

A Comprehensive Evaluation of Machine Learning Pipelines for Hormone-related Toxicity Prediction

Lisa-Marie Rolli

Saarland University

Loulwah Arnaout

Saarland University

Andrea Volkamer

`volkamer@cs.uni-saarland.de`

Saarland University

Research Article

Keywords: toxic endpoint prediction, benchmarking, statistical analysis, hormone receptor activity prediction, compound representations, dimension reduction, machine learning

Posted Date: September 30th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10470980/v2>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

A Comprehensive Evaluation of Machine Learning Pipelines for Hormone-related Toxicity Prediction

Lisa-Marie Rolli¹, Loulwah Arnaout¹, Andrea Volkamer^{1*}

¹Data Driven Drug Design, Center for Bioinformatics, Saarland Informatics Campus, Saarland University, PharmaScienceHub (PSH), Campus, Saarbrücken, 66123, Saarland, Germany.

*Corresponding author(s). E-mail(s): volkamer@cs.uni-saarland.de;
Contributing authors: lisa-marie.rolli@uni-saarland.de;
loar00001@stud.uni-saarland.de;

Abstract

A major challenge in hazard assessment of drug and material discovery is the identification of potentially toxic compounds. To assess potential toxicity of the compounds, different assays are conducted; however, they are usually costly and time-consuming. Therefore, machine learning (ML) is often used to estimate the effect of a compound on a specific endpoint based on molecular features. However, finding an appropriate ML model, compound representation, and dimensionality reduction (DR) method remains challenging. Therefore, we benchmark a wide range of ML algorithms, molecular representations, and DR approaches to identify effective combinations for predicting toxicity across ToxCast endpoints. In particular, we consider 52 hormone receptor activity assays (e.g., estrogen or androgen receptor), each having roughly 500-9,000 tested compounds. To characterize the compounds, we compare transformer-based embeddings, structural fingerprints (Morgan and MACCS), and physicochemical descriptors. For these features, we apply different DR techniques (i.e., principal component analysis, variance-based selection, mutual information, and minimum-redundancy-maximum-relevance). Finally, we train different ML models (i.e., random forest (RF), multi-layer perceptron (MLP), CatBoost, support vector machine (SVM), and TabPFN). Using a nested 5×4 cross-validation setup, we evaluate 100 model-feature-DR combinations per assay. Performance is assessed using the Matthews correlation coefficient (MCC), one of the most appropriate metrics for imbalanced data. Despite the fact that the winning algorithm of the toxicity prediction challenge *Tox21* was a deep neural network (NN), our results show that RF trained on physicochemical properties without DR is the combination that is most often ranked as top 1 (in 12/52 assays). Moreover,

it ranks among the top 5 out of 100 combinations in half (26/52) of the tested assays. When ranking models per assay, RFs significantly outperform all other models, achieving an average rank of 1.4 and an MCC of 0.207. SVM, CatBoost, follow with ranks around 3 and MCC values near 0.15, while MLP and TabPFN show the worst performance with ranks around 4 and average MCC values around 0.13. Given that tree-based methods also show a short runtime and high interpretability, we recommend employing them for toxicity prediction.

Scientific Contribution:

This study presents the first comprehensive benchmark of ML models, molecular representations, and DR methods for hormone receptor activity assays in ToxCast. In contrast to prior work, we systematically evaluate complete modeling pipelines with rigorous hyperparameter tuning and chemically-informed data splitting to avoid similarity bias, supported by extensive statistical analysis.

Keywords: toxic endpoint prediction, benchmarking, statistical analysis, hormone receptor activity prediction, compound representations, dimension reduction, machine learning

1 Introduction

The identification of toxic compounds is a crucial step in the development of novel compounds, including drugs, materials, food additives, bio-based alternatives, and pesticides. To reduce the discovery time and cost, toxicity should be assessed as early as possible. Experimental toxicity assays are typically used to provide endpoint measurements; however, they are often costly, time-consuming, and may involve animal testing. Consequently, computational, often machine learning (ML)-based, toxicity prediction has become an important strategy to prioritize and filter compounds early in the discovery pipeline.

To train ML models that estimate the potential toxicity of a given compound, the compound must be represented in a way that the model can process. ‘Classical’ ML models (e.g., tree-based methods or support vector machines) require tabular input, where each sample (compound) is represented by a vector of fixed length. This includes compound representations such as physicochemical properties and structural fingerprints [1]. With the rise of deep learning (DL) models, e.g., graph neural networks or transformers, more complex representations of varying length, including graph or text-based encodings, can be used. However, Sultan et al. [2] showed that these more complex compound representations do not consistently improve prediction performance compared to physicochemical properties across several molecular property prediction datasets. Moreover, increased model complexity generally requires larger training datasets to adapt the parameters, while simpler models can achieve competitive performance with less data [3]. Due to the limited availability of toxicity data (i.e., ≤ 9000 compounds per assay in the ToxCast database [4]), data scarcity further constrains the potential advantages of DL approaches. This raises the question of whether complex representations or architectures truly outperform simpler, physically and chemically informed features in realistic toxicity prediction scenarios.

Throughout this article, we use the term ‘ML pipeline’ as the following workflow: First, the compounds in the dataset are encoded in a way that fits the required input for the ML algorithm, e.g. feature vectors. Then, the dataset is split into a training and a test part. This can be done randomly; however, it is common practice to perform a split based on chemical similarity to estimate predictive performance on other parts of the chemical space[5]. Once the dataset is split, one may use dimension reduction (DR) methods on the feature matrix. This can be particularly helpful if the compound representation contains noisy or redundant features. However, the effectiveness of DR methods is context-dependent and varies with the specific dataset and model. The next step is typically some hyperparameter optimization of the ML; often using a k-fold cross validation (CV) on the training set. Afterwards, the chosen model is trained on the training set and evaluated on the test set. If this pipeline is repeated for several splits, one can have several performance values per model allowing for better comparison of approaches, e.g., via statistical evaluation [6, 7].

Several benchmarking studies have evaluated subsets of this ML pipeline, including comparisons of structural fingerprints, or ML models on selected toxicity endpoints (cf. Table 1). Note that we focus here on toxic endpoint predictions only, more benchmarking studies have been performed on more general molecular property prediction tasks (e.g., [1, 8]). Existing work typically focuses on only one or two components of the pipeline and often lacks chemically informed data splitting or statistical significance testing. Wu et al. [1] is one of the most cited studies for molecular property prediction as they provide easy-to-use versions of datasets collected from literature that can be used for benchmarking - including parts of the ToxCast dataset, which is the focus of our work. In their study, they compare eight different ML models (i.e., logistic regression, XGBoost, support vector machine (SVM), influence relevance voting, bypass multitask networks, Weave models, multi-task neural network (NN), and graph convolutional models) on their processed version of the ToxCast dataset. Their results indicate that, when evaluated across all 600 endpoints, the Weave models, i.e., specialized graph convolutional networks that propagate information across all atom features during convolution, perform best. Yet, this performance is evaluated using a random data split rather than a chemically informed one. Moreover, there is no statistical test comparing the model performances. Russo et al. [9] benchmarked several ML models (i.e., Bernoulli Naive Bayes, AdaBoost decision tree, random forest (RF), and SVM) and molecular representations (i.e., ECFP6, FCFP6, ToxPrint, and MACCS keys) for estrogen receptor (ER) activity and found that RFs performed best, while the models performed similarly regardless of the compound representation used. However, the results are not supported by a statistical analysis. Zorn et al. [10] performed a follow-up study based on the Russo et al. [9] study design, i.e., they use the same models and hyperparameter tuning strategy to benchmark the different architectures. Because their earlier analysis on ER activity showed minimal differences across compound representations, the follow-up study focused solely on benchmarking model architectures. Zorn et al. [10] applied the pipeline to predict androgen receptor (AR) activity assay outcomes and performed statistical tests to compare model performances on the considered assays in a 5-fold CV. Their results indicate that the models perform similar. Belfield et al. [11] compare the performance of six different ML models (i.e.,

Table 1 Benchmarking studies for ML pipelines in toxic endpoint prediction. Columns indicate studied endpoints, number of assays, and which parts of the ML pipeline were benchmarked. Model abbreviations: LR (logistic regression), NB (naïve Bayes), BNB (Bernoulli naïve Bayes), SVM (support vector machine), RF (random forest), GBT (gradient boosting trees), XGB (Extreme Gradient Boosting), AdaBoost-DT (AdaBoost decision tree), MLP (multi-layer perceptron), NN (neural network), MTNN (multi-task neural network), GCN (graph convolutional network), GNN (graph neural network), Weave (Weave neural network), IRV (influence relevance voting), kNN (k-nearest neighbours), PSM (pretrained sequence model), SAR (structure–activity relationship reasoning).

Authors, Year	Endpoint(s)	# Assays	Feature type	DR	Model(s)
Wu [1], 2018	MoleculeNet (various)	600	–	–	LR, XGB, SVM, IRV, BPNN, Weave, MTNN, GCN
Russo [9], 2018	ER activity	8	ECFP6, FCFP6, ToxPrint, MACCS	–	BNB, AdaBoost-DT, RF, SVM
Zorn [10], 2020	AR activity	11	–	–	BNB, AdaBoost-DT, RF, SVM
Belfield [11], 2023	<i>T. pyriformis</i> growth inhibition & rat lethality	2	–	variance-based feature selection with comparison of different thresholds	RF, XGB, SVM, shallow NN, deep NN, kNN
Kim [12], 2024	Various	737	MACCS, Morgan, RDKit, Layered, Patterned	–	MLP, GBT, RF, kNN, LR, NB
Guo [13], 2026	Tox21 endpoints	12	–	–	‘simple’ models (e.g., RF), GNN, PSM, SAR
Our work , 2026	ER, AR, PR, GR activity	52	MACCS, Morgan, physico-chemical properties, SMILES embeddings	PCA, variance, MI, MRMR	SVM, RF, CatBoost, MLP, TabPFN

RF, extreme gradient boosting, SVM, shallow and deep NNs, and k-nearest neighbours) on two data sets, i.e., rat lethality and *T. pyriformis* growth rate inhibition. Moreover, they study how the number of selected features affects performance. However, they do not directly compare different DR methods. Kim et al. [12] performed a large-scale study on the ToxCast and Tox21 datasets. They evaluated different combinations of ML models (i.e., multi-layer perceptron (MLP), gradient boosting trees, RF, k-nearest neighbours, logistic regression, and naïve Bayes) and structural fingerprints (i.e., MACCS, Morgan, RDKit, Layered, and Patterned) on more than 700 assays. Although they do not conduct hyperparameter tuning or statistical tests, their work emphasizes that interpretability is important and concludes that MACCS fingerprints combined with RF provide an effective balance between performance and interpretability. Guo [13] performed a benchmarking study of different model types (i.e., ‘simple’ ML models (e.g., RFs), graph neural networks (GNNs), pretrained molecular sequence models, and rule-based structure activity relationship reasoning models) on different molecular property and activity prediction tasks, including the Tox21 endpoints. They

found that the simple ML models outperform the more complex architectures in the majority of tested tasks. In particular, the simple models also win 63% of the Tox21 tasks.

In our work, we present the first comprehensive benchmark that jointly evaluates compound representations, DR methods and ML models on 52 hormone receptor activity assays from the ToxCast database [4], a target class of high toxicological relevance that is central to endocrine disruption. The focus of this benchmark is on single-task models trained separately for each assay, with the goal of comparing different representations, DR, and ML approaches under a consistent train-test setting for different hormone activity-related assays. In particular, we consider four compound representations (i.e., physicochemical descriptors, MACCS and Morgan fingerprints, and transformer-based embeddings) with five dimensionality reduction (DR) approaches (i.e., principal component analysis, minimum-redundancy-maximum-relevance, variance-based selection, mutual information, and no DR), and five ML models (i.e., RF, SVM, CatBoost, MLP, and TabPFN). Fig. 1 illustrates the different parts of the hormone-related toxic endpoint prediction pipeline that we investigate in this study.

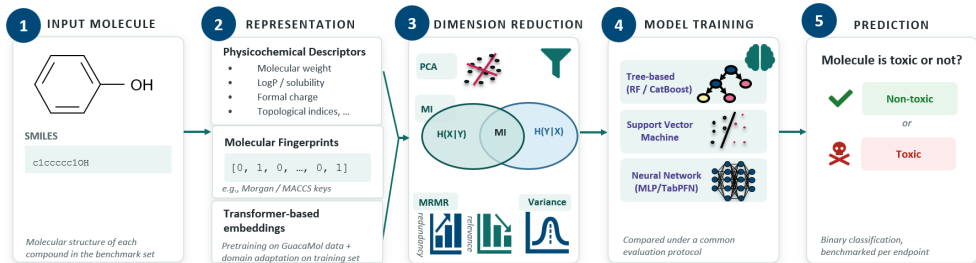


Fig. 1 Illustration of the parts of a toxicity prediction modeling pipeline investigated in this work.

To avoid similarity-based information leakage, we specifically ensure a split maximizing the chemical distance between the data points. Moreover, we use nested 5-fold CV to tune hyperparameters on the inner folds, while obtaining robust performance estimates on the outer folds, which in turn enables a statistical analysis of our results. This setup allows us to evaluate 100 distinct feature-DR-model combinations per assay; however, due to extensive hyperparameter optimization, more than 7,000 individual model instances are trained per assay.

Overall, our results show that physicochemical properties are the most informative data modality for the datasets and evaluation scheme considered. However, they do not significantly outperform transformer-based embeddings. Furthermore, while models without dimensionality reduction (DR) generally achieved the best results, our comparative analysis did not reveal a consistent advantage of either applying or omitting DR. Across all experiments, RF models demonstrated the strongest and most robust performance, achieving the highest average rank (1.4). Combining physicochemical properties with RFs and no DR resulted in the top-performing configuration among 100 evaluated approaches, ranking first in 12 and within the top five in 26 of 52 assays.

2 Materials and Methods

2.1 Data Collection and Processing

To establish a consistent dataset for downstream modeling, we systematically collected data from the ToxCast resource. For our analyses, we use the ToxCast database invit-rodb Version v4.2 (September 2024), which contains 1570 assays with 9614 unique compounds [4].

Assay endpoint selection: In this study, we focus on assays associated with hormone receptor activity measurements. Thus, we only consider assays, where the *intended target family* in the assay annotation document is *nuclear receptor* or *steroid hormone*. From these, we further restrict our selection to the 132 assays, where the *intended target family sub* is annotated as *androgens*, *estrogens*, *glucocorticoids*, *progestagens*, *steroidal*, or *steroid*.

Data collection: We automatically queried the CompTox Dashboard (Version 2.5.2, February 2025) [14, 15] to retrieve results for these assays using our own ToxCastScraper script¹. However, CompTox only contained 127 of 132 assays. Moreover, for all assays, we queried the CompTox Dashboard for mol files for all tested compounds and used RDKit (Version 2025.03.3) [16] to calculate canonical SMILES (Simplified Molecular Input Line Entry System) [17], which served as the basis for all other compound representations. Molecules without a mol file or for which RDKit could not calculate SMILES were discarded.

In the original ToxCast publication, Richard et al. [18] state that the chemical structures were curated to present the ‘chemical in the bottle’ as accurately as possible, including details such as hydration state, salt forms, complexed moieties, i.e. mixtures, and stereochemistry. Therefore, we intentionally refrained from applying common pre-processing procedures used in molecular ML workflows, such as salt removal or element filtering [19]. However, a subset of molecules, particularly metal-containing complexes, caused issues during the calculation of the RDKit physicochemical properties (cf. Table 2 and Section 2.2). To assess the impact of structure standardization on predictive performance, we conducted an additional experiment comparing models trained on the original structures with models trained on structures processed through a cleaning pipeline. For this purpose, we employed the MolPipeline package [19] including the SaltRemover and ElementFilter with default parameters. After applying this pipeline, we remove exact duplicates. Overall, we remove 825 from all 8785 available compounds.

As machine learning (ML) requires large amounts of data and we build per-assay models using a nested 5-fold cross validation (CV), we excluded all assays with fewer than 500 available compounds, leaving 59 assays (cf. Table 3).

Data splitting: To obtain statistically robust model performance estimates, we employed a nested CV strategy. Specifically, for each assay and model, we performed an outer 5-fold CV, yielding five performance estimates. Within each outer training split

¹We published the download script on GitHub: <https://github.com/volkamerlab/ToxCastScraper/releases/tag/v1>. Note that for newer versions of CompTox, the updated version of the scraper needs to be used.

Table 2 Table summarizing the smiles reported in the ToxCast dataset. Abbreviations: ‘mixture’: a mixture of more than one molecule, ‘metal complex’: a mixture, where at least one of the molecules has at least one metal atom, ‘salt’: mixture with at least one negatively and one positively charged molecule, ‘metal-containing’: single molecule with at least one metal atom

category	count	% molecules	% with issues in physicochemical properties
ordinary organic	6866	78.16	0.33
mixture	919	10.46	5.01
metal complex	489	5.57	99.8
salt	374	4.26	7.22
very small (< 5 atoms)	82	0.93	14.63
metal-containing	50	0.57	88.0
charged organic	4	0.05	0.0

Table 3 Table summarizing the steps filtering the considered assays along with the number of remaining assays at each step.

filter step	remaining assays
hormone receptor activity assays in ToxCast	132
available via CompTox Dashboard	127
at least 500 tested compounds	59
DataSail split feasible (cf. Section ‘Data splitting’)	52

(four folds), we conducted a nested 4-fold CV to evaluate hyperparameter combinations. Model selection was based on the highest mean Matthews correlation coefficient (MCC) across the inner folds. The selected model was subsequently retrained on the full outer training set and evaluated on the corresponding test fold (cf. Fig. 2).

We used DataSail (Version 1.2.3) [5] to perform a distance-based split for each assay into five folds, each accounting for approximately 20% of the respective assay dataset. DataSail formulates the data splitting task as an integer linear program (ILP). To this end, molecules are first clustered based on compound similarity, and the optimization ensures that all compounds within the same cluster are assigned to the same dataset partition. Clustering is performed using the DataSAIL’s default parameters, i.e., Tanimoto similarity computed on Morgan fingerprints [20]. As DataSail generates ILP-based splits, these splits are deterministic and optimized to satisfy the desired similarity constraints rather than being generated randomly. As a result, the repeated CV strategy proposed by Ash et al. [7] is not directly applicable while preserving the same splitting principles. Moreover, performing multiple CV runs with the high number of tested combinations and hyperparameters would be prohibitively computationally expensive. Thus, using a nested 5-fold CV instead of repeated CV is a compromise that preserves the chemically informed splitting strategy while enabling a broad comparison across methods.

Due to the high class imbalance in the assays, we additionally used the stratification parameter to ensure that the class ratio was similar across all folds. Using the default parameters, the used SCIP solver [21, 22] was only able to split 32 assays. Therefore, we

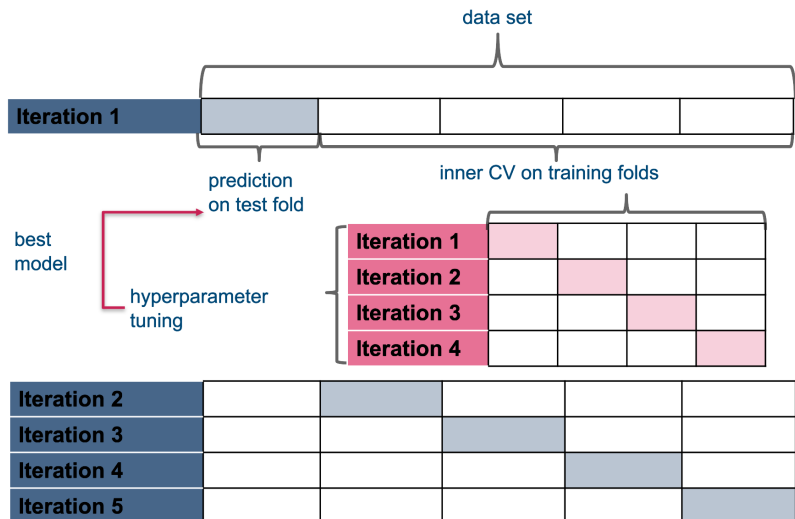


Fig. 2 Illustration of the 5-fold CV with a nested 4-fold CV for hyperparameter tuning.

relaxed the constraints according to the recommendations in the DataSail frequently asked questions by setting the epsilon (controlling the allowed error in dataset size) and delta (controlling the allowed error in the stratification constraint) parameters to 0.3, which allowed us to perform the splitting for 52 of the 59 assays.

2.2 Compound Representations

To systematically evaluate the impact of different molecular encodings on model performance, we compared several commonly used compound representations. Compounds were encoded as fixed-length numerical feature vectors suitable for tabular machine-learning models. All features were calculated from their canonical SMILES representation [17]. In our study, four different representations were evaluated: physicochemical properties, two types of structural fingerprints, and transformer-based embeddings of SMILES.

Physicochemical properties: Physicochemical descriptors comprise mostly scalar molecular features such as molecular weight or logP [16, 23]. We use the 217 RDKit descriptors, which are a fixed set of molecular properties computed from the 2D graph of the molecule, covering constitutional properties (e.g. molecular weight and atom counts), topological indices (e.g. the Hall-Kier alpha index and Chi indices), and physicochemical features such as logP, hydrogen-bond donors/acceptors, and fragment-based scores [24]. Compounds, for which RDKit descriptors could not be fully computed (i.e., containing nan or infinite values), were excluded from the experiments with physicochemical properties as features. As shown in Table 2, some compound classes, particularly metal complexes, frequently cause failures in RDKit descriptor calculations. Excluding these compounds from physicochemical property analyses could introduce bias when comparing physicochemical properties with other compound representations. To assess this effect, we performed additional experiments using only

compounds with valid RDKit descriptors across all representations. Moreover, we evaluated alternative preprocessing strategies that retain all compounds for the physicochemical properties, including median imputation of missing values and removal of the 18 affected descriptors. Lastly, we compared these approaches with compound cleaning using MolPipeline, as described above. These additional experiments were conducted using a random forest model without dimensionality reduction.

Structural fingerprints: Structural fingerprints encode molecular structure mostly as binary vectors (sometimes also count vectors are used). We considered two commonly used methods implemented in the RDKit library: MACCS (Molecular Access System) and Morgan fingerprints. MACCS fingerprints were generated using the 166-bit variant, where each bit corresponds to a predefined SMARTS pattern representing a functional group; bit values indicate presence or absence. Due to the fixed mapping between bits and substructures, MACCS fingerprints offer straightforward interpretability. Morgan fingerprints encode atom-centered neighbourhoods up to a given radius (i.e. bond length). Atom identifiers are iteratively updated to incorporate neighbouring atoms and the resulting identifiers are hashed into a fixed-length bit vector. Morgan fingerprints were generated with RDKit default parameters (2048 bits, radius 3). The initial atom identifiers of the fingerprint version that we use are based on atom connectivity and are, thus, roughly equivalent to the extended-connectivity fingerprints (ECFP)[25, 26]. Following the diameter-based naming convention of ECFP, this configuration corresponds closely to ECFP6[25], which has been shown to be suitable for assessing chemical similarity [27].

Deep learning-based embeddings of SMILES: To obtain informative fixed-length molecular representations that capture structural and chemical relationships beyond handcrafted descriptors, we generated embeddings directly from SMILES strings using transformer-based chemical language models [2, 28–30]. Specifically, we adopted the foundation model of Sultan et al. [2], pre-trained using masked language modeling on 30% of the GuacaMol dataset [31]. Following the authors’ protocol, we performed domain adaptation (DA) separately for each training set using a multi-task regression (MTR) objective. RDKit physicochemical descriptors were required for DA; compounds with missing or infinite values were excluded from this step and embedded only after model adaptation. Encoder outputs were used as tabular molecular representations with a fixed length of 768 bits for downstream models.

2.3 Dimension Reduction Techniques

To reduce the dimensionality of the input feature space, we evaluated both feature selection (FS) and feature extraction (FE) methods. Feature selection retains a subset of the original features, whereas feature extraction transforms the original feature space into a lower-dimensional representation through linear or non-linear combinations of the original features [3, 32]. We considered three FS methods and one FE method. All dimension reduction methods were fitted exclusively on the training data within each data split and subsequently applied to the corresponding test data to prevent data leakage [3]. To ensure a meaningful comparison, each dimension reduction method was evaluated using a protocol consistent with its underlying assumptions and recommended application. Rather than imposing a uniform preprocessing strategy

across all methods, which may advantage some methods while disadvantaging others, each method was assessed under conditions that preserve the statistical quantity on which it is based. The specific preprocessing protocol and its rationale are described in the corresponding method sections.

Mutual Information: Mutual information (MI) quantifies the statistical dependency between a feature and the target variable, with higher values indicating that the feature provides more information about the class labels [33]. MI is invariant to invertible and differentiable transformations in the feature space [34]. Thus, no scaling or standardization is required. We use the `mutual_info_classif` function from scikit-learn to estimate the MI between the features and the response. Based on these scores the features were ranked and the top 100 were used for model training.

Variance: Variance-based feature selection assumes that features with greater variability across samples are more likely to capture meaningful differences between classes. Since variance is sensitive to feature scale, we first applied min-max normalization to ensure that all features lie within the same range [35]. We then computed the variance of each feature on the normalized training set and selected the 100 features with the highest variance.

Minimum Redundancy Maximum Relevance: We performed Minimum Redundancy Maximum Relevance (MRMR) feature selection using the `mrselection` Python package (version 0.2.8) with its default settings. Features were selected using a greedy forward-selection procedure that maximizes relevance to the target variable while minimizing redundancy among the selected features [3]. Relevance was quantified using the Analysis of Variance (ANOVA) F-statistic, and redundancy was measured using the Pearson correlation coefficient (PCC). The ANOVA F-statistic is defined as the ratio of between-group to within-group variance [36]. Because both quantities are affected equally by linear transformations, the F-statistic is invariant to scaling and standardization. Similarly, the Pearson correlation coefficient (PCC) is invariant to linear transformations [37]. Consequently, neither ANOVA nor PCC require prior scaling or standardization, and no preprocessing was applied before MRMR.

Principal Component Analysis: Principal component analysis (PCA) was used as the FE method. PCA transforms the original features into a new set of orthogonal components, which are linear combinations of the input features [38]. As PCA is sensitive to features with different scales, standardization is widely considered necessary to prevent variables with large variances from dominating the principal components [39]. Thus, we applied a standardization using the scikit-learn `StandardScaler` with default parameters [40]. The first 100 principal components were computed using scikit-learn and used as input features for downstream models.

2.4 Machine Learning Models

ML is commonly divided into four categories: supervised, unsupervised, semi-supervised, and reinforcement learning [41]. Since toxic endpoint prediction is primarily addressed using supervised learning [42], we restrict our benchmarking pipeline to five supervised algorithms designed for tabular data. As it is often claimed that graph neural networks (GNNs) outperform classical models on tabular data, we compared in addition the best combination of our benchmarking with the recently published

CheMeleon model [43]. All models and implementation details are introduced in the following.

Support Vector Machine: We evaluated a support vector machine (SVM), a supervised classification algorithm that separates samples in the feature space using a hyperplane [38]. Since the SVM algorithm relies on distances in the feature space, model performance can be affected when features are measured on different scales. Therefore, all input features were standardized prior to model training, following the recommendation of Wiltos [44]. We used the scikit-learn implementation of the SVM and tuned three hyperparameters (Section 2.1): the kernel function (`kernel`), the kernel coefficient (`gamma`), and the regularization parameter (`C`). Because `gamma` is only applicable to non-linear kernels, the hyperparameter grid comprised 49 combinations (Table 4). To account for class imbalance, the `class_weight` parameter was fixed to ‘balanced’, which assigns class weights that are inversely proportional to the class frequencies. Additionally, `probability` was set to `True` to enable probability estimates, which are required for performance metrics such as the area under the receiver operating characteristic curve (AUROC). Appendix Table 1 lists all remaining hyperparameters that were kept at their default or fixed values.

Random Forest: A random forest (RF) is an ensemble method that consists of many decision trees that are combined to enhance robustness and predictive power [45]. For the RF, we also used the scikit-learn library. The number of trees (`n_estimators`), the maximum depth of the trees (`depth`), and the minimum number of samples per leaf (`min_samples_leaf`, cf. Table 4 for the tested ranges). In total 48 hyperparameter combinations were tested (cf. Section 2.1 for details on the hyperparameter tuning). For the RF, we also fixed the `class_weight` to ‘balanced’. As RFs have been shown to be robust regardless of the feature scaling, we did not scale the features prior to RF training [46].

CatBoost: We evaluated CatBoost (for ‘Categorical Boosting’), a gradient boosting algorithm based on decision trees [47]. In contrast to RFs, CatBoost builds trees sequentially, with each tree correcting the errors of its predecessors [48]. The algorithm natively supports categorical features through efficient encoding strategies, while also being well suited for datasets containing only continuous features, such as physico-chemical descriptors [49]. As for RF, CatBoost performance is not affected by scaling or not scaling the input features [46]. We used CatBoost version 1.2.8 and tuned the hyperparameters according to the official CatBoost recommendations. Specifically, we tuned the L2 leaf regularization (regularization strength), the bootstrap type (sampling strategy), and the maximum tree depth (tree complexity) (Table 4). In total, 56 hyperparameter combinations were evaluated.

Multi-layer Perceptron: We evaluated a multi-layer perceptron (MLP), a feed-forward neural network consisting of an input layer, one or more hidden layers, and an output layer, in which each neuron applies a weighted sum of its inputs followed by a (non-)linear activation function [50].

Because MLPs are sensitive to the scale of the input features [38], all features were standardized prior to model training. We employed scikit-learn version 1.6.1 for the MLP implementation.

Owing to the larger number of configurable hyperparameters, the MLP had the most extensive hyperparameter search, comprising 192 combinations. Specifically, we tuned the L2 regularization strength (to reduce overfitting), the activation function (to determine the non-linear transformation applied by each neuron), and the hidden layer dimensions (to control the network capacity, cf. Table 4).

Early stopping was enabled to reduce overfitting, and the adaptive learning rate schedule was used throughout (cf. Appendix Table 1).

TabPFN: TabPFN (for ‘Tabular Prior-data Fitted Network’) is a foundation model designed for tabular datasets of small to medium size [51, 52]. Unlike conventional neural networks, TabPFN relies on a pretrained prior over a large distribution of synthetic tasks rather than being trained from scratch on the target dataset.

We used tabpfm version 8.5.0 without fine-tuning limit. According to Hollmann et al. [51], TabPFN processes raw tabular data because normalization is integrated into its workflow. Therefore, we did not apply feature normalization before training. As the 2048 bit Morgan fingerprint is larger than the 2000 bit limit of TabPFN, we set the `ignore_pretraining_limits` to True so that we could still train models on the Morgan fingerprint without DR. As TabPFN is a foundation model, no hyperparameter tuning was performed. Instead, we conducted standard 5-fold CV, fine-tuning a new TabPFN model on the training folds in each iteration before evaluating it on the corresponding test fold.

CheMeleon: CheMeleon is a directed message passing neural network (D-MPNN) based on the D-MPNN implementation in the chemprop package. In contrast to chemprop, CheMeleon is pretrained on a large corpus of PubChem compounds to predict their Mordred descriptors [43]. To make task-specific predictions it must be fine-tuned on a small labelled dataset. Following the evaluation protocol of Burns et al. [43], we did not conduct hyperparameters optimization. Instead, all experiments were performed using the hyperparameter configurations reported in the original study. Thus, as for TabPFN, we fine-tuned the pre-trained CheMeleon foundation model separately for each training fold in the 5-fold CV, using chemprop version 2.3.1..

Baseline: Traditional toxicity assessment often included a method called *read-across*, which is based on the assumption that similar compounds will show similar activity [53]. Inspired by this traditional approach, we compare the ML models against a baseline, which, for each compound from the test set, ‘predicts’ the label of its nearest neighbour from the training set based on the Tanimoto similarity between the Morgan fingerprints.

2.5 Model Evaluation and Statistical Analysis

We used the MCC to evaluate model performance. Although the area under the receiver-operator-characteristics curve (AUROC) is commonly reported in toxicity prediction models, it is less suitable for highly imbalanced datasets. [54] In contrast, MCC has repeatedly been shown to provide a reliable performance measure for binary classification under class imbalance [55, 56]. MCC is a binary analog of the Pearson correlation coefficient (PCC) and is computed from the confusion matrix, consisting of true positive (TP, predicted class = actual class = 1), false positive (FP, predicted class = 1, actual class = 0), true negative (TN, predicted class = actual class = 0),

Table 4 Hyperparameter ranges for the employed models: SVM, RF, CatBoost², and MLP.

Model	Hyperparameter	Range	#Comb.
SVM [40]	Kernel function (<code>kernel</code>)	linear, poly, rbf, sigmoid	$(3 \cdot 2 + 1) \cdot 7 = 49$
	Kernel coeff. (<code>gamma</code>) ²	scale, auto	
	Regularization (<code>C</code>)	0.001, 0.01, 0.1, 1, 10, 100, 1000	
RF [40]	Number of trees (<code>n_estimators</code>)	100, 300, 500	$3 \cdot 4 \cdot 4 = 48$
	Maximum depth (<code>max_depth</code>)	10, 20, 30, 40	
	Min. samples/leaf (<code>min_samples_leaf</code>)	5, 10, 15, 20	
CatBoost [47]	L2 leaf regularization (<code>l2_leaf_reg</code>)	0.001, 0.01, 0.1, 1	$4 \cdot 2 \cdot 7 = 56$
	Bootstrap type (<code>bootstrap_type</code>)	'Bayesian', 'Bernoulli'	
	Maximum depth (<code>max_depth</code>)	4 – 10	
MLP [40]	Strength of L2 regularization (<code>alpha</code>)	0.00001, 0.0001, 0.001, 0.01, 0.1, 1	$6 \cdot 4 \cdot 8 = 192$
	Activation function (<code>activation_function</code>)	'relu', 'tanh', 'logistic', 'identity'	
	Hidden layer dimensions (<code>hidden_layer_sizes</code>)	(50,), (100,), (50, 50), (100, 50),	
		(100, 100), (50, 50, 50), (100, 100, 100), (100, 100, 50)	

² Only relevant for 'poly', 'rbf', and 'sigmoid' kernels.

and false negative (FN, predicted class = 0, actual class = 1) [55]:

$$\text{MCC} = \frac{\text{TP} \cdot \text{TN} - \text{FP} \cdot \text{FN}}{\sqrt{(\text{TP} + \text{FP}) \cdot (\text{TP} + \text{FN}) \cdot (\text{TN} + \text{FP}) \cdot (\text{TN} + \text{FN})}} \quad (1)$$

Because the MCC is not additive across cross-validation folds, we did not compute arithmetic means or medians directly. Instead, MCC values were first transformed using Fisher's transformation, yielding approximately normally distributed values [57]. Mean and median values were calculated on the transformed scale and subsequently back-transformed using the inverse Fisher transformation. Throughout this work, these values are referred to as the adjusted mean and adjusted median MCC.

To compare model performance statistically, we employed the Friedman test with the Iman–Davenport correction, followed by the Nemenyi post-hoc test [6, 7, 58, 59]. This non-parametric framework is recommended for comparing multiple machine learning algorithms across datasets and does not require assumptions such as normality or homoscedasticity. The Friedman test ranks the competing methods separately for each dataset (here: assays or cross-validation folds) and tests whether their average ranks differ significantly. The Iman–Davenport correction provides a more powerful test statistic than the original Friedman test [59]. When the global null hypothesis was rejected, pairwise differences were assessed using the Nemenyi test, which compares average ranks while controlling the family-wise error rate. Because the analysis is rank-based, statistically significant differences may occur even when absolute MCC values differ only marginally. When comparing only two models, the Wilcoxon signed-rank test is recommended [6]. This non-parametric alternative to the paired t-test accounts for the paired nature of the observations and does not assume normally distributed performance differences. Similar to the Friedman–Nemenyi test, it is rank-based and compares paired samples, i.e., model performances obtained on the same CV folds.

To compare individual pipeline components (compound representations, dimensionality reduction methods, and learning algorithms), we grouped all evaluated pipelines according to the component of interest and computed the adjusted mean MCC for each assay across the remaining pipeline configurations. These adjusted mean MCC values were then ranked per assay and compared using the Friedman–Nemenyi procedure. As recommended by Demšar [6], we use a two-sided Wilcoxon signed rank test, when performing comparisons between two classifiers, i.e., when comparing the best ML model with the baseline and CheMeleon, respectively.

Lastly, we analyse the dataset characteristics that influence model performance. In particular, we focus on (i) class imbalance, (ii) local structure-activity consistency around active compounds (the minority class) measured via Tanimoto similarity on the Morgan fingerprints and (iii) the detection of *activity cliffs*. Roughly speaking, activity cliffs are similar compounds that do not show similar activity. Usually activity cliffs are defined via the ratio of difference in activity to the distance between the compounds in a chemical space [60]. While activity cliffs are traditionally defined using continuous potency measurements, the assays considered in this study provide only binary activity labels. Thus, we define an activity cliff if a compound shows a discordant label with its partners in a matched molecular pair (MMP). MMPs are calculated using the `rdkit.Chem.rdMMPA`. We tested several values for the number of heavy atoms and the ratio of the core size that define the maximum size of the swapped substituent. In particular, we tested 6, 8, and 10 as maximum number of heavy atoms and 0.2, 0.5, and 0.8 as maximum fraction of the core size (cf. Appendix Table 3). As a balance between the two values, we show results for ≤ 8 heavy atoms and 0.5 times the core size.

3 Results

The aim of this study is to benchmark ML pipelines for toxicity prediction with a particular focus on hormone receptor activity. First, we will investigate the general performance of the models. To be comparable to the literature, we will report the achieved AUROC values. However, as discussed above, we will mainly focus on the Matthews correlation coefficient (MCC), as it accounts for the class imbalance.

To investigate which particular design choices influence the performance, we systematically compare different compound representations, dimensionality reduction (DR) methods, and ML models across the evaluated assays. We first present the results for the individual pipeline components, followed by an analysis of their combinations to identify the best-performing pipelines. The best-performing pipeline will then be compared against the D-MPNN model CheMeleon to investigate whether graph-based models can outperform the rather simple tabular models. Moreover, we analyse the influence of preprocessing strategies and compound removal on the model performance. Finally, we investigate how dataset characteristics, including the number of available samples, class balance, and the chemical diversity of the tested compounds, relate to predictive performance.

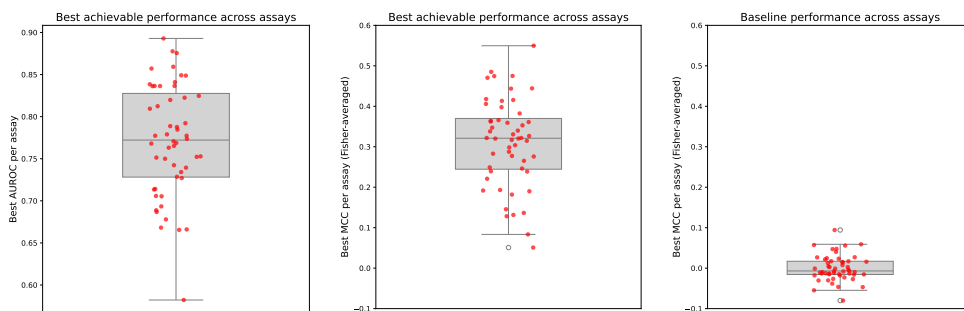


Fig. 3 Boxplot showing the best achieved performance for each assay in terms of AUROC (left) and MCC (middle). The right plot shows the MCC performance of the baseline model that always predicts the label of the closest neighbour in the training fold.

3.1 Toxic Endpoint Prediction Performance Revisited

Several ML methods have been proposed for toxicity prediction, ranging from more classical methods including RFs, SVMs, and gradient boosting approaches to deep neural networks [53, 61–64]. Reported performances on benchmark datasets often imply that toxicity endpoints can be predicted reasonably well, with AUROC values frequently exceeding 0.7 or 0.8 for selected endpoints. However, these studies typically differ in evaluation protocols, molecular representations, and train–test splitting strategies, making direct comparisons difficult, and are to a vast extent missing strategic evaluation of significant improvements over baseline performance. Moreover, AUROC can provide an overly optimistic values on highly imbalanced datasets, as it does not directly account for the correct classification of the minority class. Nevertheless, it remains one of the most commonly reported metrics for toxicity prediction models and often no other metric, which is more suitable for imbalanced datasets is reported.

For reference, we therefore also investigated AUROC performance on the 52 hormone receptor activity data sets investigated in this study, using the DataSail split. DataSail [20] generates chemically dissimilar train-test partitions and therefore represents a substantially more challenging evaluation setting than conventional random splits. As a consequence, predictive performance is typically lower than that observed in random split experiments, but the resulting estimates better reflect a model’s ability to generalize to previously unseen chemical space and are therefore more indicative of real-world applicability. Fig. 3 shows the best achieved AUROC values for each assay across all evaluated combinations. Despite the increased difficulty of this splitting strategy, the achieved AUROC values range from 0.6 to 0.9 and are therefore within the range reported for several toxicity prediction tasks [62, 63].

Importantly, the corresponding MCC values remained substantially lower. When considering the best model per endpoint, only moderate overall MCC values are observed ranging from less than 0.1 to barely above 0.5, with a median MCC value of 0.32 across all 52 assays. This discrepancy illustrates why relying solely on AUROC may lead to overly favourable conclusions when evaluating highly imbalanced toxicity datasets. Consequently, the following analyses will be based on MCC. Comparing the

MCC values of the nearest-neighbour baseline, which can be viewed as a simple read-across strategy assigning the label of the most similar compound [53], to those of the ML models shows that the evaluated approaches consistently outperform this baseline, which achieves a median MCC of around 0 (right side of Fig. 3). We performed a two-sided Wilcoxon signed rank test to compare the ML models with the baselines, and the ML models significantly outperform this baseline ($p\text{-value} \approx 3.5 \cdot 10^{-10}$). The baseline exploits a strong prior assumption commonly used in toxicology, namely that structurally similar compounds tend to exhibit similar biological and toxicological properties [53, 65, 66]. The fact that ML models are able to improve upon this nearest-neighbour strategy demonstrates that they capture some additional structure–activity relationships beyond direct chemical similarity. At the same time, the median MCC of 0.32 for the ML models is still modest, which highlights the importance of understanding the factors that constrain predictive performance. We therefore look into the individual components of the modeling aspect and how they influence the performance (see Sections 3.2-3.5); as well as the dataset characteristics, and examine how properties such as chemical diversity, class imbalance, and distribution of active compounds influence model performance (see Section 3.7).

3.2 Compound Representations

The molecular representation constitutes the foundation of any toxicity prediction pipeline, as it determines which chemical information is available to the learning algorithm [54]. Several studies have suggested that transformer-based embeddings can rival or even outperform classical descriptors and fingerprints on several molecular property prediction tasks, while other studies claimed the opposite [2, 28]. Therefore, we first investigate which type of molecular representation is most informative for predicting hormone receptor activity across the evaluated ToxCast assays and ML models.

To compare molecular representations across the 52 assays, we require a performance measure that is robust to differences in endpoint difficulty. Directly averaging MCC values across assays would allow particularly easy or difficult assays to dominate the comparison. Therefore, for each assay, we first determine the relative ranking of the four representations based on their adjusted mean MCC across all evaluated model and DR combinations. We then compare these assay-specific ranks using the Nemenyi test. This analysis allows us to assess whether one representation consistently outperforms the others across endpoints, rather than merely achieving a higher average MCC due to a small number of favourable assays.

The results of the Nemenyi test are shown in Fig. 4. The critical difference plot (left) shows the average rank of the representations across all these tests as well as the average of all fisher-transformed MCC values back transformed to the original scale. If a group of representations is connected by a vertical line it means that there is no significant difference in their average rank according to the Nemenyi test, i.e., the difference in ranks did not exceed the critical difference. The p -values for each pairwise comparison can be found on the right side of Fig. 4.

The results of this analysis reveal that physicochemical properties achieve the best average rank of 1.63. However, they only significantly outperform the other two fingerprint representations and not the learned embeddings. MACCS fingerprints and

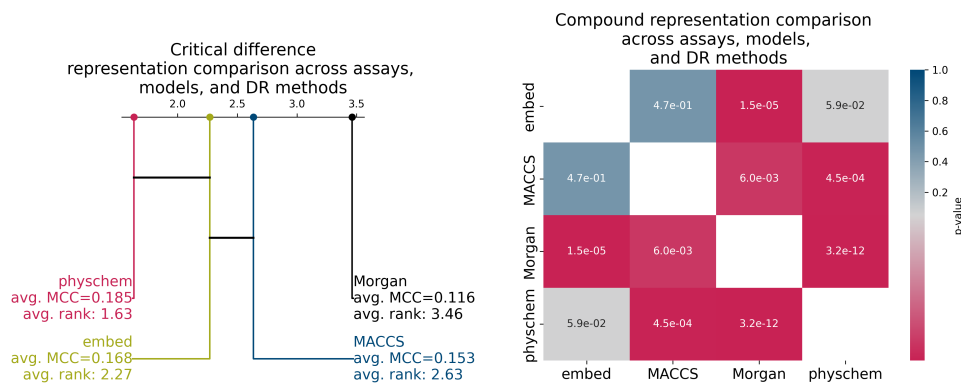


Fig. 4 Results of the Nemenyi test performed to compare compound representations across all DR method-model combinations for all tested assays. Left: Critical difference plot of the compound representations; if there is no significant difference between the average rank of different compound representations, they will be connected by a vertical line. Right: P-value heat map; a red p-value indicates that the difference was significant with significance level of 0.05. Abbreviations: ‘physico-chemical properties’ with ‘physchem’, ‘embeddings’ with ‘embed’.

embeddings from a pretrained DL model both show ranks close to 2.5 and are not significantly different, while Morgan fingerprints with an average rank of 3.46 perform significantly worse than all other representations. To further investigate these differences, we ranked each of the 100 tested combinations for each of the 52 assays. Then, we counted how often each of the descriptors was in the top 1, top 3, and top 5 of all combinations (see Fig. 5). Notably, since we are now considering combinations, for the top 3 and top 5, the total amount of considered data points is $3 \cdot 52 = 156$ and $5 \cdot 52 = 260$, respectively. In particular, the descriptors can occur in the ranking multiple times in each assay, i.e., if the descriptor was particularly informative, it might occur 5 times in the top 5 for one assay with different combinations of model and DR method. We report the fraction of top k combinations, in which the respective representation was used. We can observe that physicochemical properties are part of the best combination in half (26/52) of the tested assays. A similar pattern can also be observed for the top 3 and top 5 combinations, where the physicochemical properties are involved 86 (55%) and 138 (53%) times, respectively. In numbers, the embeddings and MACCS keys show only slight differences, while Morgan fingerprints are the least often involved in the top performing combinations, which aligns with the result of the Nemenyi test. Interestingly, however, although the Nemenyi test did not find a significant difference between physicochemical properties and embedding, the physicochemical properties are considerably more often involved in the top performing combinations.

This is consistent with the findings of Sultan et al. [2], who showed that chemically and physically-informed compound representations yield better performances across model types and different ADME (Absorption, Distribution, Metabolism, and Excretion) datasets. Yet, we have to acknowledge that the worse performance of Morgan fingerprints may be partially explained by our data splitting strategy: We use the default parameters of DataSail, which means that the datasets are constructed so that

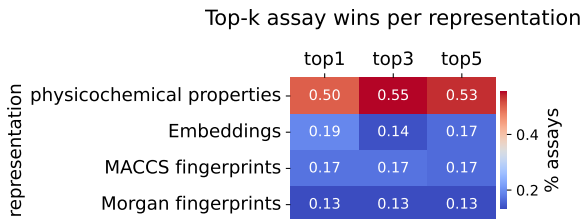


Fig. 5 Compound representations along with the fraction (rounded to two decimal places) of combinations that include the respective representations and were among the top k ($k \in \{1, 3, 5\}$) approaches across the 52 tested assays.

the dissimilarity of the Morgan fingerprints between the folds is maximized. Potentially, this makes it particularly difficult for the models to predict using this descriptor (cf. Section 4.4 for a more detailed discussion).

Interestingly, the observed ranking differences are larger than the corresponding differences in absolute predictive performance. While physicochemical descriptors show an advantage in terms of rank consistency across assays, the adjusted mean MCC values remain relatively close, ranging from 0.19 for physicochemical properties to 0.12 for Morgan fingerprints. This suggests that the primary benefit of physicochemical descriptors lies in their robust performance across diverse endpoints rather than in large performance gains for individual assays. Their superiority might further indicate that global molecular properties remain highly informative for the evaluated toxicity endpoints, potentially because characteristics such as molecular weight, and or charge may capture biologically relevant aspects of hormone receptor interactions more directly than purely structural representations.

3.3 Dimension Reduction

Many molecular representations contain hundreds or even thousands of features, potentially introducing redundancy, noise, and an increased risk of overfitting [3, 54, 67]. Dimensionality reduction (DR) is therefore frequently applied in cheminformatics and toxicity prediction workflows to improve model generalization, reduce computational complexity, and enhance interpretability. However, it remains unclear whether reducing the feature space is beneficial for hormone receptor activity prediction across diverse toxicity endpoints. Therefore, we investigate the impact of different DR strategies on predictive performance.

As for the compound representations, we use the same rank-based comparison strategy. For each assay, we aggregate the performance of all pipelines employing the same DR method, assign ranks based on the adjusted mean MCC values, and compare these ranks using the Nemenyi test. This allows us to evaluate whether a DR strategy consistently improves performance across endpoints (see Fig. 6).

We observe that applying no DR method has the best average rank (1.90); however it is not significantly different from MRMR and variance-based FS. Moreover, among the DR approaches, there is no significant difference between MRMR, variance-based, and mutual information-based feature selection, with average rank values of 2.42, 2.62, and 3.21, respectively. In contrast, PCA exhibits the worst average rank and is

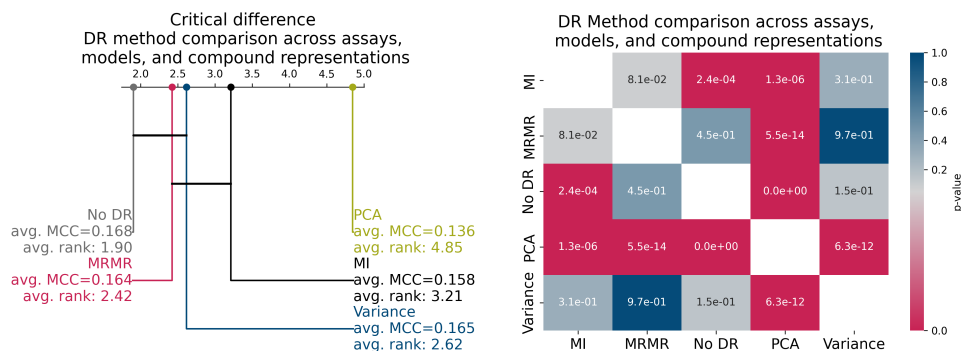


Fig. 6 Results of the Nemenyi test performed to compare DR methods across all compound representation-model combinations for all tested assays. Left: critical difference plot; if there is no significant difference between the average rank of different DR methods, they will be connected by a vertical line. Right: p-value heat map; a red p-value indicates that the difference was significant with significance level of 0.05.

significantly outperformed by all alternative approaches. Overall, none of the evaluated DR methods consistently improved predictive performance. However, our analysis is restricted to a latent dimensionality of 100, and the relative effectiveness of DR may depend on both the molecular representation and the chosen target dimensionality. Therefore, while omitting DR yielded the strongest performance in our benchmark, these results should not be interpreted as evidence that DR is generally detrimental. Given that DR can substantially reduce training time and memory requirements, applying an appropriate DR method may still be advantageous in practice, particularly for computationally demanding settings.

Interestingly, the ranking differences are again more pronounced than the corresponding differences in absolute MCC values. The advantage of using the full feature set suggests that the evaluated models can more effectively exploit the available feature dimensions without substantial overfitting. Moreover, toxicity-relevant information appears to be distributed across many features, such that feature reduction may remove complementary predictive signals. Notably, we always reduce the dimensionality of the features to 100 descriptors independent on the original feature space. This was done because otherwise the computational cost of optimizing the number of descriptors for dimensionality reduction would not be feasible. Consequently, our analysis cannot exclude the possibility that alternative dimensionalities would lead to improved performance for specific dimensionality reduction methods. In particular, the benefit of applying DR methods might even differ between different compound representations. Therefore, we repeated the same analysis as above, only considering combinations with physicochemical properties, as they were the best performing representation (cf. Appendix Fig. 1).

We observe that for this subset of combinations, there is no significant difference between mutual information (MI)-based FS and no DR. Moreover, there is no significant difference between MRMR, MI-, and variance-based FS. Thus, the choice of DR method does not appear to have a large influence for the models trained on physicochemical properties.

3.4 Machine Learning Models

We next compare the evaluated classifiers across all compound representations and DR strategies to identify the most performant approach for hormone receptor activity prediction. Following the analysis strategy used in the previous sections, we aggregate the performance of all pipelines employing the same ML model, derive assay-specific ranks from the adjusted mean MCC values, and compare these ranks using the Nemenyi test (see Fig. 7). RF achieves the best average rank (1.35) and significantly outperforms all other evaluated models. In addition to its superior ranking, RF also achieves the highest adjusted mean MCC of 0.2. SVM and CatBoost form a second performance tier, while MLP and TabPFN show the weakest performance overall. Although all models achieve relatively modest average MCC values, RF consistently obtains the best ranks across assays, indicating robust performance across diverse endpoints. Combined with its high interpretability and low computational cost, these results suggest that classical tree-based ensemble methods remain a strong choice for toxicity prediction.

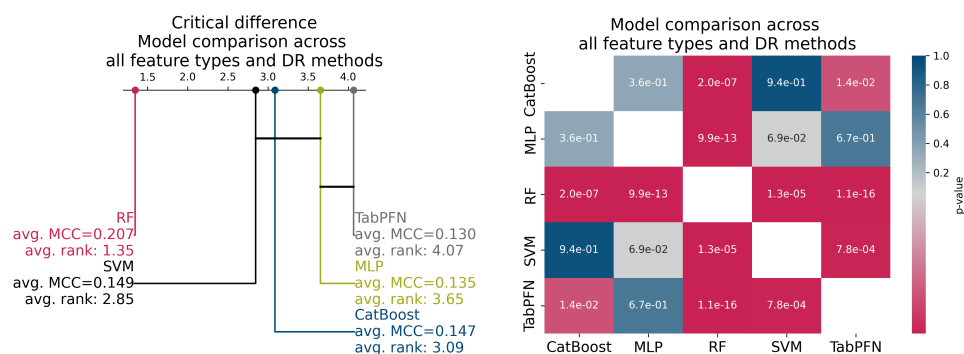


Fig. 7 Results of the Nemenyi test performed to compare models across all compound representation-DR method combinations for all tested assays. Left: Critical difference plot of the model. Right: p-value heatmap; a red p-value indicates that the difference was significant with significance level of 0.05.

One potential explanation of the particular weak performance of the deep learning models (MLP and TabPFN), is the low dataset size. As NNs are complex models with many parameters, they tend to overfit; especially, if the training dataset is too small. Therefore, we investigate which model was in the best performing combination for each assay in relation to the performance and the assay size (cf. Fig. 8). Surprisingly, for none of the assays with a high number of tested samples a deep learning model was performing best. For 8 out of 10 largest assays the RF was the best model and the SVM for the remaining 2. MLP and TabPFN were each only the best model in two assays with less than 4000 samples. In the two assays, where TabPFN performs best, it also achieves decent performance values with MCC above 0.45. Notably, there seems also not to be an obvious correspondence between the assay size and the achieved performance. This is counter-intuitive, as one would usually expect that models trained on larger datasets would also perform better. A more in depth-analysis of the dataset

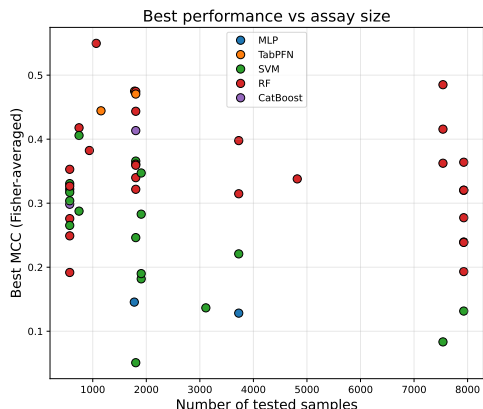


Fig. 8 Scatter plot showing best achieved performance and assay size. The individual points are coloured by the model that was used in the best combination for this assay.

characteristics that show a correlation to the performance are shown below (cf. Section 3.7).

3.5 Full Pipeline Evaluation

Our previous findings indicate that physicochemical properties, no DR, and RF, each outperform all tested alternatives for their individual pipeline part. However, this does not necessarily mean that they perform well in combination. Therefore, we now analyse for each of the 100 tested combination in how many assays it was among the top performing approaches. To this end, we again take the adjusted mean MCC for each combination and assay and rank the combinations. Then, we count for each combination how often it was among the top 1, top 3, and top 5 approaches. The results of this analysis are shown in Fig. 9. Notably, we only show combinations with at least 2 top 1 wins, as otherwise the plot would be too large. The RF trained on physicochemical properties without DR is the best combination in 12 out of 52 assays. Moreover, it ranks among the top 5 in 26 assays. This aligns with our previous findings that each of the individual components is the best in their respective category. The combination with the second highest number of top 1 rankings, i.e., RF on Morgan fingerprints without DR, is only ranked among the top 5 in 8 assays, indicating that it does not have a consistently strong performance. However, the plot reveals that the RF trained on physicochemical properties with MI feature selection only shows the best performance in 2 assays but considering the top 5 rankings, it performs competitive with 22 assays.

Our analyses show that the combination of RF with physicochemical properties and no DR is often among the best performing combinations. Yet, it was not consistently the best performing method across all assays. Consequently, it can be seen as a strong baseline and starting point for future method development. However, other modeling choices should still be considered and tailored to the specific application.

Given that RF models trained on physicochemical properties without DR achieved the strongest overall performance, we further compared this configuration with the

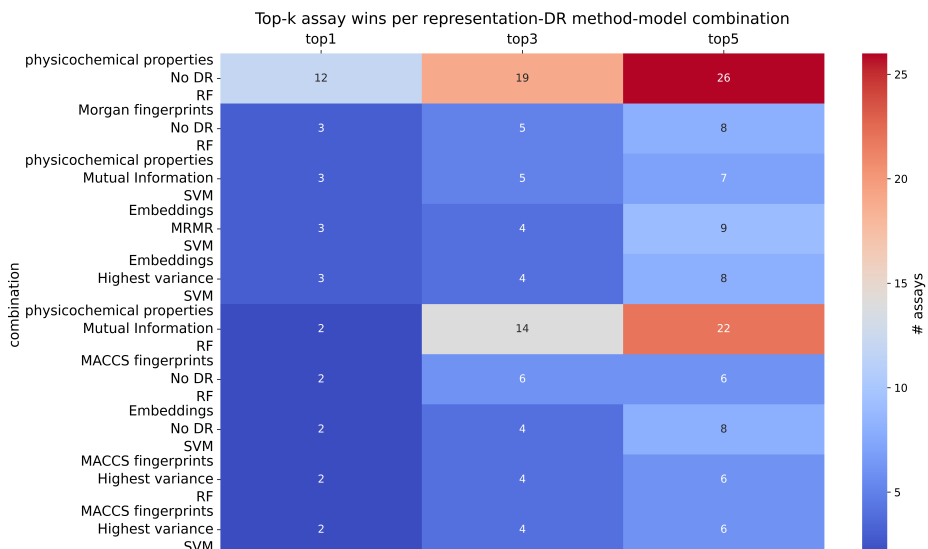


Fig. 9 Combinations along with the number of assays, for which the were among the top k ($k \in \{1, 3, 5\}$) approaches. Notably, we only show combinations with at least 2 top 1 wins.

D-MPNN foundation model CheMeleon. Statistical significance was assessed using a two-sided Wilcoxon signed-rank test (Fig. 10). We found that the RF model significantly outperformed CheMeleon across the evaluated assays, achieving a median MCC of 0.26 compared with 0.01 for CheMeleon. This finding highlights that, despite recent advances in foundation models for molecular machine learning, comparatively simple descriptor-based approaches can still provide strong predictive performance on ToxCast hormone-related prediction tasks. While CheMeleon was fine-tuned following the protocol of Burns et al. [43], future work could investigate whether dataset-specific hyperparameter optimization improves its performance on this benchmark.

3.6 Influence of Preprocessing on Model Performance

One potential concern is that the superior performance of physicochemical descriptors may be partly influenced by differences in compound handling across representations. As described in Section 2.2, compounds for which RDKit could not compute all descriptors were excluded from the corresponding analyses, whereas they remained available for the other molecular representations. To assess whether this discrepancy introduced a systematic bias, we examined the relationship between the percentage of excluded compounds and the adjusted mean performance difference between physicochemical descriptors and the other representations (Fig. 11). No statistically significant association was observed (Pearson correlation (PCC): 0.105, p-value: 0.457; Spearman correlation (SCC): 0.130, p-value: 0.358), indicating that models on assays with more excluded compounds did not systematically show larger performance advantages for physicochemical descriptors. Thus, we found no strong evidence that the exclusion of

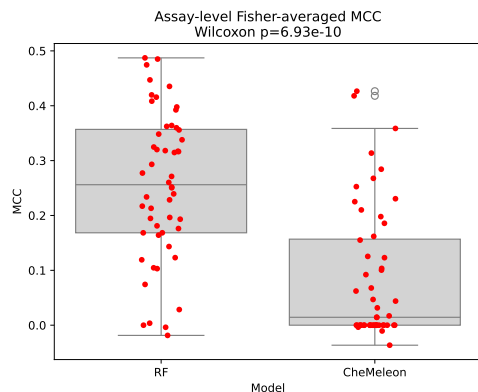


Fig. 10 Performance of RF trained on physicochemical properties and no DR versus CheMeleon across all assays.

compounds substantially influenced the observed ranking of molecular representations. Nevertheless, to further investigate the robustness of our findings, we next evaluate the influence of different preprocessing and compound curation strategies on model performance.

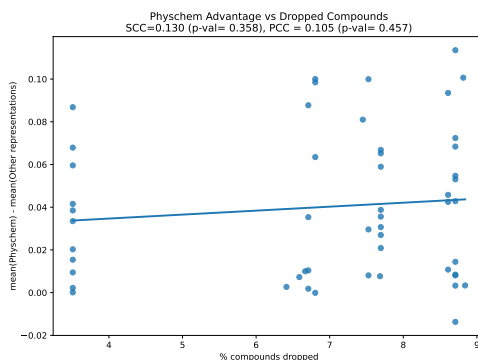


Fig. 11 Percentage of dropped compounds for the physicochemical properties experiments along with the observed adjusted mean performance difference physicochemical properties and the other representations in terms of MCC.

The analyses presented above were conducted on the original ToxCast structures without additional preprocessing (see ToxCast paper [18]), such as salt removal or filtering of metal-containing compounds. Because such curation steps are common in cheminformatics workflows, we investigated their impact by comparing three scenarios: (1) using all compounds, (2) using all compounds except those excluded from the RDKit descriptor experiments, and (3) using all compounds after cleaning with MolPipeline [19]. This analysis allows us to assess whether the observed performance

differences are robust to compound curation and preprocessing decisions. For physicochemical descriptors, scenario (1) is not directly feasible because RDKit could not calculate all descriptor values for some compounds. Therefore, we considered two alternatives: median imputation of missing values or exclusion of the 18 descriptors that contained missing or infinite values. For none of the molecular representations did we

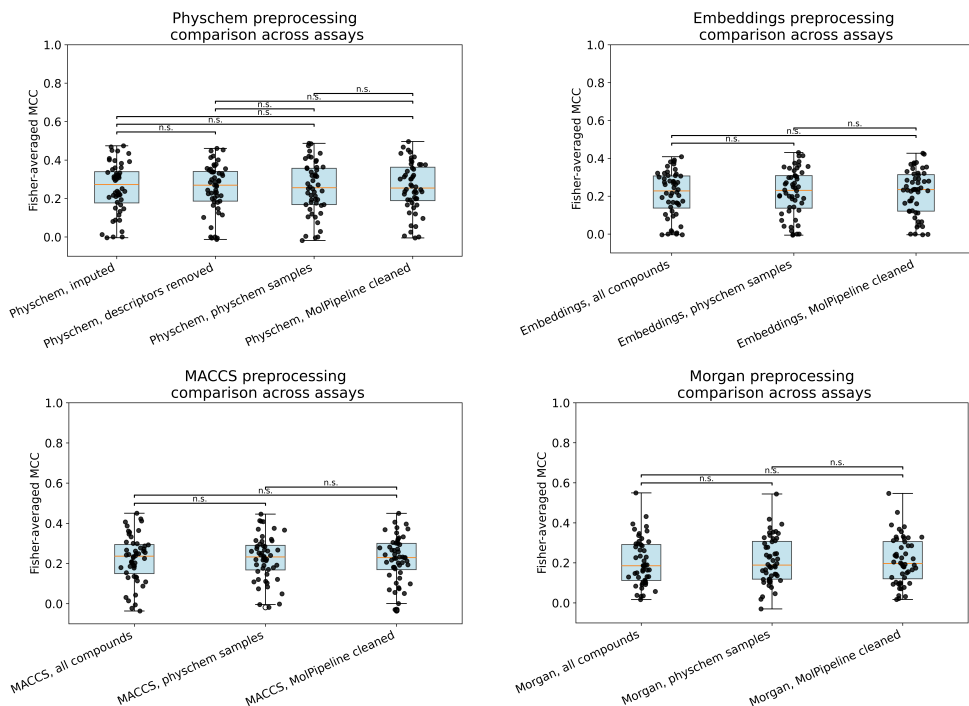


Fig. 12 Boxplots showing the effect of different preprocessing strategies on model performance for the different compound representations. Physicochemical Properties (Physchem): upper left, Embeddings: upper right, MACCS: lower left, Morgan: lower right. The annotation ‘n.s.’ indicates that there is no significant difference between the connected pair of settings using the Friedman test with a Nemenyi post-hoc procedure.

observe a significant difference in model performance between preprocessing strategies (cf. Fig. 12). These results suggest that the choice of preprocessing strategy had only a minor influence on predictive performance in our benchmark and that the exclusion of compounds with missing RDKit descriptors did not introduce a substantial systematic bias.

3.7 Influence of Dataset Characteristics on Performance

As mentioned in Section 3.1, the overall model performance in terms of MCC was modest with average values around 0.3 (cf. Fig. 3). Notably, there are assays, where the models achieved MCC values above 0.5, while for other assays the maximum MCC

was below 0.1. Therefore, we investigated the characteristics of the different assay datasets to better understand the large performance gaps.

Since dataset size alone showed only weak relationship with MCC performance (cf. Section 3.4), we investigated three additional dataset characteristics that may help explain this variation (computation details in methods section 2): (i) class imbalance, (ii) local structure-activity consistency around active compounds (the minority class) measured via whole-molecule fingerprint similarity, and (iii) structure-activity consistency measured using matched molecular pairs (MMPs). For local neighbourhood (ii), we calculate the fraction of active compounds that have a concordant label with their nearest neighbour. Moreover, we average the fraction of label concordance of the 5 nearest neighbours across all active compounds in the assay dataset. As a complementary measure based on chemically interpretable structural changes, we identified MMP analogs for active compounds and computed the fraction whose analog carried the opposite label (MMP cliff fraction). The results of these analyses are shown in Fig. 13, additional dataset characteristics, e.g., dataset size, are shown in the Appendix.

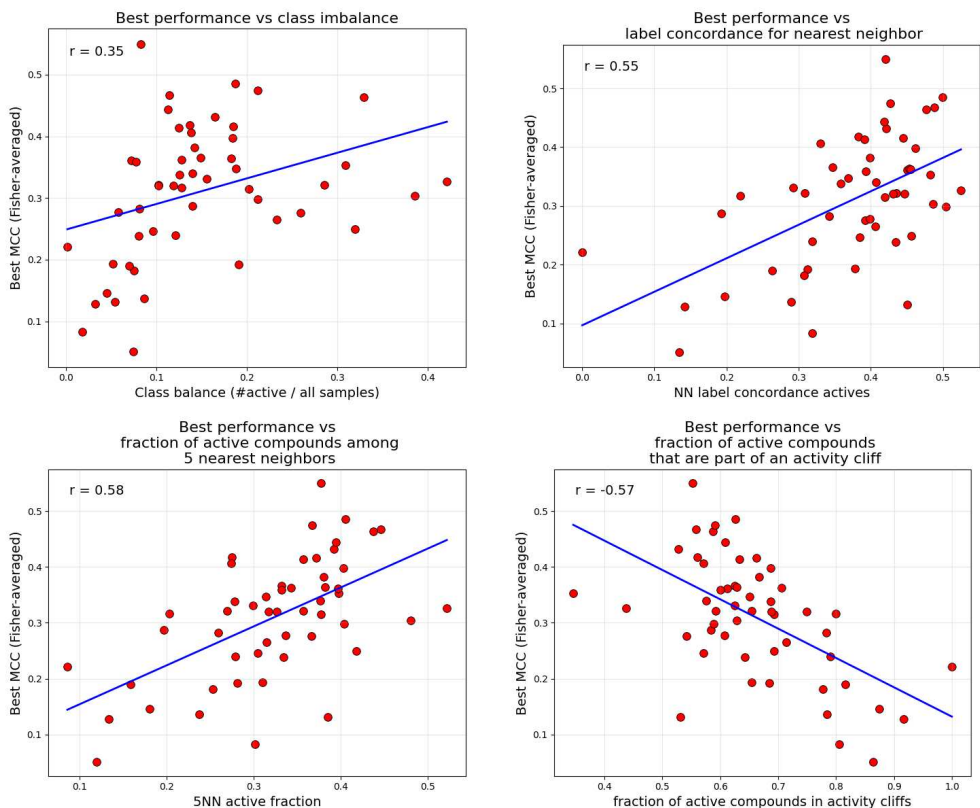


Fig. 13 Relationship between best achieved performance and assay size (upper left), class imbalance (upper right), or compound similarity (lower row).

All four characteristics showed a relationship with predictive performance. While class imbalance exhibited only a moderate correlation with MCC ($\rho = 0.35$), the three measures related to local structure-activity consistency showed substantially stronger associations with model performance. Considering only active compounds, i.e. the minority class in all assays, the fraction of nearest neighbours sharing the same label correlated positively with MCC ($\rho = 0.55$), as did the fraction of active compounds among the five nearest neighbours ($\rho = 0.58$). Conversely, the MMP cliff fraction correlated negatively with MCC ($\rho = -0.57$), i.e., more activity cliffs in the data set make predictions on this dataset harder, thus, resulting in lower performance. Notably, in Fig. 13, we use ≤ 8 heavy atoms and 0.5 times the size of the shared core as hyperparameters for our MMP definition to provide a good balance between the two. However, in Appendix Table 3, we provide the observed correlations for the other tested hyperparameter combinations. The correlations for all tested values lie between -0.32 and -0.58. Therefore, our conclusion that there is a negative correlation between the presence of activity cliffs and model performance did not change.

The consistency of these correlations across both fingerprint-based and MMP-based similarity measures suggests that predictive performance is more strongly associated with local structure-activity relationships of active compounds than with overall dataset composition. Assays in which minority class compounds have many analogs with different labels tend to be harder to predict. Taken together, these results indicate that activity cliffs represent one of the primary challenges across the assay collection, thus, the challenge is rooted in the data itself, rather than in the model architecture. Our results align with the findings of Van Tilborg et al. [68], who showed that the presence of activity cliffs poses a particular challenge for ML models in various bioactivity datasets. Notably, they took three different definitions of activity cliffs into account, where one is based on Tanimoto similarity on Morgan fingerprints, corresponding to our analysis of structure-activity consistency around active compounds. Interestingly, they also demonstrate that classical ML models trained on simple molecular descriptors outperform deep learning models in the presence of activity cliffs across the tested datasets.

To further investigate our observation, we take a closer look at the compounds in the two assay datasets with the worst performance (cf. Fig. 14), i.e., OT_AR_ARELUC_AG_1440 (MCC: 0.05), and TOX21_ERb_BLA_Agonist_ratio (MCC: 0.08), and with the best performance (cf. Fig. 15), i.e., NVS_NR_bER (MCC: 0.55), and TOX21_AR_LUC_MDAKB2_Antagonist_0.5nM_R1881 (MCC: 0.49) label.

We observe that for the worst performing assays, there are very similar compounds – sometimes only distinguishable by one stereoisomer – that show different activity, illustrating the difficulty of the task. Interestingly, compound DTXSID7020687 in Fig. 14 appears to be tested as a stereoisomeric mixture rather than as a pure form. In this case, the observed activity is dominated by the active isomer. This introduces additional complexity for predictive modeling, as the measured activity does not directly correspond to a single well-defined stereochemical structure.

In other cases, the activity changes despite only a single substitution on an aromatic ring. For example, compounds that differ only by the replacement of a hydroxyl group with bromine, chlorine, or hydrogen at the para position can switch the molecule

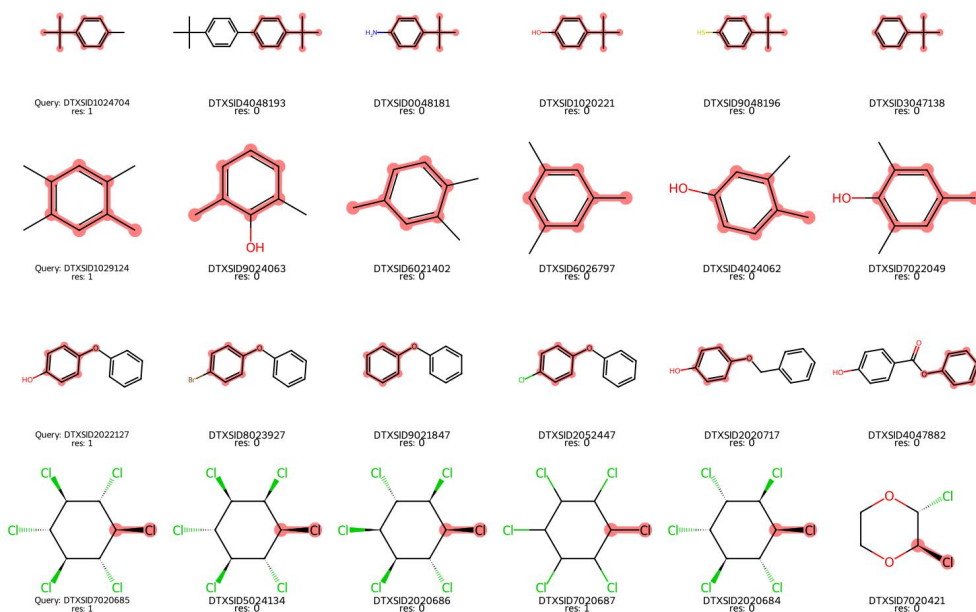


Fig. 14 Example compounds from the two assays with the worst achieved performance: OT_AR_ARELUC_AG_1440 (upper two rows), and TOX21_ERb_BLA_Agonist_ratio (lower two rows). Abbreviations: ‘res’ is the label (1: active, 0: inactive).

from active to inactive for $ER\beta$ agonism, as for the 4-phenoxyphenol in Fig. 14. Such examples highlight that small structural changes may substantially alter the underlying binding mode. Similar observations have been reported for bisphenol analogs that interact with $ER\alpha$, where docking studies showed that differences in hydrogen-bond formation with key receptor residues were important for distinguishing active and inactive compounds despite their highly similar core structures, suggesting that fingerprint-based representations alone may not fully capture the relevant binding determinants [69].

On the contrary, for the best performing assays (Fig. 15), we see that similar compounds often tend to have the same response, which potentially makes the prediction task easier for the models.

4 Benchmark Design and Evaluation Considerations

In this study, we systematically benchmarked compound representations, dimensionality reduction (DR) strategies, and ML models for predicting hormone receptor activity across 52 ToxCast assays. Overall, our results demonstrate that random forests (RFs) consistently outperform all other tested approaches evaluated across all representation-DR combinations. Moreover, across all experiments physicochemical properties are the best compound representation and we show that the application of any tested

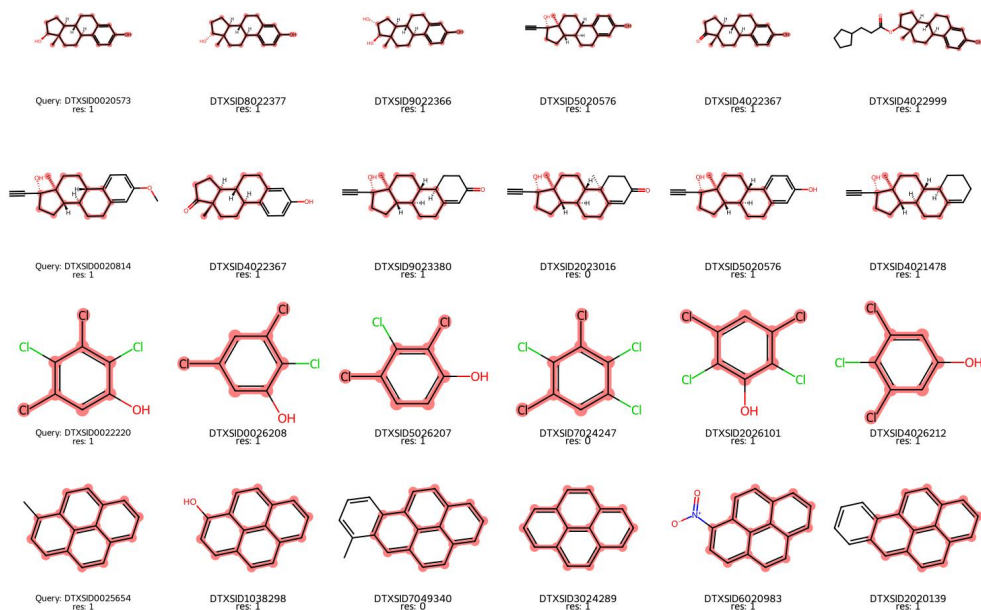


Fig. 15 Example compounds from the two assays with the best achieved performance: NVS_NR_bER (upper two rows), and TOX21_AR_LUC_MDAKB2.Antagonist.0.5nM_R1881 (lower two rows). Abbreviations: ‘res’ is the label (1: active, 0: inactive).

DR method is significantly worse than applying no DR. This reinforces earlier findings that classical, tree-based methods and simple compound representations remain highly competitive for small to medium-sized chemical datasets, even in comparison to more complex approaches.

4.1 Exclusion of Small Assays and Considered Training Schemes

In this study, we focused on relatively large assays comprising at least 500 compounds. Since our objective was to compare single-task models, i.e., one model trained per assay, we defined a minimum dataset size to ensure sufficient data for model development and nested stratified 5-fold cross-validation. This threshold was chosen to provide a robust basis for performance estimation and to reduce the influence of sampling variability, thereby enabling more reliable comparisons across methods.

We acknowledge that many toxicity prediction tasks involve substantially smaller datasets and that learning under data-scarce conditions remains an important challenge in the field. Consequently, a range of approaches, including transfer, few-shot, and multi-task learning, have been proposed to address limited data availability [70]. While our benchmark includes pre-trained foundation models such as TabPFN and CheMeleon, this study was not designed to evaluate methods specifically in low data

regimes, nor to assess the benefits of transfer or few-shot learning. Importantly, our findings on datasets with at least 500 compounds, align with those reported by Snyder et al. [71], who systematically investigated the relationship between dataset size and model performance and found that classical ML approaches generally outperform few-shot and transformer-based methods once datasets contain more than approximately 240 samples. Thus, focusing on assays with at least 500 compounds allowed us to investigate model performance in a regime where sufficient data are available for robust evaluation.

4.2 Dimension Reduction

Overall, our results did not demonstrate a consistent performance benefit from applying a DR method. However, our results should not be interpreted as evidence that DR is generally inferior to using the original feature space. Rather, they indicate that, under the specific conditions investigated here, namely the evaluated descriptors, datasets, models, DR methods, and a target dimensionality of 100 features, DR did not provide a consistent performance benefit. In higher dimensional spaces, DR can be beneficial for reducing redundancy, improving computational efficiency, and potentially enhancing model performance [3].

A key design choice of this benchmark was to reduce all representations to a fixed dimensionality of 100 features, irrespective of the compound representation, DR method, prediction model, or assay dataset. The goal of this study was to benchmark different DR techniques under a standardized and practically feasible setting. Using a fixed target dimensionality ensured that differences in performance could be attributed to the dimensionality-reduction method itself rather than to extensive optimization of the number of retained components for each individual scenario. Furthermore, identifying the optimal number of retained dimensions would require hyperparameter optimization not only for each compound representation, but also for every assay dataset and prediction model, which was computationally infeasible within the scope of this study. Consequently, our conclusions are limited to this standardized setting and should not be interpreted as evidence that alternative target dimensionalities could not yield different results.

4.3 Cross Validation

Although Ash et al. [7] recommend a repeated 5x5 CV for datasets of our size, we only perform the statistical analysis on a 5-fold CV and have a nested 4-fold CV for hyperparameter tuning. Given the high number of representations, DR methods, and hyperparameter combinations, a full repeated design would have been prohibitively expensive, especially for CatBoost and TabPFN. The nested CV setup enabled robust hyperparameter tuning and performance estimation while keeping runtimes manageable. Moreover, because our fold generation is based on maximizing chemical dissimilarity, obtaining multiple independent CV partitions would have required relaxing the optimization constraints further or using a completely different splitting scheme [7]. We therefore opted for a single, chemically informed 5-fold split per assay to ensure

meaningful separation between training and test data. We used stratification to maintain class proportions across folds. This was necessary due to the high class imbalance in some assays. However, stratification can also counteract the objective of maximizing chemical dissimilarity; in particular, for rare positive compounds that cluster structurally. As a result, some assay splits may represent a compromise between chemical diversity and class ratio maintenance rather than an optimal configuration for either.

4.4 Impact of Chemically Informed Data Splitting

We use an ILP-based split that maximizes distances between folds using pairwise Tanimoto similarities of Morgan fingerprints [20]. While chemically informed splitting is essential to reduce similarity-based information leakage, this choice may partly explain the comparatively poor performance of Morgan fingerprints in our benchmark. Although the fingerprints used for splitting (1024-bit, radius 2) differ from the ones used as ML features (2048-bit, radius 3), they still represent the same family of descriptors. As a result, models trained on Morgan fingerprints may be exposed to more challenging splits than models trained on other representations. Prior work has also shown that MACCS keys and physicochemical properties can perform strongly and even outperform more complex representations [2, 72], but a targeted evaluation using alternative splitting criteria would be necessary to disentangle methodological biases from inherent representational differences.

Although chemically informed splits provide a more realistic evaluation setting than random CV [7, 54], they also introduce an important conceptual limitation: they implicitly assume that models should generalize out-of-distribution (OOD); yet, the models are not explicitly trained for OOD scenarios. Standard supervised ML methods rely on the assumption that training and test data are independently and identically distributed (i.i.d.). While the fulfilment of the i.i.d. assumption is questionable on chemical datasets, our distance-maximizing split intentionally violates this assumption. Thus, it effectively turns the evaluation into an OOD robustness test without equipping the models with mechanisms designed for robustness, such as adversarial robustness [73]. As we are, however, highly interested in models with a good generalization across the chemical space, it will be crucial for future work to develop approaches that explicitly target this type of robustness rather than relying solely on OOD-style splitting during evaluation.

4.5 Statistical Analysis

We used the Friedman test followed by the Nemenyi post-hoc analysis to investigate statistical differences between the different compound representations, DR methods, and ML models that we considered for our benchmarking. This test was chosen as it is commonly used for ML model comparison if the assumptions of parametric alternatives such as ANOVA (ANalysis Of VAriance) are not fulfilled [6, 7]. However, the performance estimates across CV folds are not strictly independent, as the training sets overlap substantially across the iterations. This dependence means that the formal assumptions of the Friedman test are not fully satisfied, making its applicability in this setting debatable. Nevertheless, we consider the test useful as an indicator

of whether observed performance differences are systematic rather than attributable to random variation. Therefore, we still believe that a statistical analysis, despite its limitations, is preferable to relying solely on raw performance values..

5 Conclusion

This manuscript presented a comprehensive benchmarking pipeline comparing compound representations, dimension reduction (DR) methods, and machine learning (ML) models for toxic endpoint prediction on hormone-related assays from the ToxCast database. We evaluated each of the three parts individually and in combination. In total, we compared 100 distinct feature-DR-model combinations per assay; however, due to extensive hyperparameter optimization, we trained more than 7,000 individual model instances per assay. While AUROC values, which are often used in literature imply good performance (ranging from ≈ 0.6 to ≈ 0.9), more realistic MCC results for imbalanced data are only modest with a median MCC of the best model performance of 0.32.

The statistical comparison of the individual pipeline parts revealed that, given our data and the structure-based split we are using, physicochemical properties are the most informative data type and rank significantly better than all other tested compound representations. Surprisingly, we showed that applying any of our tested DR methods is significantly worse than using no DR. From all tested models, the random forest (RF) consistently and significantly outperformed all other tested models – including deep neural networks. While the combination of RF with physicochemical properties and no DR also performs particularly good compared to the other 100 tested combinations, we note that it is not the best method in the majority of the assays.

Furthermore, we showed that the model performance is most strongly correlated to the class imbalance and the presence of activity cliffs. Activity cliffs represent a fundamental challenge for toxicity prediction because compounds with highly similar structures can exhibit different biological responses. This was exemplified for a ER β hormone receptor assays, where even small structural modifications can alter receptor interactions and lead to different activity labels. Consistent with this observation, a simple nearest-neighbour read-across baseline achieved a median MCC of 0.28, only slightly below the best ML models. This can be explained by the fact that read-across solely relies on the assumption that structurally similar compounds show the same activity. However, in the presence of activity cliffs this assumption does not hold. While the ML approaches outperformed this baseline, the relatively small margin suggests that much of the predictive signal is already captured by local chemical similarity. Other effects, such as protein-ligand interactions, are difficult to capture with conventional molecule-centric encodings. Thus, further research needs to be done to capture these effects in more physics-based representations.

Overall, our results suggest that further progress in toxicity prediction will likely require approaches that go beyond (global) structure similarity. Future work should focus on improving performance for highly imbalanced assays, while explicitly addressing activity cliffs. At the same time, the available data itself remains an important limitation. Many assays contain only a limited number of compounds and cover

restricted regions of chemical space. Expanding these datasets with additional compounds and broader chemical diversity will likely be essential for improving the generalization capabilities of future models. To this end, it might be beneficial to incorporate the performance on MMPs with different labels explicitly in the training objective. Moreover, instead of training ‘global’ ML models that are expected to perform well across different classes of molecules, it may be beneficial to focus on ‘local’ models tailored to a specific class of molecules of interest.

Despite these challenges, our study provides a comprehensive benchmarking analyses of hormone receptor activity prediction. By systematically evaluating representations, DR methods, ML models, and a read-across baseline as well as chemical space analysis, we establish a robust reference point for future method development and highlight both the opportunities and current limitations of ML-based toxicity prediction.

List of Abbreviations. A full list of abbreviations used in the manuscript can be found in Table 5.

Table 5 List of Abbreviations: Abbreviations and corresponding meanings used in the text alphabetically ordered.

abbreviation	meaning
ADME	absorption, distribution, metabolism, and excretion
ANOVA	analysis of variance
AR	androgen receptor
AUROC	area under the receiver-operator-characteristics curve
CatBoost	categorical boosting
CV	cross validation
DA	domain adaptation
DL	deep learning
D-MPNN	directed message passing neural network
DR	dimension reduction
ECFP	extended connectivity fingerprint
ER	estrogen receptor
FE	feature extraction
FS	feature selection
GNNs	graph neural networks
ILP	integer linear program
MACCS	molecular access system
MCC	Matthews correlation coefficient
MI	mutual information
ML	machine learning
MLP	multi-layer perceptron
MRMR	minimum redundancy maximum relevance
MTR	multi-task regression
NN	neural network
PCA	principal component analysis
PCC	Pearson correlation coefficient
RF	random forest
SMILES	simplified molecular input line entry system
SVM	support vector machine
TabPFN	tabular prior-data fitted network

Declarations

Supplementary information

Extra information on the hyperparameters, used package versions, and additional results can be found in the Appendix.

Acknowledgements

We thank Dr. Sebastian Dunst for guiding us with the selection of hormone-related ToxCast assays.

Moreover, we thank Ana Isabel Barreiros Mirandela for proofreading Section 3.6.

Funding

This work was performed in the context of the European RADAR project and has received funding from the the European Union’s Horizon program under grant agreement No. 101178148. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Competing interests

There are no competing interests to declare.

Ethics approval and consent to participate

Not applicable.

Data availability

We used publicly available data from ToxCast [4]. Moreover, the processed data as used for our analyses as well as our results are also uploaded on our GitHub repo <https://github.com/volkamerlab/ToxCastBenchmark>.

Code availability

The code used for our experiments as well as for generating the results plots is publicly available on our GitHub repo <https://github.com/volkamerlab/ToxCastBenchmark>.

Large Language Model Usage

Note that large language model (LLM) software (i.e., Microsoft Copilot and Claude Code) has been used in the preparation of this article exclusively for the generation of plotting scripts and for reviewing the text with a focus on grammar and wording. We confirm that everything was carefully checked and that we take the full responsibility of the submitted work.

Author contribution

LMR and AV conceptualized the study. LMR implemented the code, and wrote the initial draft of the manuscript. LMR, supported by LA, performed the experiments. LA generated the transformer-based embeddings of the molecules. AV supervised the study. All authors analyzed the results, reviewed and agreed to the manuscript.

References

- [1] Wu, Z., Ramsundar, B., Feinberg, E.N., Gomes, J., Geniesse, C., Pappu, A.S., Leswing, K., Pande, V.: Moleculenet: a benchmark for molecular machine learning. *Chemical science* **9**(2), 513–530 (2018)
- [2] Sultan, A., Rausch-Dupont, M., Khan, S., Kalinina, O., Klakow, D., Volkamer, A.: Transformers for molecular property prediction: Domain adaptation efficiently improves performance. *arXiv preprint arXiv:2503.03360* (2025)
- [3] Eckhart, L., Lenhof, K., Rolli, L.-M., Lenhof, H.-P.: A comprehensive benchmarking of machine learning algorithms and dimensionality reduction methods for drug sensitivity prediction. *Briefings in bioinformatics* **25**(4) (2024)
- [4] US EPA ORD, Center for Computational Toxicology: ToxCast Database: invit-rodb version 4.2. Dataset. Retrieved July 2025. (2024). <https://doi.org/10.23645/epacomptox.6062623.v13>
- [5] Joeres, R., Blumenthal, D.B., Kalinina, O.V.: Data splitting to avoid information leakage with datasail. *Nature Communications* **16**, 3337 (2025) <https://doi.org/10.1038/s41467-025-58606-8>
- [6] Demšar, J.: Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research* **7**(Jan), 1–30 (2006)
- [7] Ash, J.R., Wognum, C., Rodríguez-Pérez, R., Aldeghi, M., Cheng, A.C., Clevert, D.-A., Engkvist, O., Fang, C., Price, D.J., Hughes-Oliver, J.M., *et al.*: Practically significant method comparison protocols for machine learning in small molecule drug discovery. *Journal of chemical information and modeling* **65**(18), 9398–9411 (2025)
- [8] Jiang, X., Tan, L., Cen, J., Zou, Q.: Molbench: A benchmark of ai models for molecular property prediction. In: *International Symposium on Benchmarking, Measuring and Optimization*, pp. 53–70 (2023). Springer
- [9] Russo, D.P., Zorn, K.M., Clark, A.M., Zhu, H., Ekins, S.: Comparing multiple machine learning algorithms and metrics for estrogen receptor binding prediction. *Molecular pharmaceutics* **15**(10), 4361–4370 (2018)
- [10] Zorn, K.M., Foil, D.H., Lane, T.R., Hillwalker, W., Feifarek, D.J., Jones, F., Klaren, W.D., Brinkman, A.M., Ekins, S.: Comparison of machine learning

- models for the androgen receptor. *Environmental science & technology* **54**(21), 13690–13700 (2020)
- [11] Belfield, S.J., Cronin, M.T., Enoch, S.J., Firman, J.W.: Guidance for good practice in the application of machine learning in development of toxicological quantitative structure-activity relationships (qsars). *PloS one* **18**(5), 0282924 (2023)
 - [12] Kim, D., Jeong, J., Choi, J.: Identification of optimal machine learning algorithms and molecular fingerprints for explainable toxicity prediction models using toxcast/tox21 bioassay data. *ACS omega* **9**(36), 37934–37941 (2024)
 - [13] Guo, J.: Do larger models really win in drug discovery? a benchmark assessment of model scaling in ai-driven molecular property and activity prediction. *arXiv preprint arXiv:2604.26498* (2026) <https://doi.org/10.48550/arXiv.2604.26498>
 - [14] Williams, A.J., Grulke, C.M., Edwards, J., McEachran, A.D., Mansouri, K., Baker, N.C., Patlewicz, G., Shah, I., Wambaugh, J.F., Judson, R.S., *et al.*: The comptox chemistry dashboard: a community data resource for environmental chemistry. *Journal of cheminformatics* **9**(1), 61 (2017)
 - [15] U.S. Environmental Protection Agency: CompTox Chemicals Dashboard. <https://comptox.epa.gov/dashboard/>. Version 2.5.2, released February 2025. Accessed July 2025 (2025)
 - [16] Landrum, G., *et al.*: RDKit: Open-source cheminformatics (version 2025.03.3) <https://doi.org/10.5281/zenodo.15605628>
 - [17] Weininger, D.: Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences* **28**(1), 31–36 (1988)
 - [18] Richard, A.M., Judson, R.S., Houck, K.A., Grulke, C.M., Volarath, P., Thillainadarajah, I., Yang, C., Rathman, J., Martin, M.T., Wambaugh, J.F., Knudsen, T.B., Kancharla, J., Mansouri, K., Patlewicz, G., Williams, A.J., Little, S.B., Crofton, K.M., Thomas, R.S.: Toxcast chemical landscape: Paving the road to 21st century toxicology. *Chemical Research in Toxicology* **29**(8), 1225–1251 (2016) <https://doi.org/10.1021/acs.chemrestox.6b00135> <https://pubs.acs.org/crtoec/article-pdf/29/8/1225/13068454/tx6b00135.pdf>
 - [19] Sieg, J., Feldmann, C.W., Hemmerich, J., Stork, C., Sandfort, F., Eiden, P., Mathea, M.: Molpipeline: A python package for processing molecules with rdkit in scikit-learn. *Journal of Chemical Information and Modeling* **64**(24), 9027–9033 (2024)
 - [20] Joeres, R., Blumenthal, D.B., Kalinina, O.V.: Clustering. <https://datasail>.

- readthedocs.io/en/latest/workflow/clustering.html, [Online; accessed 02-March-2026]
- [21] Bolusani, S., Besançon, M., Bestuzheva, K., Chmiela, A., Dionísio, J., Donkiewicz, T., Doornmalen, J., Eifler, L., Ghannam, M., Gleixner, A., Graczyk, C., Halbig, K., Hedtke, I., Hoen, A., Hojny, C., Hulst, R., Kamp, D., Koch, T., Kofler, K., Lentz, J., Manns, J., Mexi, G., Mühmer, E., Pfetsch, M.E., Schlösser, F., Serrano, F., Shinano, Y., Turner, M., Vigerske, S., Weninger, D., Xu, L.: The SCIP Optimization Suite 9.0. Technical report, Optimization Online (February 2024). <https://optimization-online.org/2024/02/the-scip-optimization-suite-9-0/>
 - [22] Bolusani, S., Besançon, M., Bestuzheva, K., Chmiela, A., Dionísio, J., Donkiewicz, T., Doornmalen, J., Eifler, L., Ghannam, M., Gleixner, A., Graczyk, C., Halbig, K., Hedtke, I., Hoen, A., Hojny, C., Hulst, R., Kamp, D., Koch, T., Kofler, K., Lentz, J., Manns, J., Mexi, G., Mühmer, E., Pfetsch, M.E., Schlösser, F., Serrano, F., Shinano, Y., Turner, M., Vigerske, S., Weninger, D., Xu, L.: The SCIP Optimization Suite 9.0. ZIB-Report 24-02-29, Zuse Institute Berlin (February 2024). <https://nbn-resolving.org/urn:nbn:de:0297-zib-95528>
 - [23] Yap, C.W.: Padel-descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of computational chemistry* **32**(7), 1466–1474 (2011)
 - [24] Landrum, G., et al.: RDKit Documentation - rdkit.Chem.Descriptors module. <https://www.rdkit.org/docs/source/rdkit.Chem.Descriptors.html>. [Online; accessed 11-November-2025]
 - [25] Landrum, G., et al.: Feature Definitions Used in the Morgan Fingerprints. <https://www.rdkit.org/docs/GettingStartedInPython.html#feature-definitions-used-in-the-morgan-fingerprints>. [Online; accessed 05-November-2025]
 - [26] Rogers, D., Hahn, M.: Extended-connectivity fingerprints. *Journal of chemical information and modeling* **50**(5), 742–754 (2010)
 - [27] O’Boyle, N.M., Sayle, R.A.: Comparing structural fingerprints using a literature-based similarity benchmark. *Journal of cheminformatics* **8**(1), 36 (2016)
 - [28] Fabian, B., Edlich, T., Gaspar, H., Segler, M., Meyers, J., Fiscato, M., Ahmed, M.: Molecular representation learning with language models and domain-relevant auxiliary tasks. *arXiv preprint arXiv:2011.13230* (2020)
 - [29] Chithrananda, S., Grand, G., Ramsundar, B.: Chemberta: large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885* (2020)

- [30] Sultan, A., Sieg, J., Mathea, M., Volkamer, A.: Transformers for molecular property prediction: Lessons learned from the past five years. *Journal of Chemical Information and Modeling* **64**(16), 6259–6280 (2024)
- [31] Brown, N., Fiscato, M., Segler, M.H., Vaucher, A.C.: Guacamol: benchmarking models for de novo molecular design. *Journal of chemical information and modeling* **59**(3), 1096–1108 (2019)
- [32] Cios, K.J., Swiniarski, R.W., Pedrycz, W., Kurgan, L.A.: Feature Extraction and Selection Methods, pp. 133–233. Springer, Boston, MA (2007). https://doi.org/10.1007/978-0-387-36795-8_7 . https://doi.org/10.1007/978-0-387-36795-8_7
- [33] Kwak, N., Choi, C.-H.: Input feature selection for classification problems. *IEEE transactions on neural networks* **13**(1), 143–159 (2002)
- [34] Vergara, J.R., Estévez, P.A.: A review of feature selection methods based on mutual information. *Neural computing and applications* **24**(1), 175–186 (2014)
- [35] Ali, P.J.M., Faraj, R.H., Koya, E., Ali, P.J.M., Faraj, R.H.: Data normalization and standardization: a technical report. *Mach Learn Tech Rep* **1**(1), 1–6 (2014)
- [36] St, L., Wold, S., *et al.*: Analysis of variance (anova). *Chemometrics and intelligent laboratory systems* **6**(4), 259–272 (1989)
- [37] Koike, T., Lin, L., Wang, R.: Invariant correlation under marginal transforms. *Journal of Multivariate Analysis* **204** (2024) <https://doi.org/10.1016/j.jmva.2024.105361> . Publisher Copyright: © 2024 The Author(s)
- [38] James, G., Witten, D., Hastie, T., Tibshirani, R., Taylor, J.: An introduction to statistical learning: With applications in python (2023) <https://doi.org/10.1007/978-3-031-38747-0>
- [39] Jolliffe, I.T., Cadima, J.: Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **374**(2065), 20150202 (2016) <https://doi.org/10.1098/rsta.2015.0202> <https://royalsocietypublishing.org/rsta/article-pdf/doi/10.1098/rsta.2015.0202/1381479/rsta.2015.0202.pdf>
- [40] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011)
- [41] Lenhof, K.: Machine learning-based anti-cancer drug treatment optimization (2024)
- [42] Guo, W., Liu, J., Dong, F., Song, M., Li, Z., Khan, M.K.H., Patterson, T.A.,

- Hong, H.: Review of machine learning and deep learning models for toxicity prediction. *Experimental Biology and Medicine* **248**(21), 1952–1973 (2023)
- [43] Burns, J.W., Zalte, A.S., Abreu, C.R., Sieg, J., Feldmann, C., Mathea, M., Green, W.H.: Deep learning foundation models for low-data regimes from classical molecular descriptors. *Journal of Chemical Information and Modeling* (2026) <https://doi.org/10.1021/acs.jcim.6c01546>
- [44] Wiltos, K.: Impact of feature standardization on classification process using pca and svm algorithms. In: Lopata, A., Gudonienė, D., Butkienė, R., Čeponis, J. (eds.) *Information and Software Technologies*, pp. 3–13. Springer, Cham (2025)
- [45] Breiman, L.: Random forests. *Machine learning* **45**(1), 5–32 (2001)
- [46] Pinheiro, J.M.H., Oliveira, S.V.B., Silva, T.H.S., Saraiva, P.A.R., Souza, E.F., Godoy, R.V., Ambrosio, L.A., Becker, M.: The impact of feature scaling in machine learning: Effects on regression and classification tasks. *IEEE Access* **13**, 199903–199931 (2025)
- [47] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A.: Catboost: unbiased boosting with categorical features. In: *Advances in Neural Information Processing Systems*, vol. 31. Curran Associates, Inc., Red Hook, NY, USA (2018)
- [48] Friedman, J.H.: Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232 (2001)
- [49] The CatBoost team: Transforming categorical features to numerical features. <https://catboost.ai/docs/en/concepts/algorithm-main-stages-cat-to-numeric>. [Online; accessed 18-November-2025]
- [50] Hornik, K., Stinchcombe, M., White, H.: Multilayer feedforward networks are universal approximators. *Neural networks* **2**(5), 359–366 (1989)
- [51] Hollmann, N., Müller, S., Purucker, L., Krishnakumar, A., Körfer, M., Hoo, S.B., Schirrmeister, R.T., Hutter, F.: Accurate predictions on small data with a tabular foundation model. *Nature* **637**(8045), 319–326 (2025)
- [52] Hollmann, N., Müller, S., Eggensperger, K., Hutter, F.: TabPFN: A transformer that solves small tabular classification problems in a second. In: *International Conference on Learning Representations 2023* (2023)
- [53] Morger, A., Mathea, M., Achenbach, J.H., Wolf, A., Buesen, R., Schleifer, K.-J., Landsiedel, R., Volkamer, A.: Knowtox: pipeline and case study for confident prediction of potential toxic effects of compounds in early phases of development. *Journal of Cheminformatics* **12**(1), 24 (2020)

- [54] Seal, S., Mahale, M., García-Ortegón, M., Joshi, C.K., Hosseini-Gerami, L., Beatson, A., Greenig, M., Shekhar, M., Patra, A., Weis, C., *et al.*: Machine learning for toxicity prediction using chemical structures: Pillars for success in the real world. *Chemical research in toxicology* **38**(5), 759–807 (2025)
- [55] Chicco, D., Jurman, G.: The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics* **21**(1), 6 (2020)
- [56] Chicco, D., Jurman, G.: The matthews correlation coefficient (mcc) should replace the roc auc as the standard metric for assessing binary classification. *BioData Mining* **16**(1), 4 (2023)
- [57] Fisher, R.A.: Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika* **10**(4), 507–521 (1915)
- [58] Friedman, M.: The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the american statistical association* **32**(200), 675–701 (1937)
- [59] Iman, R.L., Davenport, J.M.: Approximations of the critical region of the fbi-etkan statistic. *Communications in Statistics-Theory and Methods* **9**(6), 571–595 (1980)
- [60] Maggiora, G.M.: On outliers and activity cliffs why QSAR often disappoints. *ACS Publications* (2006)
- [61] Cavasotto, C.N., Scardino, V.: Machine learning toxicity prediction: latest advances by toxicity end point. *ACS omega* **7**(51), 47536–47546 (2022)
- [62] Mayr, A., Klambauer, G., Unterthiner, T., Hochreiter, S.: Deeptox: Toxicity prediction using deep learning. *Frontiers in Environmental Science* **Volume 3 - 2015** (2016) <https://doi.org/10.3389/fenvs.2015.00080>
- [63] Banerjee, P., Siramshetty, V.B., Drwal, M.N., Preissner, R.: Computational methods for prediction of in vitro effects of new chemical structures. *Journal of cheminformatics* **8**(1), 51 (2016)
- [64] Banerjee, P., Eckert, A.O., Schrey, A.K., Preissner, R.: Protox-ii: a webserver for the prediction of toxicity of chemicals. *Nucleic acids research* **46**(W1), 257–263 (2018)
- [65] Shah, I., Liu, J., Judson, R.S., Thomas, R.S., Patlewicz, G.: Systematically evaluating read-across prediction and performance using a local validity approach characterized by chemical structure and bioactivity information. *Regulatory Toxicology and Pharmacology* **79**, 12–24 (2016)

- [66] Chushak, Y.G., Shows, H., Gearhart, J., Pangburn, H.: In silico identification of protein targets for chemical neurotoxins using toxcast in vitro data and read-across within the qsar toolbox. *Toxicology research* **7**(3), 423–431 (2018)
- [67] Chen, L.: In: LIU, L., ÖZSU, M.T. (eds.) *Curse of Dimensionality*, pp. 545–546. Springer, Boston, MA (2009). https://doi.org/10.1007/978-0-387-39940-9_133 . https://doi.org/10.1007/978-0-387-39940-9_133
- [68] Van Tilborg, D., Alenicheva, A., Grisoni, F.: Exposing the limitations of molecular machine learning with activity cliffs. *Journal of chemical information and modeling* **62**(23), 5938–5951 (2022)
- [69] Klutzny, S., Kornhuber, M., Morger, A., Schoenfelder, G., Volkamer, A., Oelgeschlaeger, M., Dunst, S.: Quantitative high-throughput phenotypic screening for environmental estrogens using the e-morph screening assay in combination with in silico predictions. *Environment international* **158**, 106947 (2022)
- [70] Wang, Y., Yao, Q., Kwok, J.T., Ni, L.M.: Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)* **53**(3), 1–34 (2020)
- [71] Snyder, S.H., Vignaux, P.A., Ozalp, M.K., Gerlach, J., Puhl, A.C., Lane, T.R., Corbett, J., Urbina, F., Ekins, S.: The goldilocks paradigm: comparing classical machine learning, large language models, and few-shot learning for drug discovery applications. *Communications Chemistry* **7**(1), 134 (2024)
- [72] Eckhart, L., Lenhof, K., Herrmann, L., Rolli, L.-M., Lenhof, H.-P.: How to predict effective drug combinations—moving beyond synergy scores. *Iscience* **28**(6) (2025)
- [73] Lenhof, K., Rolli, L.-M., Buhr, L., Roth, S., Binkyte, R., Schicktanz, S., Fritz, M., Volkamer, A., Beerenwinkel, N.: The trustworthiness landscape in machine learning: a conceptual guide with applications in medicine (2025)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [appendix.pdf](#)