154	4,722	5,539	6,290	5,709	Positive
684-16-2	Hexafluoroacetone	6,785	4,590	5,021	6,277	5,450	5,072	5,410	5,107	4,524	5,886	6,110	5,873	Positive
51146-57-7	(2R)-2-(4-Isobutylphenyl)propanoic acid	6,383	4,952	4,996	5,995	5,662	4,908	5,443	5,168	5,119	5,964	6,300	5,894	Positive
30745-55-2	Hydroxyaluminium bis(2-ethylhexanoate	6,522	4,998	4,795	6,188	5,569	4,730	5,108	5,023	4,465	5,788	6,263	5,761	Positive
302-25-0	11β,17,21-trihydroxypregna-1,4-diene-3	6,342	5,355	5,098	5,978	5,015	5,732	5,572	5,954	5,157	6,015	5,957	6,024	Positive
1070-11-7	[S-(R*,R*)]-2,2'-(ethylenediimino)dibutan	6,160	5,261	4,683	6,643	5,506	5,908	4,996	5,458	4,954	5,980	5,968	5,985	Negative
814-89-1	Cobalt oxalate	6,100	5,468	4,438	6,085	5,269	4,052	5,392	5,167	5,051	5,753	5,942	5,754	Negative
33629-47-9	4-(tert-butyl)-N-sec-butyl-2,6-dinitroanil	6,356	5,205	4,536	6,171	5,467	6,193	5,361	5,637	5,356	6,527	5,577	5,772	Positive
1257-78-9	Ethane-1,2-disulphonic acid, compound	5,993	6,037	6,110	5,755	5,596	6,035	5,456	5,637	5,224	7,139	5,964	5,875	Positive
41556-26-7	Bis(1,2,2,6,6-pentamethyl-4-piperidyl) se	5,981	5,169	5,834	6,317	5,210	6,459	5,430	7,553	5,199	6,459	6,213	5,921	Positive
9002-61-3	Gonadotropin, chorionic	6,277	4,558	4,846	6,474	5,513	5,158	5,543	5,871	5,805	6,337	5,686	5,379	Positive
1638-05-7	2,7,11-trimethyl-13-(2,6,6-trimethylcyclc	5,850	4,834	6,314	6,551	5,398	5,636	5,287	5,811	5,215	6,399	5,817	5,790	Positive
29094-61-9	Glipizide	5,952	5,030	4,916	5,711	5,272	5,969	5,475	6,015	6,344	6,682	5,693	6,094	Positive
107534-96-3	tebuconazole (ISO) | 1-(4-chlorophenyl)-4	5,645	6,669	5,091	6,437	5,405	6,115	5,927	5,812	5,506	6,860	5,821	5,773	Positive
36282-47-0	trans-(±)-2-[(dimethylamino)methyl]-1-(3	6,685	5,710	6,016	6,932	5,431	5,812	5,509	5,555	5,451	6,444	5,988	5,769	Positive
108-95-2	phenol | carbolic acid | monohydroxybenz	6,279	4,942	4,441	6,480	5,464	4,620	5,448	5,108	4,658	5,686	6,104	5,926	Positive
61417-49-0	Tris(isooctadecanoato-O)(propan-2-olat	6,219	4,991	5,665	6,633	5,192	5,435	6,093	5,859	5,403	6,175	5,860	5,505	Positive
125306-83-4	N,N-diethyl-3-(2,4,6-trimethylbenzenesu	5,898	5,599	4,963	6,010	5,250	6,389	5,990	5,842	5,186	6,974	5,797	6,146	Positive
688-84-6	2-ethylhexyl methacrylate	6,319	5,309	4,723	6,177	5,643	4,753	5,170	5,030	5,097	5,929	6,185	5,738	Positive
3385-03-3	Flunisolide	6,402	4,554	5,382	6,028	4,884	5,746	5,198	6,101	4,517	6,129	6,045	6,519	Positive
133454-47-4	Iloperidone	5,818	5,574	6,148	6,282	5,193	7,091	4,994	6,110	6,646	7,304	5,643	7,096	Positive
86290-81-5	Gasoline | Low boiling point naphtha - un	6,362	5,596	6,284	6,222	5,487	6,039	5,604	5,629	5,976	6,666	5,987	5,848	Positive
13406-89-8	Lead(2+) 2,4-dinitroresorcinolate	6,449	4,895	4,354	6,062	5,374	4,991	5,521	5,385	5,007	5,769	5,419	5,904	Negative
6639-99-2	Estra-1,3,5,7,9-pentaene-3,17α-diol	7,123	5,597	7,353	6,965	5,466	6,001	5,975	5,594	4,900	6,490	6,143	5,676	Positive
5630-53-5	Tibolone	6,967	6,121	5,966	6,720	4,865	6,683	6,119	5,394	4,676	6,059	6,104	5,721	Positive
2425-79-8	1,4-bis(2,3 epoxypropoxy)butane | butan	6,168	5,241	4,463	6,303	5,342	6,007	4,745	5,615	4,649	5,908	5,777	5,964	Negative
1020668-59-0	5-[3-[(1S)-2-(difluoromethoxy)-1-methyl	5,986	5,220	4,607	6,009	5,447	6,194	5,728	6,086	6,144	7,084	6,057	7,044	Positive
80-92-2	5-β-pregnane-3-α,20-α-diol	6,525	5,635	5,457	6,211	4,926	6,152	5,459	5,439	4,733	6,042	6,168	5,911	Positive
123997-26-2	EPRINOMECTIN	6,074	4,660	7,044	6,546	5,123	5,616	5,330	7,245	5,817	6,282	6,371	6,493	Positive
684-93-5	1-methyl-1-nitrosourea	6,156	5,857	4,234	6,169	5,466	5,601	5,325	5,232	5,509	5,736	5,975	5,854	Negative
3693-39-8	Fluclorolone acetonide	6,414	5,008	5,022	6,203	4,930	5,860	5,456	6,087	4,695	6,084	6,297	6,145	Positive
143322-57-0	(R)-5-bromo-3-(1-methyl-2-pyrro	6,622	4,852	5,831	6,203	5,938	7,328	5,525	5,445	4,897	6,830	6,042	5,931	Positive
4721-69-1	Oxabolone	7,686	5,231	5,763	6,211	4,892	5,949	5,332	5,545	4,629	5,912	6,124	5,654	Positive
107724-20-9	(7α,11α,17α)-9,11-Epoxy-17-hydroxy-3	5,754	4,284	5,031	6,081	4,852	5,912	5,445	5,974	4,939	6,154	6,124	6,339	Positive
121-25-5	Amprolium	6,120	5,486	4,705	6,147	5,682	6,075	6,710	5,515	4,640	6,386	5,914	6,123	Positive
50-44-2	Mercaptopurine	6,210	5,252	4,669	6,011	5,240	4,869	6,023	5,276	4,510	6,039	6,146	5,911	Positive
80-43-3	bis(α,α-dimethylbenzyl) peroxide	5,546	5,166	6,070	6,550	5,705	4,924	5,399	5,539	5,308	6,351	5,940	5,861	Positive
202825-46-5	(+)-(S)-2-[[p-[(m-fluorobenzyl)oxy]benzyl	5,719	5,383	6,112	6,645	5,579	6,008	5,368	5,603	5,757	6,577	5,714	6,181	Positive
108-57-6	1,3-divinylbenzene	6,173	4,987	4,236	5,966	5,852	4,347	5,502	5,111	5,033	5,852	6,342	5,948	Positive
19774-82-4	2-butyl-3-benzofuryl 4-[2-(diethylamino)	5,639	5,165	5,829	7,100	5,263	5,683	6,254	6,154	6,158	6,882	6,100	6,202	Positive
1255-49-8	17β-hydroxyandrost-4-en-3-one 3-phen	6,670	6,086	5,684	5,987	4,789	6,380	5,377	5,977	4,912	6,341	6,288	5,748	Positive
108-98-5	Benzenethiol	6,279	5,459	4,257	5,954	5,806	4,469	4,860	5,082	4,936	5,755	6,173	5,872	Positive
68-12-2	N,N-dimethylformamide | d	6,588	5,568	4,506	6,119	5,410	5,841	4,968	5,081	4,842	5,803	6,267	5,775	Positive
33467-79-7	(2E,4E)-hepta-2,4-dien-1-ol	6,181	5,184	4,730	6,101	5,427	5,988	5,150	5,000	4,449	5,734	6,377	5,889	Positive
9002-70-4	Gonadotropin, pregnant mare serum	7,135	6,586	5,796	6,261	5,369	5,586	5,941	5,511	5,258	6,269	6,361	5,648	Positive
61413-54-5	Rolipram	6,574	5,693	5,045	6,755	5,476	6,985	5,566	5,696	4,808	6,380	5,709	5,852	Positive
466-99-9	Hydromorphone	7,000	6,000	5,586	6,924	5,177	6,237	5,349	5,844	4,787	6,457	5,880	5,626	Positive
15262-86-9	17β-hydroxyandrost-4-ene-3-one 4-met	7,314	5,717	5,832	6,215	4,645	6,275	5,413	5,849	4,880	6,154	6,246	5,873	Positive
16463-74-4	11β,17,21-trihydroxypregn-4-ene-3,20-c	6,605	5,061	5,501	6,011	4,817	5,834	5,062	5,914	4,563	6,046	6,062	6,100	Positive
13838-16-9	Enflurane	6,841	5,648	5,045	6,325	5,511	4,987	5,097	5,084	4,810	5,949	6,086	5,862	Positive
2426-08-6	butyl glycidyl ether | butyl 2,3-epoxypropy	6,397	5,820	4,589	6,167	5,518	5,610	5,021	5,174	4,569	5,832	6,241	5,907	Positive
64681-08-9	(S)-dichloro[2-[[(2,3-dihydroxypropoxy)h	6,145	5,187	4,845	6,492	5,339	4,455	4,743	5,651	4,629	6,134	5,622	5,839	Positive
26761-45-5	2,3-epoxypropyl neodecanoate	6,344	4,770	4,620	6,636	5,680	5,382	5,528	5,518	5,291	6,111	6,122	5,791	Positive
16630-66-3	Methyl (methylthio)acetate	6,719	5,516	4,654	6,095	5,451	5,444	5,248	5,104	5,008	5,998	6,150	5,883	Positive
68892-13-7	1-phenyldecane-1,3-dione	6,044	5,168	4,935	6,272	5,436	5,378	5,553	5,591	5,715	6,494	6,098	5,779	Positive
552-94-3	Salsalate	5,704	5,361	5,488	6,600	5,271	4,752	5,595	5,333	5,224	6,259	5,600	6,314	Positive
843-55-0	4,4'-cyclohexylidenebisphenol	6,094	5,308	6,737	7,219	5,715	5,869	5,603	5,570	5,122	6,485	5,967	5,959	Positive
407-25-0	Trifluoroacetic anhydride	6,625	5,133	4,849	6,239	5,377	4,519	4,797	5,235	5,000	6,001	6,121	5,879	Positive
5234-06-0	[(2-naphthyloxy)methyl]oxirane	6,170	4,453	4,975	7,206	5,579	5,811	5,377	5,223	4,962	5,894	6,228	5,957	Positive
29205-06-9	6α-fluoro-11β,21-dihydroxy-16α-methyl	6,613	5,334	5,656	6,025	4,923	5,829	5,469	6,018	5,332	5,979	6,372	6,436	Positive
862-89-5	17β-hydroxyestr-4-en-3-one 17-undecar	7,030	5,381	5,912	6,291	4,836	6,617	5,197	6,494	4,990	6,151	6,232	5,962	Positive
88495-63-0	Artesunate	6,428	5,319	4,257	5,908	4,944	5,762	5,117	6,117	4,897	6,175	5,884	5,932	Positive
2582-30-1	Aminoguanidinium hydrogen carbonate	6,454	6,205	4,354	6,309	5,493	5,119	5,262	5,319	5,405	5,581	6,032	5,689	Positive
94249-01-1	Acetyl dihydrogen phosphate, lithium pot	6,149	5,678	4,483	6,018	5,286	4,833	4,833	5,298	4,714	5,927	5,894	5,717	Negative

Figure 35: A representative subset of results of the DART-list. Comment = classification based on IDG threshold. Blue highlighted fields = predicted pChEMBL > IDG threshold.

Since only DART-positive classifications were available, only true positives and false negatives could be determined, and sensitivity was the only performance metric that could be calculated. The values reported in Table 6 show that 97% of the substances classified as reproduction-toxic were correctly identified as positive. This result demonstrates excellent performance for risk identification and highlights the robustness of the model.

	ECHA
TP	4722
FN	169
Sensitivity	97%

Table 6: Results of validation with DART-list. TP = true positive. FP = false positive. FN = false negative.

5. Conclusion

In summary, this thesis shows that the AOP-Wiki provided a valuable and comprehensive base for understanding the underlying mechanistic pathways leading to DART. The systematic analysis revealed that several MIEs and KEs occur in multiple AOPs, which indicates that AOP-Wiki highlights the linked nature of toxicological mechanisms and supports the concept of AOP networks. However, the frequent occurrence is limited to a subset of MIEs which could mean that those MIEs have a central role in multiple pathological and toxicological processes but may also imply higher number of experimental research and thereby be overrepresented. Similar trends were observed for the occurrence of biological targets and their number of bioactivity data in ChEMBL. Only 66% of the MIEs have specific molecular targets and only 56% of those targets are single proteins and thereby suitable for QSAR modeling in this work. This led to a loss of some mechanistic information encoded in AOPs that could not be included into the modeling framework. Furthermore, out of the fifteen targets associated to DART, only twelve were eligible for modeling since the excluded three ones did not provide sufficient biological IC50 data in ChEMBL. These limitations could mean that not all compounds leading to DART may be detected, since DART can arise through different mechanisms that are not fully captured by the twelve targets included in this thesis. This shows how model performance is highly dependent on the amount of experimental data.

Another limitation of this work was the grouping of AOs into body systems or organs. This process was carried out manually which introduced subjectivity and potential for misclassification. It would be interesting if AOP-Wiki linked every AO automatically to a standardized group or Globally Harmonized System (GHS) hazard classes enabling users to systematically query AOPs associated with certain toxicological endpoints.

The internally validated QSAR models showed robustness and stability with nine out of twelve targets resulting in R2 values above 0,5 as well as similar values between cross-validation and holdout test set which indicates limited overfitting. Yet, the ability for generalization of these models is clearly limited, as seen in the low or negative R2 results

of eight targets during prospective validation with the updated ChEMBL version. These outcomes are expected when models are applied to new compounds that are outside the chemical space of the training set through which a reliable structure-activity relationship cannot be performed. This suggests that the random splitting of datasets into training and test set may not reflect realistic chemical spaces. Instead, selecting compounds located at the edge of the chemical space and using them as the training set could have led to more realistic predictions and increased robustness. The consideration of applicability domain might further strengthen predictive performance.

Nevertheless, the validation of our models using two different external datasets showed high sensitivities (96% and 97%) meaning that the models successfully identified almost all truly toxic substances as toxic. A high sensitivity is especially critical for toxicological screenings. The results of this work also suggest that CHEMBL3522 (CYP17A1) plays a crucial role for the high sensitivity since an exclusion of this target led to increased false negatives and decreased true positives. This observation is consistent with the enzyme's pivotal function in steroid hormone synthesis, a key biological process underlying DART. In addition, the IDG threshold provided the best results in comparison to increasing the threshold which resulted in less false positives at the cost of increased false negatives.

Overall, the models presented in this thesis showed robust and promising performance while leaving clear room for improvement. Nevertheless, this work supports the statement that AOP-derived QSAR models represent valuable tools for early, cost-effective and fast hazard identification. By supporting the principles of 3R (reduce, replace, refine) of animal testing, such approaches are important for the advancement of NAMs in risk assessment.

List of Abbreviations

DART	developmental and reproductive toxicity
ROS	reactive oxygen species
SOD	superoxide dismutase
QSAR	quantitative structure-activity relationship
REACH	Registration, evaluation, authorisation and restriction of chemical
ECHA	european chemicals agency
NAM	new approach methodologies
AOP	adverse outcome pathway
MIE	molecular initiating event
KE	key event
KER	key event relationship
AO	adverse outcome
AOP-KB	AOP-Knowledgebase
OECD	Organisation for Economic Co-operation and Development
FDA	food and drug administration
ML	machine learning
IDG	illuminating the druggable genome
MAE	mean absolute error
RMSE	root mean squared error
DNT	developmental neurotoxicity
C&L	classification & labelling
CLP	classification, labelling, packaging
AhR	aryl hydrocarbon receptor
ACE2	angiotension converting enzyme 2
IC50	inhibitory concentration 50%
Ki	inhibtion constant
HMG CoA	3-hydroxy-3-methylglutaryl-coenzyme A
TP	true positive
FP	false positive
FN	flase negative
GHS	globally harmonized system
StAR	Steroidogenic Acute Regulatory Protein

List of Figures

List of Tables

Appendix

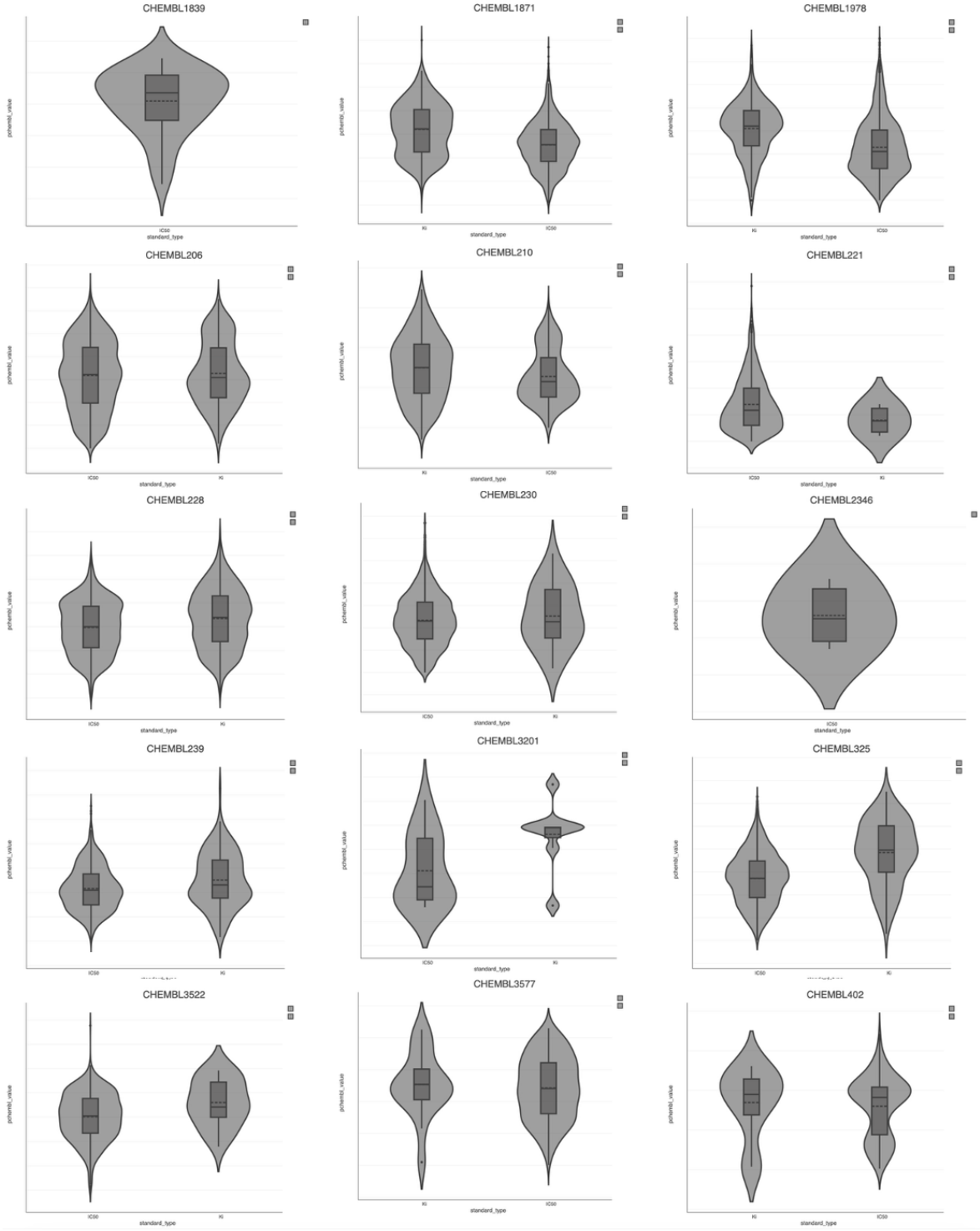

Figure 36: Violin plots of IC50 and Ki for each target

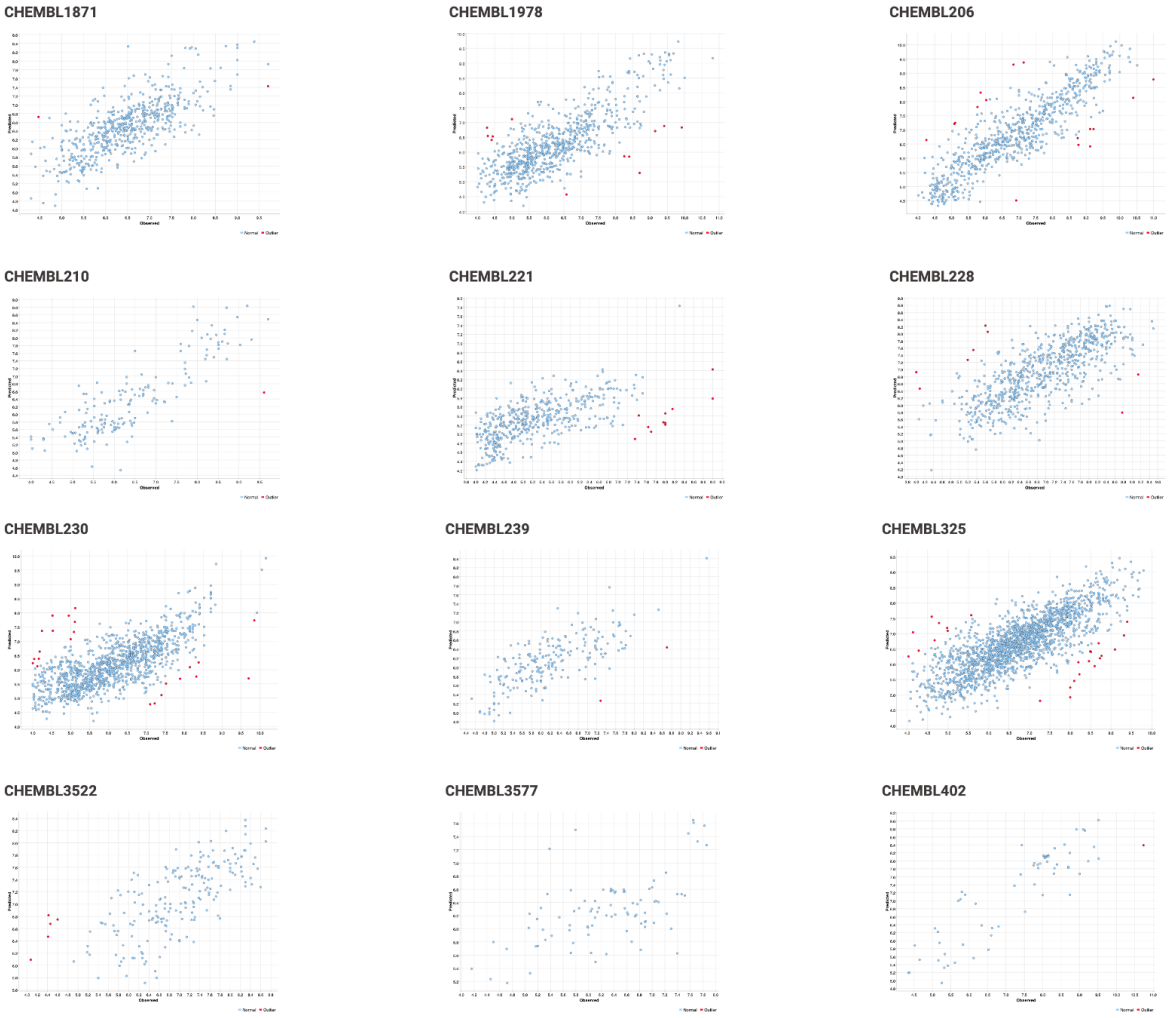

Figure 37: Scatter plot for outlier detection for every target. x-axis = observed, y-axis = predicted.

Bibliography

Abubakar, Souley Boukari, Abdulsalam Ya'u Gital, Fatima Umar Zambuk, 2025. Evaluation of a Hyperparameter Tuned Random Forest Algorithm Based on Artificial Bee Colony for Improving Accuracy and Precision of Crime Prediction Model. AJBES. https://doi.org/10.11648/j.ajbes.20251103.16

Ahmad, S., Bano, N., Raza, K., 2025. RCSB Protein Data Bank: revolutionising drug discovery and design for over five decades. Med Data Min 8, 8. https://doi.org/10.53388/MDM202508008

"AOP-Wiki." n.d. Accessed on June 29, 2025. https://aopwiki.org

Ball, T., Barber, C.G., Cayley, A., Chilton, M.L., Foster, R., Fowkes, A., Heghes, C., Hill, E., Hill, N., Kane, S., Macmillan, D.S., Myden, A., Newman, D., Polit, A., Stalford, S.A., Vessey, J.D., 2021. Beyond adverse outcome pathways: making toxicity predictions from event networks, SAR models, data and knowledge. Toxicology Research 10, 102–122. https://doi.org/10.1093/toxres/tfaa099

"Chembl35." n.d. Accessed on July 7, 2025. https://www.ebi.ac.uk/chembl/

"Chembl36." n.d. Accessed on November 1, 2025. https://www.ebi.ac.uk/chembl/

Druggable Genome Initiative, n.d. IDG Protein Families. Accessed on August 1, 2025. https://druggablegenome.net/IDGProteinFamilies

Dugershaw, B.B., Aengenheister, L., Hansen, S.S.K., Hougaard, K.S., Buerki-Thurnherr, T., 2020. Recent insights on indirect mechanisms in developmental toxicity of nanomaterials. Part Fibre Toxicol 17, 31. https://doi.org/10.1186/s12989-020-00359-x

"ECHA." n.d. Accessed on November 9, 2025. https://echa.europa.eu/information-on-chemicals/cl-inventory-database

Feng, H., Zhang, L., Li, S., Liu, L., Yang, T., Yang, P., Zhao, J., Arkin, I.T., Liu, H., 2021. Predicting the reproductive toxicity of chemicals using ensemble learning methods and molecular fingerprints. Toxicology Letters 340, 4–14. https://doi.org/10.1016/j.toxlet.2021.01.002

Fillbrunn, A., Dietz, C., Pfeuffer, J., Rahn, R., Landrum, G.A., Berthold, M.R., 2017. KNIME for reproducible cross-domain analysis of life science data. Journal of Biotechnology 261, 149–156. https://doi.org/10.1016/j.jbiotec.2017.07.028

Gaulton, A., Bellis, L.J., Bento, A.P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., Overington, J.P., 2012. ChEMBL: a

large-scale bioactivity database for drug discovery. Nucleic Acids Research 40, D1100–D1107. https://doi.org/10.1093/nar/gkr777

Gaulton, A., Hersey, A., Nowotka, M., Bento, A.P., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L.J., Cibrián-Uhalte, E., Davies, M., Dedman, N., Karlsson, A., Magariños, M.P., Overington, J.P., Papadatos, G., Smit, I., Leach, A.R., 2017. The ChEMBL database in 2017. Nucleic Acids Res 45, D945–D954. https://doi.org/10.1093/nar/gkw1074

Goodsell, D.S., Zardecki, C., Di Costanzo, L., Duarte, J.M., Hudson, B.P., Persikova, I., Segura, J., Shao, C., Voigt, M., Westbrook, J.D., Young, J.Y., Burley, S.K., 2020. RCSB Protein Data Bank: Enabling biomedical research and drug discovery. Protein Science 29, 52–65. https://doi.org/10.1002/pro.3730

Handelman, G.S., Kok, H.K., Chandra, R.V., Razavi, A.H., Lee, M.J., Asadi, H., 2018. ED octor: machine learning and the future of medicine. J Intern Med 284, 603–619. https://doi.org/10.1111/joim.12822

Jiang, C., Yang, H., Di, P., Li, W., Tang, Y., Liu, G., 2019. In silico prediction of chemical reproductive toxicity using machine learning. J of Applied Toxicology 39, 844–854. https://doi.org/10.1002/jat.3772

Kleinstreuer, N.C., Sullivan, K., Allen, D., Edwards, S., Mendrick, D.L., Embry, M., Matheson, J., Rowlands, J.C., Munn, S., Maull, E., Casey, W., 2016. Adverse outcome pathways: From research to regulation scientific workshop report. Regulatory Toxicology and Pharmacology 76, 39–50. https://doi.org/10.1016/j.yrtph.2016.01.007

Knapen, D., Vergauwen, L., Villeneuve, D.L., Ankley, G.T., 2015. The potential of AOP networks for reproductive and developmental toxicity assay development. Reproductive Toxicology 56, 52–55. https://doi.org/10.1016/j.reprotox.2015.04.003

Miller, L.B., Feuz, M.B., Meyer, R.G., Meyer-Ficca, M.L., 2024. Reproductive toxicology: keeping up with our changing world. Front. Toxicol. 6, 1456687. https://doi.org/10.3389/ftox.2024.1456687

Myden, A., Cayley, A., Davies, R., Jones, J., Kane, S., Newman, D., Payne, M.P., Ude, V.C., Vessey, J.D., White, E., Fowkes, A., 2024. A developmental and reproductive toxicity adverse outcome pathway network to support safety assessments. Computational Toxicology 31, 100325. https://doi.org/10.1016/j.comtox.2024.100325

Peng, Y., He, Q., 2024. Reproductive toxicity and related mechanisms of micro(nano)plastics in terrestrial mammals: Review of current evidence. Ecotoxicology and Environmental Safety 279, 116505. https://doi.org/10.1016/j.ecoenv.2024.116505

Rainio, O., Teuho, J., Klén, R., 2024. Evaluation metrics and statistical tests for machine learning. Sci Rep 14, 6086. https://doi.org/10.1038/s41598-024-56706-x

"RCSB PDB." n.d. Accessed on July 29, 2025. https://www.rcsb.org

Sandunil, K., Bennour, Z., Ben Mahmud, H., Giwelli, A., 2024. Effects of tuning decision trees in random forest regression on predicting porosity of a hydrocarbon reservoir. A case study: volve oil field, north sea. Energy Adv. 3, 2335–2347. https://doi.org/10.1039/D4YA00313F

Scialli, A.R., 2008. The challenge of reproductive and developmental toxicology under REACH. Regulatory Toxicology and Pharmacology 51, 244–250. https://doi.org/10.1016/j.yrtph.2008.04.008

Shi, W., Zhang, Z., Li, M., Dong, H., Li, J., 2024. Reproductive toxicity of PFOA, PFOS and their substitutes: A review based on epidemiological and toxicological evidence. Environmental Research 250, 118485. https://doi.org/10.1016/j.envres.2024.118485

Sydow, D., Wichmann, M., Rodríguez-Guerra, J., Goldmann, D., Landrum, G., Volkamer, A., 2019. TeachOpenCADD-KNIME: A Teaching Platform for Computer-Aided Drug Design Using KNIME Workflows. J. Chem. Inf. Model. 59, 4083–4086. https://doi.org/10.1021/acs.jcim.9b00662

The UniProt Consortium, Bateman, A., Martin, M.-J., Orchard, S., Magrane, M., Ahmad, S., Alpi, E., Bowler-Barnett, E.H., Britto, R., Bye-A-Jee, H., Cukura, A., Denny, P., Dogan, T., Ebenezer, T., Fan, J., Garmiri, P., Da Costa Gonzales, L.J., Hatton-Ellis, E., Hussein, A., Ignatchenko, A., Insana, G., Ishtiaq, R., Joshi, V., Jyothi, D., Kandasaamy, S., Lock, A., Luciani, A., Lugaric, M., Luo, J., Lussi, Y., MacDougall, A., Madeira, F., Mahmoudy, M., Mishra, A., Moulang, K., Nightingale, A., Pundir, S., Qi, G., Raj, S., Raposo, P., Rice, D.L., Saidi, R., Santos, R., Speretta, E., Stephenson, J., Totoo, P., Turner, E., Tyagi, N., Vasudev, P., Warner, K., Watkins, X., Zaru, R., Zellner, H., Bridge, A.J., Aimo, L., Argoud-Puy, G., Auchincloss, A.H., Axelsen, K.B., Bansal, P., Baratin, D., Batista Neto, T.M., Blatter, M.-C., Bolleman, J.T., Boutet, E., Breuza, L., Gil, B.C., Casals-Casas, C., Echioukh, K.C., Coudert, E., Cuche, B., De Castro, E., Estreicher, A., Famiglietti, M.L., Feuermann, M., Gasteiger, E., Gaudet, P., Gehant, S., Gerritsen, V., Gos, A., Gruaz, N., Hulo, C., Hyka-Nouspikel, N., Jungo, F., Kerhornou, A., Le Mercier, P., Lieberherr, D., Masson, P., Morgat, A., Muthukrishnan, V., Paesano, S., Pedruzzi, I., Pilbout, S., Pourcel, L., Poux, S., Pozzato, M., Pruess, M., Redaschi, N., Rivoire, C., Sigrist, C.J.A., Sonesson, K., Sundaram, S., Wu, C.H., Arighi, C.N., Arminski, L., Chen, C., Chen, Y., Huang, H., Laiho, K., McGarvey, P., Natale, D.A., Ross, K., Vinayaka, C.R., Wang, Q., Wang, Y., Zhang, J., 2023. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Research 51, D523–D531. https://doi.org/10.1093/nar/gkac1052

Tiwari, S., Chadha, D.S., Chauhan, D.R., 2025. Exploring Climate Change Dynamics Using Machine Learning and Deep Learning Approaches. https://doi.org/10.52783/jisem.v10i52s.10677

Toragall, M., C. Ghagane, S., B. Nerli, R., B. Hiremath, M., 2022. Reproductive Toxicology: An Update, in: Wu, W. (Ed.), Male Reproductive Anatomy. IntechOpen. https://doi.org/10.5772/intechopen.101404

"Uniprot." n.d. Accessed on July 7, 2025. https://www.uniprot.org

Wangikar, P., Ahmed, T., Vangala, S., 2011. Toxicologic pathology of the reproductive system, in: Reproductive and Developmental Toxicology. Elsevier, pp. 1003–1026. https://doi.org/10.1016/B978-0-12-382032-7.10076-1

Weyrich, A., Joel, M., Lewin, G., Hofmann, T., Frericks, M., 2022. Review of the state of science and evaluation of currently available *in silico* prediction models for reproductive and developmental toxicity: A case study on pesticides. Birth Defects Research 114, 812–842. https://doi.org/10.1002/bdr2.2062

Zhang, H., Ren, J.-X., Kang, Y.-L., Bo, P., Liang, J.-Y., Ding, L., Kong, W.-B., Zhang, J., 2017. Development of novel in silico model for developmental toxicity assessment by using naïve Bayes classifier method. Reproductive Toxicology 71, 8–15. https://doi.org/10.1016/j.reprotox.2017.04.005